Tag Based Bayesian Latent Class Models for Movies:

Economic Theory reaches out to Big Data Science

Lucy Amez

A thesis presented for the degree of

Doctor of Philosophy

Statistics, Operational Research and Mathematics (StORM)

London Metropolitan University

March 2017

LONDON METROPOLITAN UNIVERSITY

Tag Based Bayesian Latent Class Models for Movies:

Economic Theory reaches out to Big Data Science

Lucy Amez

A thesis presented for the degree of
Doctor of Philosophy

Statistics, Operational Research and Mathematics (StORM)
London Metropolitan University
March 2017

Copyright ©Lucy Amez

To those that were, and are always there,

to those that are, and were always there

Contents

Preface

People who know me will confirm I don't have a jealous nature. More the type that likes to see that everyone lives up to their full potential. That said, the ink of the first words is softly dipped in a drop of envy towards those who have an array of thank you notes to write. It most likely signals that their PhD path was not all that solitary. Mine, it is fair to say, was not paved with roses. Too many setbacks, too much of this and that. Family, friends, work, all joined forces to push my life forward. In the end, it was the persistent will to finish what I had once started that prevailed over everything else. However, it isn't to myself I dedicate this work to, nor to anyone else, but to those I lost in the course of this endeavour. They remain my ultimate source and reference.

It was a winter day in the early years 2000 when I arrived in Vienna for what was to be my first cultural economics conference. There was snow on the tarmac and my quickly compiled suitcase didn't carry what was needed to accommodate to the climatic demands. I didn't experience any cold. Discussing my topic with like minded, my first visit to Wiener Staatsoper and the burned calories of the Sachertorte, it all provided me with sufficient energy. The conference dinner took place at a cosy restaurant and seats were allocated in a rather random fashion. At the table with me was Professor John Sedgwick, very English, from the University of North London.

When I reach the age of retirement, I will reflect further on the non randomness of randomness. Fact is that John took an immediate and genuine interest in my work,

followed up on it during the next conferences and years later offered me to pursue my doctoral research with him. I'm sure he didn't foresee all the consequences of this invite - the rational expectations hypothesis once more rejected - but he stayed supportive till the last administrative form. I will never know why John helped me the way he did. Maybe because, not unlike myself, he resents wasted potential. For me, the tag altruism is one to one connected to him. When he decided to leave London Met, I felt devastated at first, but he pointed me to Professor Stasinopoulos. Because my approach of things was getting more probabilistic, this proved to be a wise choice.

This thesis has the annotation creative good attached to it. Partly because it reflects on it, partly because it wants to be one. Creativity is intelligence having fun, Einstein said. If so, than this work is likely to qualify. My thinking went through multiple exciting loops before landing on a Bayesian mixture of economics and recommender theory. It captures artistic commodities that don't allow themselves to get framed in formal models. Art and science united in content, like they should be more often and like they often were during this project. It reminds me to mention Johann Sebastian, Georg Friedrich and Giuseppe by Josep, for their inspiring notes accompanying me during the long writing hours.

I don't wish to blow the cream off the cup now, but artists rarely pay your bills. My employer was in that respect a far better ally by rewarding me punctually for my services. While the copyright of this research is fully attributed to a private company called Lucy's Free Time, an entity my job did steal a lot from, I owe the Vrije Universiteit Brussel for providing me a stable and intellectually challenging work environment over the past decade. My acquired big data expertise will most certainly find its feedback into my job. Many thanks to them and to my truly great colleagues.

And so you see, once you start spelling it out, there is actually a lot to be thankful for. Indeed, taking part in global scientific communication is a privilege, at least if

you let intelligence enjoy itself and allow space for randomness. Time to let the black ink dry and pass my words on to the reader. Professor Alan Collins of Portsmouth University and Professor Bob Gilchrist of London Metropolitan University, kindly accepted to be the first judges of its content. I do hope that they, and anyone reading this work thereafter, experience an interesting journey, going from economic theory to big data science and back, and label its contribution as a worthy addition to the research topic of cultural economics.

Abstract

For the past 50 years, cultural economics has developed as an independent research specialism. At its core are the creative industries and the peculiar economics associated with them, central to which is a tension that arises from the notion that creative goods need to be experienced before an assessment can be made about the utility they deliver to the consumer. In this they differ from the standard private good that forms the basis of demand theory in economic textbooks, in which utility is known ex ante. Furthermore, creative goods are typically complex in composition and subject to heterogeneous and shifting consumer preferences. In response to this, models of linear optimization, rational addiction and Bayesian learning have been applied to better understand consumer decision- making, belief formation and revision. While valuable, these approaches do not lend themselves to forming verifiable hypothesis for the critical reason that they by-pass an essential aspect of creative products: namely, that of novelty. In contrast, computer sciences, and more specifically recommender theory, embrace creative products as a study object. Being items of online transactions, users of creative products share opinions on a massive scale and in doing so generate a flow of data driven research. Not limited by the multiple assumptions made in economic theory, data analysts deal with this type of commodity in a less constrained way, incorporating the variety of item characteristics, as well as their co-use by agents. They apply statistical techniques supporting big data, such as clustering, latent class analysis or singular value decomposition.

This thesis is drawn from both disciplines, comparing models, methods and data sets. Based upon movie consumption, the work contrasts bottom-up versus top-down approaches, individual versus collective data, distance measures versus the utility-based comparisons. Rooted in Bayesian latent class models, a synthesis is formed, supported by the random utility theory and recommender algorithm methods. The Bayesian approach makes explicit the experience good nature of creative goods by formulating the prior uncertainty of users towards both movie features and preferences. The latent class method, thus, infers the heterogeneous aspect of preferences, while its dynamic variant- the latent Markov model - gets around one of the main paradoxes in studying creative products: how to analyse taste dynamics when confronted with a good that is novel at each decision point. Generated by mainly movie-user-rating and movie-user-tag triplets, collected from the Movielens recommender system and made available as open data for research by the GroupLens research team, this study of preference patterns formation for creative goods is drawn from individual level data.

Keywords: consumer segmentation, creative goods, experience goods, movie choice, Bayesian statistics, latent class models, recommender theory, online data, tags

Chapter 1

Research Outline

1.1 Introduction

Consumption patterns for creative goods in general and movies in particular were the topic of multiple studies in various disciplines and speciality fields of the social sciences and humanities. Also policy studies took a high interest in the economic importance of the creative sectors, where the movie industry stands out as a key sector, contributing substantially to the overall value added creation. Where sociologists studied demand in terms of class segregation, rooted in social relations, as the force behind cultural capital formation, economic theory and more specifically cultural economists focused mainly on movie sector characteristics and their impact on revenue and box office structure. The evolution of those performance indicators followed a capricious pattern which is lead back to the unpredictable nature of demand. The expression "nobody knows anything", a saying by William Goldman (1983), and later rephrased by Caves (2000) in his book *"Creative Industries"*, became the philosophical basso continuo accompanying mainstream reflection of the movie market. It emphasizes the near impossibility to predict hits nor failures in advance. The randomness of the supply side of the movie market is attributed to elements shaping the demand side, where agents face uncertainty each time they have to go through the process of deciding what product to opt for. Movies have

been labelled as experience goods, a concept going back to Nelson (1970), denoting a group of commodities where the quality or the compository features are hard to assess at a pre-purchase phase, but are partly revealed by undergoing the experience. If uncertainty stays in place, one speaks of credence goods. Also the concept of information good, put forward by Varian (1998), has been brought in connection with creative products, referring to items where not the material carrier determines value, rather the information attached to it. All those conceptual classes point to the strive of consumers to gain sufficient knowledge in order to make adequately informed choices. The determining factors for movie choice that are singled out in cultural economics literature largely fall under the nominator of "quality certifiers", providers of knowledge to others, such as reviews or Oscars, or formats such as sequels that are known from the past.

Experience economy was a trending concept during the past decades. Experience is endowed with a dual meaning where it can refer to immediate joy or satisfaction as well as to long term accumulated capital formation. The first has been of interest to marketing researchers studying the nature of the experience and its main drivers. The second view was covered by branches of economic science and psychology, incorporating elements of the cognitive. They look at it from the perspective of a rational agent who makes decisions given an information set, which is adjusted based on past information, but can be evoked in the light of future decisions. Those two stands are the subject of the influential paper by Holbrook & Hirschman (1982), placing the first in the sphere of the symbolic, hedonic, having connotations to sensations and emotions, while the latter connects to cognition, linked to processes of belief formation and learning. The duality runs in parallel to a second dichotomy in the study of preferences for creative products, that of cumulative taste formation versus innate stable preferences. Part of scholarly research looks at taste as innate and stable, Peltoniemi (2015), a view that is supported by studies of genre attachment over time, while others picture taste as acquired through time, which is revealed by consumers getting an interest or becoming specialist in a particular style. The concept of experience goods enriched research in cultural economics with a conceptual underpinning for a large number of empirical movies studies. It served

as a container concept whilst largely passing by the elements determining quality uncertainty nor its consequences in terms of post-consumption assessment resulting in potential taste shifts. Consumers are faced with products that are novel at each consumption point. The innovation aspect is what makes a creative product to what it is. However, what meaning can be attributed to consistency, or adverse shifts in tastes, with respect to goods which exhibit high degrees of novelty each time a choice is made, a decision coming about after answering what elements are cognitively and emotionally lightening up as crucial features to opt in or opt out. Previous studies included objective elements such as the presence of actors, directors, location, season, genre or subjective explanatory variables like self-reflection or arousal. However, results have been inconclusive, particularly so when the research has been intertemporal in nature.

Shifts in taste as well as the multi-featured nature of creative goods challenge the neoclassical economical paradigm. Addressing a number of caveats in a review article *"The new science of pleasure, consumer choice behaviour and the measurement of well-being"*, McFadden (2014) explicits the fact that, when dealing with larger scale micro data on consumer behaviour, the neoclassical econometric demand systems based on a representative consumer are "uncomfortably restrictive" and show difficulties in dealing with preference heterogeneity, acquired taste, shifting hedonic attributes of commodities,.., time, space and uncertainty. He shows how, in the light of those challenges, theories were developed, preserving the core ideas of consumer sovereignty and utility maximization while at the same time incorporating broader components. They include hedonic models such as developed by Lancaster (1966), household production functions or contingent valuation models to deal with the featured nature of commodities, life-time utility discounting to handle consumer dynamics or theoretical models adding an experience variable into the utility function to introduce heterogeneity in preference based on acquired taste. It remains problematic however to measure and integrate all the varied experiences of consumers. McFadden (2014) offers a solution through a discrete choice model based on random utility maximization (RUM), thereby shifting the focus from individual preferences to the distribution of preferences. However, taste variations by or across individuals

remain troublesome as they undermine revealed preference results. The RUM does translate nicely into a logistic regression model, which is an attractive tool to handle the estimation of the characteristics in a binary choice model. In an expanded version, latent class logistic regression, potentially deals with consumer heterogeneity, be it under the assumption of categories of individuals behaving in a similar way in a defined consideration class. Its intertemporal variant, latent Markov models allow to consider whether or not consumers change segments over time, thereby testing the presence of taste stability or shift.

The study of consumer behaviour for creative goods puts pressure on extant economic theory by exposing its conceptual and methodological limitations and forcing theorists to adjust in order to deal with its challenging features. At the same time, empirical cultural econometrics bounced against the borders of data availability to assure a satisfactory mapping of consumer's motives and long term choice patterns, asking for the presence of longitudinal micro level data. Movie economics performance studies mainly investigate box-office distributions, coming from aggregated data, where the analysis of choice dynamics heavily relies on country level data. The study of box-office data has to be acknowledged, as it provided valuable insights in the skewed nature of the performance distributions characterizing film revenues, pointing to the asymmetrical chances of creating hits or failures and clarifying some of the main causes determining the observed phenomena. Given that those causes find their origin at the demand side, more specifically in consumer choice, it seems that looking at micro level data is the only way to gain sufficient insight in how agents might respond to a new product being offered to the market. Only a few movie economics studies use micro level data, despite the observation that movie forums are a rich information source on consumer's opinion. This state of affairs contrasts strongly with that of the computer science literature where data analytic tools for market segmentation of creative goods, in the form of recommender systems that make automatic recommendations to users based on the abstraction of his/her preferences, have a firm presence. These tools are used by e-commerce companies and rely heavily on online data of creative products, and more specifically on movie ratings, to build behavioural models, to test and to predict consumer choice. The

worldwide web has been expansive when it came to user involvement. Clickstream data, search terms, all generate big information flows used to improve interfaces or steer consumer choice. Through blogs or opinion sites, consumers share thoughts on the products they come to experience. Especially tags are a promising addition for those investigating features that influence consumer choice. Tags are user generated keywords to annotate an object. When added by a community of users, a vocabulary arises, named a folksonomy, Vander Wal (2007). It refers to the general public attaching a collection of terms, characterized by its own dynamics of word generation and reuse. Being freely annotated, tags can be expected to reflect, at least partly, what features individuals label as important in their relation to the product.

In contrast to econometric studies, recommender theory is based on a bottom up approach to data. While specifying in vague terms the notion of consumer preferences, researchers start from looking at the data to discover distinctive patterns. Not limited by the multiple assumptions made in economic theory, data scientists deal with this type of commodities in a less constrained way, incorporating their characteristics as well as their co-use by agents. They take into account the heterogeneous nature of users, expressed in consumer subgroups. Also the statistical techniques presented in their literature fundamentally differ, employing methods that support the treatment of big data, such as clustering, latent class analysis or singular value decomposition. Where those methods did certainly influence cultural economics research, the bulk of studies is based on linear regression analysis. Models for online recommendation follow two major streams of analysis. A first approach relies on the assumption of intertemporal consistency in taste and looks at the content of the product in terms of its features. Content based systems examine what type of features were desirable to an individual in the past and proposes new items with similar characteristics. A second strand looks at the community of users, singles out groups showing similar taste on items, and uses the average rating of peers to advise individuals. A particular subset, namely model based systems, adhere the idea that latent classes or segments steer consumer groups. In a hybrid formulation, they also allow the integration of object related features. Especially the probabilistic latent class recommender system put forward by Hofmann & Puzicha (1999),

bridges recommender theory with economic models.

The main objective of this thesis is to investigate preference profiles for creative goods, considering movies as a typical case. That implies addressing a number of conceptual and methodological shortcomings which currently prevent proper empirical investigation of an individual's decision making process for experience goods. Starting from the sketched state of affairs, a number of research questions can be formulated that will be addressed in this investigation:

1. Upon acceptance of the definition of movies as an experience good, in what way can the distinct elements of that concept, the multi-characteristic nature, the uncertainty and the belief formation involved, be integrated in a consistent theoretical framework that allows empirical verification of micro-economic behavioural consumer patterns?

2. What insights, techniques or data derived form computer sciences, and more specifically from the theory of recommender systems, can be transferred to economics and enrich research on consumer preferences for creative goods?

3. In what way can the study of online social data generated by consumers through their tagging behaviour provide information on the main dimensions that steer their choice for movies?

4. Is it meaningful to introduce heterogeneity or perform preference segmentation based on distinct patterns of attachment of consumers towards an array of predefined features detached from social information?

5. What is the value added of Bayesian models based on latent class theory to discover typical preference patters and to deal with the intertemporal nature of experience good consumption?

6. How to deal with the paradox of studying dynamic consumer behaviour when dealing with novelty goods: can both be incorporated into a single analytical framework?

This PhD project is intrinsically multidisciplinary. A study of how to treat creative goods from the perspective of economic theory is matched against a cross section of recommender theory offered in computer science literature. The aim is to investigate which concepts, methods or data are available and can be transferred to the study of individual consumer patterns for movies. In doing so, the analysis touches upon psychological and philosophical matters in respectively discussing a user's recognition of stereotypes and the nature of genre concept. The investigative path is preceded with a literature review, recognising the contributions of key economists over the discipline and the multiple insights offered over the past decades on movie sector characteristics both from the demand and supply side. Following this, some gaps are identified to be rephrased in terms research opportunities. To begin with, the concept of experience goods and the multi-featured nature of creative products are unravelled. Because commodity composition differs at each time point, new experience can only be judged based on its feature similarity when compared to prototypes. One will have to fall back on insights of classification and categorization. Here, the contrast model initiated by Tversky (1977) offers a valuable perspective, where the degree of similarity and dissimilarity between features is judged upon in additive way. When integrating this back into a theory of utility, as explained, the Random Utility Model and its empirical offspring, logistic regression are foregrounded as the conceptual core building blocks of this work. However, in order to allow translation into an empirical setting, the question remains what features are included by individuals when selecting a movie. To answer this, the potential of online social information is investigated and in particular the value added of tags as signals of user's motives is further examined. This is done in reference to the recommender theories, reviewed in chapter 3, and to tag quality studies, part of data chapter 4. A selection of the most relevant tags will be used as explanatory variables in the proposed latent class models.

The statistical techniques used in this study are fundamentally Bayesian. Not only because Bayesian statistics offers a coherent tool set to structure and reduce massive data, but because mixture models present a consist theoretical framework on how to infer different levels of posterior probability distributions given a set of observations. In doing this, Bayesian statistics appears to be best equipped to detect the uncertain relationship of users versus their decision making aspects and to pin down the dynamics of the underlying belief process. In this thesis, Bayesian models were tested on a sample taken from the MovieLens dataset. MovieLens (section 4.1) is a recommender system initiated by the Grouplens research group. They gathered a big stream of online information, that was structured and made available as open data for research purposes. A large amount of academic papers rely on this particular data. Three Bayesian models are applied, each generating a chapter of the thesis. The first, treated in chapter 5, is a topic model or a Latent Dirichlet Allocation model. This tool is mainly used for text mining purposes, bringing together terms having a large probability of co-appearance. It not only divides the set of tags into semantic groups, thereby estimating the likelihood of a tag belonging to a class, by setting a Dirichlet prior over the topic distribution, it estimates the chance of an individual being assigned to a class. This exercise serves as a first investigation in the nature of tags with the aim of selecting the most relevant keywords. Considering only the most frequent terms might exclude tags of value to a particular segment of users. A selection of tags is used as explanatory variables in the Bayesian Latent Class Logistic Regression model of Chapter 6. This can be considered the core chapter, as it investigates the relationship between movie choice and some key tags, integrating potential heterogeneity in taste through the presence of latent classes or segments. It allows to discover the main consumer patters, to compare a multi-class with a uni-class model in terms of model fit and to establish the value added of tags compared to genre as a class separator.

The models of chapter 5 and 6 are static, looking for patterns when the data are taken over the entire time period. The last chapter introduces dynamics by allowing users to switch segments between periods. To do so, a Latent Class Markov Model is applied on the data. Apart from estimating the relationship between choice and

tags, this type of models generates transition probabilities, indicating the probability users stay in a particular segment or move away form it. By letting the transition depend on rating, a model of intertemporal belief change is tested, taking us back to the idea of experience goods as a dynamic construct based on belief revision following past evaluation. It is an initial experimental trial to approach persistency in taste not in terms of products, but in terms of loyalty to segments. It completes a line of thought that starts with the economics of experience goods, featuring similarity and random utility models, and passes to computer science, where open datasets and the use of tags are invoked. Finally the two disciplines are integrated through probabilistic latent class models.

1.2 Literature review

From an academic perspective, a discipline is a particular branch of knowledge which unifies a theoretical paradigm. It defines a set of notions of academic credibility and intellectual substance, and agrees largely, though not fully, to a taught subject or a university department, Becher & Trowler (2001), Krishnan (2009). A research specialty is a less outlined, self organized network of researchers who study the same research topics, attend the same conferences, publish in the same journals and read and cite each others research papers, Morris & Van der Veer Martens (2008). Small (1973) speaks of "a consensual structure of concepts in a field, employed through its citation and co-citation network". The counterpart in the publication sphere is that of invisible college referring to networks of literature being connected by reference to each other without being linked by a formal institutional structure, de Solla Price (1963), Crane (1972), Lievrouw (2014). It is an ensemble of high profile researchers or their works clustered by the way they are referred to in connection. They are linked in what can be described as schools of thought, communities of the mind, or networks of acquaintances that focus on similar questions. In newer, more technological terminology it can be defined as "cliques". Amez (2010) shows that in the field of cultural economics, movie economics appears as a separate interlinked

structure.

This literature overview uses bibliometric techniques to discover the main literature hubs, their importance and how they are linked. One of the basic idea's in bibliometric linking is that the co-occurrence of items in publications tells something about the extent to which the publications or items belong together in the knowledge domain. Items can be the entire indexed reference, any part of the reference, titles or keywords. When it comes to analysing a knowledge base and detect the core literature of a research speciality, it is common to use co-citation techniques. It consist of identifying pairs of co-occurring items in all references of the publications in your bibliographic database or your extracted publication set. The pairs of co-occurrent items can be listed in order of importance, however it is common to represent the items in a matrix structure. A co-occurrence matrix contains counts of the number of times two bibliographic entities of the same entity type are associated with some other entity type, Morris & Van der Veer Martens (2008). The co-citation matrix is a useful tool for data mining techniques such as clustering, methods that are suited to group items into different substructures, that way partitioning your specialty.

The publication set used for this literature study is selected from Thomson Reuters Web of Science. The search strategy was performed using the online Web of Science Social Science Citation Index Expanded data. The filter applied was based on six topic words: movie, film, cinema, motion picture, box office and Hollywood, applied on the fields of economics, business, business finance. The selection was limited to articles, letters, notes, reviews and proceedings papers. The exercise starts from the selected publication list of 329 publications in movie economics. Using Bibexcel, the references were extracted and further decomposed into cited items. They come in the form [SEDGWICK J, 1998, V35, P196, EXPLOR ECON HIST]. To perform the co-citation analysis, only the top 150 cited publications where considered. This list is the input to determine the co-occurrence of publications in the references of the 329 selected publications. Apart from statistical analysis of co-occurrence results, the matrix can be used as an input to visualize results in a way that the

most the influential works are highlighted and the linkage between the core works is manifested. It provides easy insight into the structure of the specialty. The visualized partitioning of movie economics topics is shown in figures 1.1 and 1.2. The bibliometric visualisation was performed using VOSviewer. The size of the circles is proportional to the total number of citations. The distance between items reflects the strength of the relationship in terms of reference co-occurence.

The analysis distinguished five clusters. The dominant one, coloured red, was labelled as *empirical movie performance studies*. The basic keywords of that cluster are success, prediction, reviews, critics, star, box-office, information, consumer. When looking at the publications that are at the forefront of the visualization, it is clear that it involves mainly studies aimed at predicting box-office success. The main factors under research are those providing consumers with information such as reviews or awards and marketing factors that are under control of the producer such as the cast. A second cluster features around the literature of Arthur De Vany and David Walls, dealing with matters of distribution characterizing the movie industry and was labelled as *nobody knows*. Partition three, referring to words such as industry, sharing, contracts, prices, release and booking, points to an *organizational perspective* of movie economics. The green cluster touches the supply and demand side with *historical studies of the movie industry*. A last group, standing slightly on its own, is a cluster treating trade, import and export movements of movies.

The partitioning largely agrees with the movie economics overview paper published in the Journal of Economic Survey, McKenzie (2012). Here, a micro- and macro-economics structure was maintained. The micro part treating, under the scope of demand uncertainty, the role of stars, critics, review, ratings, awards and genre. The supply part covers a production, distribution and exhibition angle. The macro economic section includes aggregate demand, trade of motion pictures, industry structure and copyright issues. As will be explained however, in absence of micro level data, often aggregate indicators are used as supply or demand proxies, that way shading the lines between micro and macro studies, borders that seem to disappear

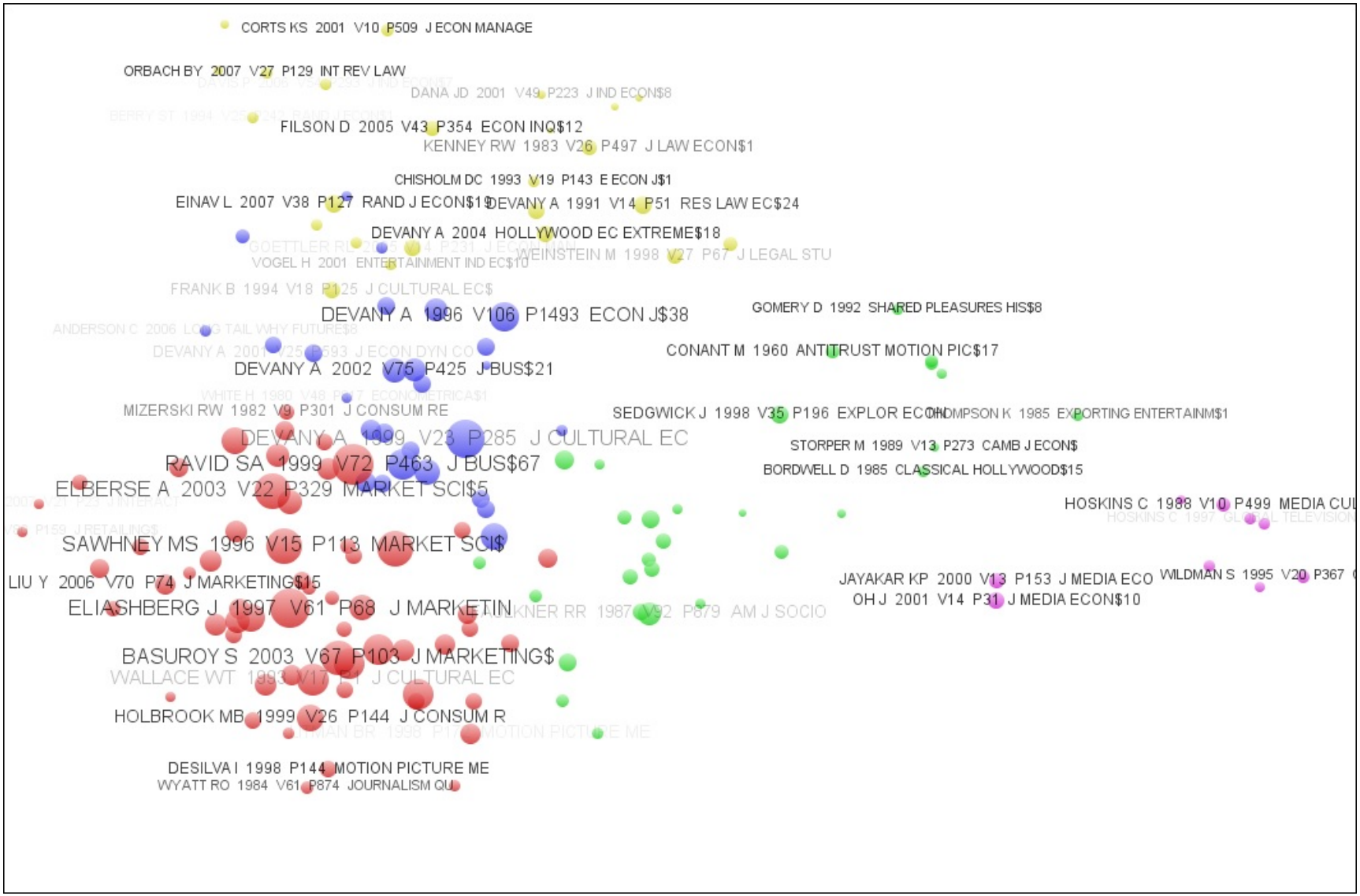

Figure 1.1: Science Map of 150 highest cited publications in movie economics with full reference information

The size of the circle is proportional to the citation importance, the colours represents the partitions. Red=Empirical movie performance studies, Blue=De Vany Walls Nobody Knows, Yellow=Industrial Organisation, Green=Historical studies of the movie industry, Purple=Trade in Movies

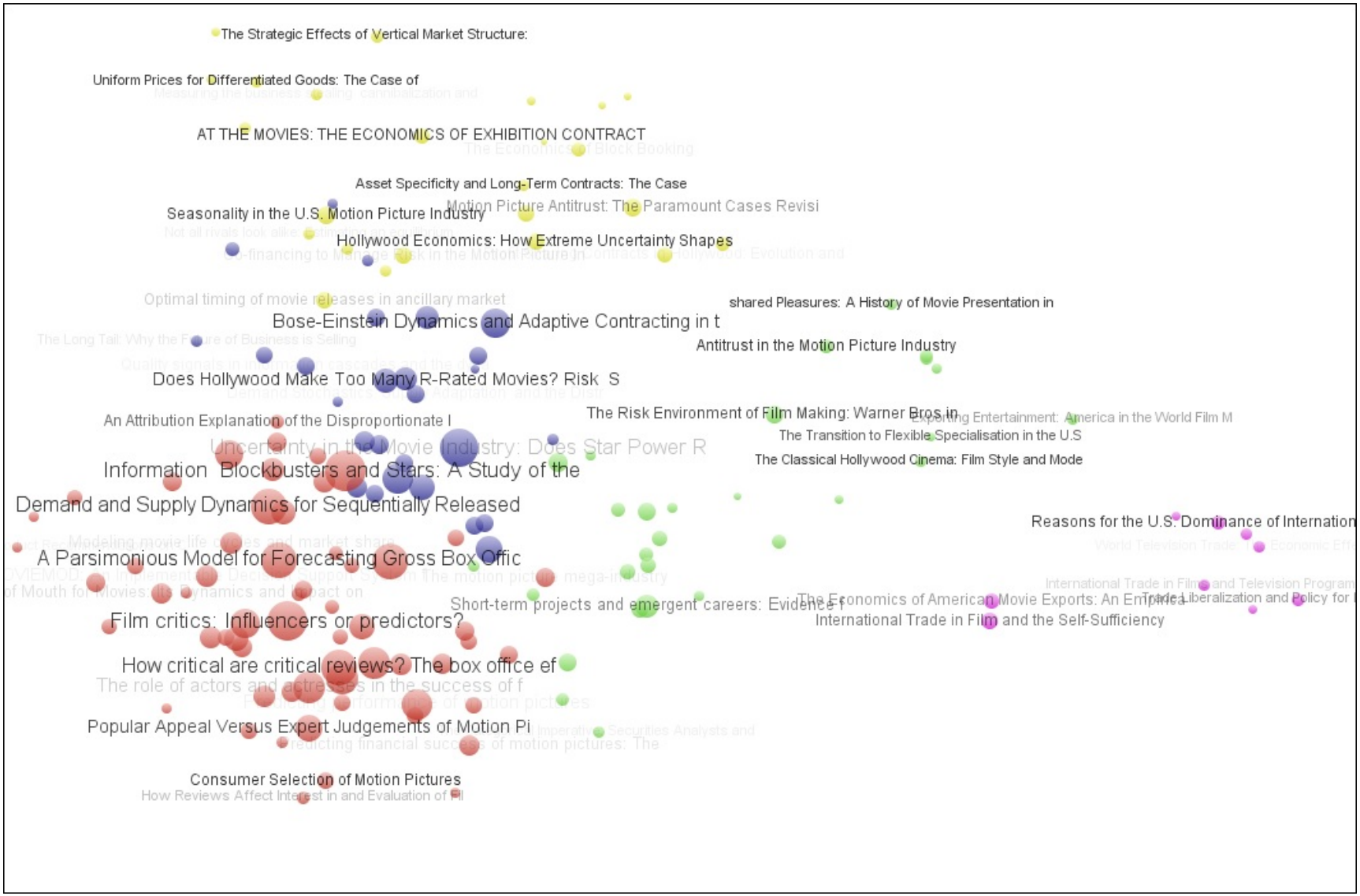

Figure 1.2: Science Map of 150 highest cited publications in movie economics with title information

The size of the circle is proportional to the citation importance, the colours represents the partitions. Red=Empirical movie performance studies, Blue=De Vany Walls Nobody Knows, Yellow=Industrial Organisation, Green=Historical studies of the movie industry, Purple=Trade in Movies

in a co-citation structure. The first two clusters, gathering the literature on movie performance and the models by De Vany are most directly relevant to this thesis. However, the other clusters touch decision elements such as seasonality, pricing and historicity, and are therefore important elements following on in a comprehensive overview. The next sections will dig deeper in the content of the various clusters. Following from this, the main conclusions will be summarized within a framework of movie choice decisions. This will lead to the re-formulation of a number of research questions and a path to deal, at least partly, with a number of shortcomings.

Performance studies dominate the literature of movie economics as a research specialty. The article by Litman (1983) was one of the first comprehensive empirical writings on movie success. The explanatory variables include genre, rating, star, academy award and release date. Those variables do set the standard for many studies that follow. The cluster density at 1.2 is intense around movie performance papers. Of the ten most cited papers, all but one belong to that cluster. The core papers, as well as their offsprings, are quite stereotypical. They consist of a regression type analysis relating a performance measure to a set of potentially explanatory factors. The main performance measures of success or failure are box office revenue or attendance, more than profit on investment or rent. Generally, this type of studies does not bear largely on a body of supporting theories or models. The explanatory dimensions can be grossly classified into film factors, including production, distribution and informational variables. Production factors are generally seen as movie attributes, items that filmmakers decide upon such as the story, cast, genre, and sequel. Not all of production factors are controllable however. One counterexample here is nationality of the movie. A second category consist of distribution and exhibition factors such as advertising, screen coverage decisions and timing, Hennig-Thurau et al. (2007). These show a cross-over with the industrial organisation (IO) literature. A third group are consumer information or certification factors that go beyond supplier control such as ratings, professional critics, peer review and awards. The latter group also encompasses information spread by consumers themselves defined as word of mouth.

The publication with the highest number of occurrences in the major cluster and overall is Ravid (1999). The article mainly investigates the effect of the presence of stars on movie performance. A star is identified by the fact an Academy award was won by that actor or that the actor participated in a top-ten grossing movie during the ten years preceding the release of the movie. The study bears on some other highly cited publications, such as Sawhney & Eliashberg (1996) and the influential empirical studies by Litman (1983) and Litman & Kohl (1989) on predicting the success of theatrical movies. Ravid (1999) borrows from the mentioned articles the control variables but at the same time challenges those studies on two levels. First, the presence of star power, although not new in empirical movie research, is now encompassed into a theory of quality signalling and secondly, the nature of both dependent and explanatory variables are questioned. The presence of stars is explained as a quality statement, not so much towards the consumer, as is mostly the case in models of quality uncertainty. Rather it is a signalling device regarding the quality of a potential movie project from the executive towards the studio or to outside financiers. This hypothesis is tested against the "rent capture hypothesis" stating that the value of stars is reflected by the market. To do so, return on investment is opted for as dependent variable.

The presence of stars is intuitively strongly connected to the movie business and researchers have kept a vast interest in what Vogel (1998) called "the bankability" of stars. Besides the article of Ravid (1999), the issue is also covered in the prominent article by Wallace et al. (1993) and the variable is included in the majority of performance studies. Despite the high interest in the role of stars as an investigative factor, the positive effect on movie performance is far from empirically conclusive. Nor the founding articles by Litman (1983), Litman & Kohl (1989) or Smith & Smith (1986), nor the top cited article by Ravid (1999) are able to establish a significant positive impact of the presence of stars on box-office performance. However, equally influential papers such as Wallace et al. (1993), Prag & Casavant (1994), Sawhney & Eliashberg (1996), Neelamegham & Chintagunta (1999), Basuroy et al. (2003) and Elberse & Eliashberg (2003) reveal positive effects on opening -, weekly - as well as cumulative revenues. Differences in results partly stem from methodological

differences. Simonton (2009) reveals how the studies differ in the samples that are used, the choice of variables included and the way star power is represented. Star influence in particular is measured in a variety of ways: took part in high performing movies, Litman (1983), won a best acting award, Basuroy et al. (2003), Delmestri et al. (2005), or ranking high in consumer surveys/industry magazine list, Sawhney & Eliashberg (1996), Elberse & Eliashberg (2003). In more recent literature, the way the star index is shaped has been challenged by the use of new media, with star phenomenon modelled using star buzz, reflected by the intensity of internet searches, Karniouchina (2011).

A second dominant keyword in the cluster, besides star power, is critics or critical review. The most influential paper here is *"Film Critics: influencers or predictors"* by Eliashberg & Shugan (1997). Studying the nature of this relationship was not new. Critics are seen as privileged watchers who can express their views through divers media. King (2007) locates the force of critics reviews at three levels 1. The omnipresence of opinions in popular press creates positive or negative buzz in the opening week 2. They send explicit recommendation like do consumer reports 3. Those reports are considered objective because too much bias would be detected. Although a vast percentage of viewers read reviews before attending, it is not an established fact that they express mainstream preferences. It is therefore a re-occurring question how strong their influence on movie performance actually is. Critics opinions might be overpowered by promotion activities where critics views can be considered elitist and not reflecting general taste. The earlier writing by Litman (1983) also included critics reviews as an explanatory variable as does the study by Prag & Casavant (1994). The article by Eliasberg and Shugan specifically focus on this particular relationship. First, they contribute to the study of the subject by looking at the effect of critics' reviews at different points of the movie life cycle. Second the authors experiment with two visions on the relation between movie reviews and movie success, namely that of critics as influencers and that of critics as predictors which they oppose as two verifiable alternatives. The critic can be seen as an opinion leader whose views are taken as objective and therefore influence the decision of consumers to attend a movie or not. Being experienced watchers,

their opinion is taken as knowledgeable and their views are taken over by others. Although the conceptual frame of the movie as an informational good is dismissed at the beginning of their article, the idea of a critic as an information provider is very much into their way of thinking. The second view is that of a critic as mere a predictor. Potential consumers are as such not influenced by their opinions, but the views of critics can serve as leading indicators signalling the path the movie's success will follow. The empirical results provided in the article show little proof of the influencer view. The relationship between critical reviews and box office revenues is not significant during the first four weeks. They do establish a positive relation between the critics' view and total accumulated revenue, which they motivate as an underpinning of the predictor view. That is put into question in a highly cited paper by Basuroy et al. (2003). They establish that, considered over an eight weeks period, positive as well as negative reviews affect weekly box office results. Empirical results reveal how the impact of negative reviews is stronger, which the authors attribute to negative bias in impression formation. This however holds only the first weeks after the release as the effect is mitigated by studios' systematic leverage of positive reviews. The article is also important in studying the interrelationship between star power and movie budgets on the one side and reviews on the other. It questions why, since the effect of both star power and budget is proven ambiguous, there is still large investment in both. They conclude that both variables indeed seem to show little effect when reviews are positive, however, the opposite is true when negative criticism is expressed.

Like the effect star power, the role of critics remains ambiguous. The paper of Basuroy et al. (2003) is important in showing how not one variable in particular affects box-office returns, but how the interaction between variables can play a role. This point was also brought to attention by Reinstein & Snyder (2005) who state that the effect of critics on performance cannot be analysed discarding correction for the underlying quality of the movie. They compare a difference in difference approach to including reviews as an explanatory variable in a global statistical specification, McKenzie (2012). Hennig-Thurau et al. (2007), confronted with the lack of homogeneity in studies, expands this idea and looks how the diverse items driving perfor-

mance are interrelated. The majority of preceding studies, with Elberse & Eliashberg (2003) as exception, were performed using regression techniques, assuming that the underlying factors are statistically independent. The study of Hennig-Thurau et al. (2007) acknowledges potential autocorrelation which can exist between the explanatory factors and separates immediate from mediated effects. The empirical results obtained do confirm the lack of relationship between star power and movie success. The indirect effect is negative even. It stems from the negative effect of star power on quality. The latter can be due to disconfirmation effects - Consumers set their expectations higher and are more likely to be disappointed-. Star power does not seem to positively influence critics review. Critics opinions have little impact on short nor long term box-office performance, but in agreement with the results of Reinstein & Snyder (2005), it correlates with consumer quality perception. Through that, they indirectly affect in the longer run. This result contradicts with the study of Eliashberg & Shugan (1997) stating that critics only have predictive power. However, it is not established that the indirect relation is of a causal nature. More recent publications in the cluster include web information to investigate opinions of experts as well as those of consumers. Online reviews allow to monitor consumer responses in real time. The internet serves as an open forum where consumers reveal their opinions on movies they came to attend as well as their expectations on movies to come. It has become one of the most important ways through which information spreads from consumer to consumer, from critic to consumer and from producer to consumer. Sites such as the Internet Movie Database (IMDB), Rotten Tomatoes (p. 226) or Yahoo aggregate opinions and introduce their aggregate representations into online scores. One of the papers represented in the 150 top cited papers, written by Dellarocas et al. (2007) emphasizes that the presence of online information available to producers on a continuous basis offers material to improve forecasting models. Given the large uncertainty that characterizes the movie industry, the search for adequate forecasting models in support of decision making has always been a motivation for this type of research. As is clear from looking at the cluster, besides terms such as star and critic, the word "prediction" is very much at the forefront. An important example of a decision support model is Moviemod by Eliashberg et al. (2000), which is a pre-release model. It allows to estimate the effects of controlling

extra advertising, extra magazine articles, extra TV commercials and higher trailer intensity on box office revenues. Also earlier published and highly cited papers such as Litman (1983) and Zufryden (1996) serve as predictive pre-release models. Others, like Sawhney & Eliashberg (1996), Neelamegham & Chintagunta (1999) focus on later-week revenues which usually generates better results because they can be supported by more updated explanatory variables. The model of Dellarocas et al. (2007) classifies as an early post release revenue forecasting model. The authors include variables of volume, valence and dispersion of online conversations, proxied by the number of posted reviews, the average ratings posted and the entropy of age and gender. It is shown that adding online metrics to other factors such as pre-release marketing, theater availability and professional critic reviews significantly improves forecasting precision. The volume of online reviews serves as a leading indicator for early sales and can be used before sales reports are out.

The increased use of online data matches a new stream in empirical research. Finding its origin in computer sciences as much as in economics, the process of analysis is often accompanied by the use of other statistical techniques more apt to deal with the big data streams. It contrasts with the dominantly OLS type estimates that characterize the eighties and nineties literature. The model of Dellarocas et al. (2007) uses a bass diffusion model where the internal and external factors are bass diffusion parameters. The internal forces are endogenous and come dynamically from the past observations which are related to word of mouth, such as valence, the spread of user reviews and MPAA rating. The latter is seen as steering the word of mouth process. To estimate this, they use a two-level hierarchical Bayesian estimation model that includes the influence of past on future through a process of conditional updating.

Word of mouth (WOM) refers to all informal communication between parties concerning the evaluation of goods and services before making an opinion, Anderson (1998), Westbrook (1987). It can take the form of buzz, contribution to forums or conversation among peer groups. The web 2.0. has created extended opportunities

for users to post and exchange information. One of the most recent publications on that subject represented in the cluster is the article by Liu (2006). Because every movie product is new and awareness has to be built, the importance of word of mouth has always been acknowledged as an influencer of consumer decisions. The author offers a double motivation for that. First, movies categorize as popular cultural goods and receive great public attention. Therefore active communication around them is to be expected. Secondly, the nature of movies as an intangible experience good creates uncertainty about its quality before the product is actually viewed. Consumers therefore rely on gathering all sorts of ex ante information and experiences by others who saw the movie before and made their opinion public. WOM comes from other moviegoers and may therefore be perceived as more trustworthy than advertising, Faber & Oguinn (1984). At the same time, it might be a better reflection of popular taste than critical reviews. The study of Liu relates to the study of Dellarocas et al. (2007) in the observation that both articles use the same types of data and measures. The study of Liu (2006) uses criteria of volume and valence of 12.000 posted messages from the Yahoo Movie Message board. Volume is expected to have a positive influence on consumer awareness, the more messages, the more likely the consumer will hear from it. Conversely, valence is expected to be more affective. One of the interesting features shown in the study is that, depicted dynamically, volume and valence do not necessarily follow the same pattern. Although the study is innovative in many ways, methodologically it follows the patterns of the older and previously mentioned studies in the linear way relationships are modelled. This is also the case for the choice of the control variables which consist of critical reviews and the number of screens. The model is intertemporal and it is used to forecast box office revenues for the opening weeks as well as the seven weeks following. Like the model of Dellarocas et al. (2007), it shows that online WOM indicators improve prediction accuracy. The value added to the forecasting model however comes mainly from volume rather than from valence. It thus influences awareness more than attitude. A second observation is that there are very few antecedents to word of mouth. Neither star power nor critical reviews seem to affect volume and valence of the posted messages. Word of mouth indicators themselves prove endogenous in the sense that a week of active word of mouth is

followed by another active week. Finally the paper confirms the hypothesis that the valence of pre-release word of mouth is positive and on average higher than in opening weekends. The author claims this is an illustration of the confirmation or disconfirmation theory of expectations. Intensive advertising might generate high awareness for the movie, attracting people who would initially not go and see a film. That group of consumers might, faced with a lower than expected experience, spread negative opinions afterwards.

The theory of movies as an experience good underpins the paper by Liu (2006) and this holds for the majority of the other literature that composes this cluster. The idea of movies as a product characterized by uncertainty where consumers are looking for ex ante information is manifestly present and explains how factors such as critics revue and stardom come into play. They signal quality or inform people on what to expect. Despite the fact that consumer theory of experience and information goods is latent, it is striking to observe that looking at the most cited papers in the movie literature, there is hardly any theoretical literature on the decision making process of consumers towards cultural goods, on the nature of movies as a product or on the nature of the experience. Some exceptions are located at the bottom of the cluster. It involves the papers by DeSilva (1998) and Holbrook (1999), the latter inspired by the sociological writings on how members of different social classes will dispose of different preferences in artistic objects. The differences find their origin in economic and cultural capital as the product of education and learning. The theory is empirically tested by relating movie features such as genre, objectionability, origin and starpower to differences in popular versus expert appeal. Besides writings on popular taste, Holbrook (1999), (2005), Holbrook & Addis (2007), there is the highly referenced paper *"The experiential aspects of consumption: consumer fantasies feeling and fun"*. Although this paper is not on movie consumption in particular, it was highly influential by offering a framework to reflect on consumer products where so called multi-sensory psychophysical relationships steer consumer behaviour. The experiential view on consumer relationships offers counterarguments to the informational process theory on consumer behaviour but actually deepens it in many respects. The paper will be discussed in more detail in the next chapter.

The idea of movies as a good endowed with uncertainty was also at the heart of the *"nobody knows"* views, expressing how uncertainty rules the film industry. The second cluster manifested in the mapping exercise is mainly composed around the articles by Arthur De Vany and David Walls (DW), some of them co-authored by Cassey Lee or Ross Eckert. It is as such not the largest cluster nor the most dominant, but it is important for a number of reasons. First because Arthur De Vany and David Walls are amongst the highest cited authors in the literature on movie economics. Secondly, because their articles do offer a theoretical framework that reflects on some important characteristics of the movie business. The core document in the cluster is the article titled *"Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry"* published in 1996 in the Economic Journal. This article was also integrated in the handbook *"Hollywood Economics"*, published in 2005, along with other core articles of the authors. Central to the writings by DW are the insights that the time cycle of box office revenues is highly demand driven and that the nature of movie industry contracting is organized to follow the pattern of demand. The definition of a movie as experience good is at the very core of the analysis. Before a movie is attended, there is uncertainty about what to expect and the quality of the product is only exposed through experience. The thesis of DW is that the dominant stream of information that steers the demand pattern is that coming from previous viewers whose opinions and evaluations can be consulted prior to one's own decision process. This may come under the form of reviews, expert reports, friends' conversations or from online forums in which movies are widely discussed. That way, information is transmitted from consumer to consumer and global demand develops dynamically over time.

The literature is specific in that the ideas of information dynamics are formally modelled through a Bayesian decision process where the consumer has prior uncertainty about the quality of the movies which can only be fully resolved after experience. The uncertainty relies on the individual's type and on the information set available to the consumer. That information set is composed out of the quality signals revealed by previous attendants. The model is inspired by Jovanovic (1987) and the specifications of information cascades presented in the paper by Bikhchandani et al.

(1992). This conditional choice logic is introduced in a setting where consumers are faced with a portfolio choice over a number of movies. Starting with a Dirichlet prior, the demand dynamics, expressed in terms of the probability that a consumer selects a particular movie given previous trials, is demonstrated to be drawn from a Bose-Einstein distribution. The aggregated choice reflects the fact that customers sequentially select movies and the probability of a given choice is proportional to the fraction of all the previous moviegoers having made that choice before. This path dependence, linked through information feedback, makes that small differences in movie attendance at the beginning of a movie run can evolve in a very large spread in movie success at the end. At the same time, a movie starting broad with high expectation can, through a process of negative information diffusion, follow a path of failure. To test the hypothesis of information cascades, DW compare with potential alternatives. One alternative is the possibility that revenues are characterized by power laws. Box office revenues can easily be ranked and when a distribution is manifestly skewed, in many cases it obeys a Pareto distribution which is a power law probability distribution. Power laws are the manifestation of a model in which the growth rate is independent of size (Gibrat's law). Ijiri & Simon (1977) stated, using US firm size data, that deviation from the Pareto distribution might suggest the presence of an underlying autocorrelated process. Using box-office data of the Variety's Top-50, the authors establish that this is indeed the case. The hypothesis of a linear Pareto distribution is rejected in favour of downward concavity and along with that, the presence of autocorrelated growth in motion pictures revenues is established. The latter is claimed to be consistent with the view of increasing returns to motion pictures caused by information feedback.

The importance of the contribution of DW plays at different levels. First they provide a profound insight in what they describe as "the way individuals process and exchange information leads to complicated dynamics that create extreme differences among movie picture revenues". Secondly, their research focuses strongly on the particular shape of the movie earnings distribution as well as on the empirical verification of the evolutionary dynamics of rank and revenue. Thirdly, like is the case in performance type literature, one questions how "traditional influential factors" such

as stars and genre affect this model. Finally the framework is largely entangled with issues that are at stake in supply side movie economics such as contracting schedules. The uncertainty in the movie business is inherently present in the majority of movie literature. Attempts to discriminate the driving forces that explain and forecast movie revenues can be placed in the light of an uncertain business seeking for predictability. The writings of DW can be considered as attempts to catch and translate this uncertainty in terms of statistical distributions.

Using both descriptive analysis and Gini coefficients, they demonstrate the high degree of inequality in box-office revenues. About 20 percent of the movies stand for about 80 percent of the earnings. At the same time, the authors point at large differences between statistical means and medians as well as at the presence of very high standard deviations. This all adds to the observation that movie revenues are characterized by highly skewed and asymmetrical statistical distributions with a thick tail to the right. Those observations are confrontational for many movie studies implicitly assuming an underlying Gaussian distribution. The findings are re-established in multiple papers by their hand such as De Vany & Walls (1997), (2002), and (2004). The last two articles describe the features of the statistical distributions which they show to be stable Pareto against the alternative hypothesis of normality. The stable Pareto distribution was proposed by Mandelbrot (1963*b*), (1963*a*), (1997) and by Fama (1963), (1965) to analyse returns to financial assets. Krishnan (1998) showed that there is a class of distribution functions which follow the Pareto law asymptotically and are still consistent with the Bose-Einstein information updating process. This can be empirically verified by testing behaviour in the upper tail of the distribution, which seems to be confirmed on movie data. DW use the asymptotic Pareto property in estimating models which are restricted to revenues above a certain limit value. A particular property of the Pareto distribution is that in the upper tail, the conditional expectation is proportional to the current realization, which allows analysing dynamic behaviour. It implies that at each week of a film's run, its expected future revenue is a portion of the revenue already earned. This can be translated into a testable model of sequential weekly revenues. The authors demonstrate that the immediate past is the best predictor for revenues that will

be earned the following weeks, whereas opening and early weeks appear as weak predictors for subsequent revenues, for hits as well as not-hits. Moreover, it is empirically established that around the fourth week, a change in the mapping occurs, pointing to a potential bifurcation taking place. At that point, hit movies separate from others at increasing rates. The estimated dynamic coefficient can be translated back into Paretian tail weight coefficients. Here it is shown again that hit movies are characterized by increasingly heavier tails the longer the movie runs, while the opposite is true for non successful movies. To reach this, the movie has to run long enough to get at the upper tail point where the conditional expectation of box-office revenue is linear in the past. The results are therefore considered as an additional confirmation of the Bose-Einstein dynamics process. The endogeneity observed is not a manifestation of statistical herding. At a certain point in time, hit movies disperse from non-hit movies to achieve a peak conditional expectancy. This point is interpreted by DW as the moment where there is plenty of quality information around to guide the filmgoer to good movies and away from bad ones. If herding drove the process, consumers would imitate under all times and peak conditional expectation would be observed from the first weeks on.

Most of the explanatory variables dominating the performance literature of cluster 1 are also investigated in the research of DW. Certainly the role of stars is manifestly present, being the key topic of three articles *"Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office"*, De Vany & Walls (1999), *"Motion Picture Profit, the Stable Paretian Hypothesis and the Curse of the Superstar"*, De Vany & Walls (2004) and *"Contracting with stars when nobody knows anything"*, De Vany (2005). However this time, the explanatory variables are not captured in a linear statistical Gaussian model, but placed against the stochastic dynamics of their underlying framework. In line with the Pareto rank law models, empirical results show that the slope is flatter among star movies compared to non-star movies. Superstar movies show a higher probability mass in the upper tail in a way that the probability of positive earnings is larger compared to films that do not cast superstars. This effect however disappears at the level of individual stars and is softened by observing that star movies also have larger budgets, wider releases or

better scripts. Conclusions are similar when expressed in terms of profits, although the average profit in both cases is negative. The skewnedness of the distribution is also used for the explanation of the so-called curse of the superstar. It makes that there is a huge difference between the average and the expected profit. If expectations are formed to determine the amount paid to a superstar, the likely outcome is that of a severe loss.

Some of the modelling put forward by DW includes other control variables such as budget, screens, sequels and rating. Budgets and opening screens have their main influence in the lower quantiles and their significance diminishes when moving to the upper quantiles of the revenue distribution. DW state that huge budgets or launching a film by booking a large number of screens helps to place a floor under the box-office revenues. Sequels are, under this framework, proven to be the safest movies. The fact that the movie is a sequel pushes it up the top decile of the revenues. The rating issue is addressed in the article *"Does Hollywood make too many R-rated movies"*. The authors address the question posed in the title in methodologically the same way as described in previous paragraphs. They show that the mean revenues of G-, PG-rated movies dominate the mean revenue of R-rated movies. At the same time, G- and PG-rated distributions are more skewed to the right, so they have a higher probability of extremely high revenue outcomes. The ratio of expected over most probable value is therefore lower for R-rated movies, which makes the authors conclude that the portfolio is indeed overvalued in that category. Ravid & Basuroy (2004), in a later writing, place the R-Rated movie puzzle in a decision theoretical framework by executives in charge of movie projects. The authors show that certain subsets of R-Rates movies combine less riskiness with higher revenue and can therefore be seen as rational choices in terms of managerial decision.

The remaining literature in the cluster can be categorized into either papers that were an inspiration to DW articles or more recent papers that were inspired by them. The first group involves papers like Ijiri & Simon (1977), a core reference on the issue of the deviation from the Pareto distribution, the article on information cascades

by Bikhchandani et al. (1992), and Chung & Cox (1994) early paper on how the superstar phenomenon can be explained as a stochastic process. The category of papers influenced by the framework depicted by DW is generally more recent. The ideas are applied on national level data. Hand (2001) and Collins et al. (2002) re-establish elements on UK data. They confirm the unbounded variance and the heavy tail property and dismiss standard OLS models as proper tools for the estimation of movie prediction models. To perform the estimation, they transform the data from continuous to binary, applying a threshold level. Employing a probit like model, the probability of a hit is related to an array of film characteristics. Like DW they confirm the influence of stars to be positive though highly uncertain. The influence of reviews is smaller, but also less uncertain, while the genre characteristics are shown to be of little influence. Another study by Bagella & Becchetti (1999) investigates movie box-office performance in Italy over a period of ten years. They rely on Generalised Method of Moments (GMM) to deal with the observed non-normality. Successes in movie performance are shown to be impacted by cast choice and the popularity of the director. Noteworthy in their result is that being a subsidized movie seems to generate impact through the cast factor. Two other studies by Nelson et al. (2001) and Deuchert et al. (2005) combine the input of DW writings with insights of Rosen (1981) *"economics of superstars"* article to validate the effect of Oscars. Rosen (1981) states that small differences in quality perception and willingness to pay can translate into large differences in income. Results of both the study of Nelson et al. (2001) and Deuchert et al. (2005) show that the effect of both nominated and awarded Oscars is positive and significant on the movie's survival rate, on the return per screen and box-office revenues respectively, but only so for best actor/actress and the best picture categories. Deuchert et al. (2005) also observe an influence of the opening box-office on the performance of the weeks following, but the impact is too weak to speak of a snowball effect. The authors interpret this as the absence of word-of-mouth processes. The minor effects of word-of-mouth are contradicted by a more recent study of Moul (2007). The author complies with the insights of DW. However, where the latter claim that the presence of word-of mouth effects follow directly from the rejection of the Pareto law, Moul (2007) states that this is only true when no other explanatory variables are in play. The author frames the estimates in a demand

model including also a rude form of consumer heterogeneity. To do so, he uses a nested logit model along three dimensions: substitution between movies and other goods, between action and non-action movies and between family and non-family movies. The word-of-mouth effect is considered as an unobservable effect which exhibits as a residual and the empirical way to pin down WOM is to disentangle the serial correlation that stems from weekly and idiosyncratic disturbances. Results indicate that about 10 percent of consumer expectation can be attributed to dynamic information effects. Their heteroscedastic plots show a turn at around the fourth week indicating that information is dispersed quickly. By the fourth week, they estimate the WOM share to explain half of the unobserved, given a minimal number of people having watched the movie before.

Although the DW framework offers a demand driven model to explain box-office behaviour, their work is very much written from the perspective of movie supply side decision makers having to act in response to the information sequentially offered. The movie industry is information driven in its dependence on temporary reporting and this translates into business practices and types of contracting. One of their key contributions is revealing the mechanism of how supply dynamically interacts with unfolding consumer behaviour. Contracting in the movie sector reflects this pattern in the sense that it allows to adjust in a flexible way to changes in demand. After a film is produced, theatres exhibit the movie by renting it from the distributors at a percentage of the box-office revenue, a rate which declines over the run, but goes up to 90 percent at the beginning. When a movie shows unsuccessful, exhibitors can drop out, but generally have to respect a minimal number weeks. In case a movie is successful, there are two ways in which the distributor can adapt supply to demand, either by extending the theatres already booked or by contracting to new exhibitors. The latter minimizes risk in the sense that distributors can wait and see how the movie catches on. The cost of augmenting supply is small and is limited only by theatres' capacity and prints. In that sense, the movie supply is considered largely elastic and can react in an almost infinite way to demand. That, by its own force can strengthen the word of mouth effect. Apart from employing contingency type contracting, distributors can be influential through advertising and

launching strategies. This is based on a priori appraisal of demand, which given that nobody knows how a movie will catch on, is endowed with large uncertainty. The business can opt for a wide release, with the risk that bad word of mouth makes the movie disappear quickly and that high losses are incurred. If the movie generates a positive information stream, it can result in very high revenues. A more tailored release allows the control of risk, but at the expense of a slower spread for word of mouth, De Vany & Walls (1996).

The industrial organization related reference literature presented in the third cluster is overall more recent, ranging from the article by Kenney & Klein (1983) on block booking to an article by Einav published in 2007 on the seasonality of the US motion picture industry. Surprisingly, the highly linked works are also the most recent ones, but no article in this cluster is listed in the top 20 of highest cited records. The publication by Einav (2007) has most links. As shown in figure 1.2, the subject of seasonality stands close to the performance literature. In empirical literature, the phenomenon of seasonality is often seen as a noise factor. This however is dismissed by Einav in two papers published in 2003 and 2007. Empirical evidence shows that distributors release big budget movies mainly at the beginning of what they consider high demand moments, being the start of the summer holidays and the Christmas period. Smaller budget films are introduced outside those months. Einav (2007) estimates a nested demand model. Agents have to choose between a movie and an outside good which accounts for the opportunity cost. The gain in utility of going to the movies relative to the alternative is dependent on quality, an underlying seasonality effect and the number of weeks that have passed since the movie release. Quality is proxied by items that do not change over the course run such as cast and genre. The effect of WOM in the error term, such as brought forward by Moul (2007), is acknowledged but ignored in the analysis. The study is based on a broad dataset of releases over a period of 15 years. The idea is to benchmark the same calendar weeks, that way assuming that the consumer mix remains similar between observation periods and that the seasonal pattern is stable. The author demonstrates that, once the quality of the movie is accounted for, the seasonality effect is strongly dampened or vice versa that the quality effects amplifies the seasonality

effect by about 50 percent. The intertemporal decay in revenues is calculated to equal 3 percent. Three earlier papers in the cluster also treat seasonality effects, a preliminary paper by Einav (2003), a writing by Corts (2001) and an article by Frank (1994), the latter on the subject of optimal timing of video releases. The preliminary paper by Einav (2003) was later published under the same title in 2010 in Economic Inquiry. It provides an industrial organizational game theoretical model to determine optimal release times. Seasonality is studied as a sequential move game with private asymmetric information. The solution is a probability distribution over all possible outcomes that can be translated into a maximal likelihood estimation model. The season is taken as given, the timing within the season is the strategic decision variable. Distributors generally want to avoid placing the movie to close in time to a similar movie and the movie release moment is often changed. About 60 percent change the release date at least once, with high quality movies shifting less. The fact that the switches are small makes the author conclude that the season is targeted beforehand and that the exact date is fine tuned. As the potential number of shifts is limited, the decision process can be described as a sequential discrete choice game where distributors choose among a small number of potential entry weekends. For the estimation, Einav (2010) reuses the nested demand model suggested in his 2007 paper. Those demand patterns are employed to test the optimality of movie producers' behaviour. The results show that, conditional on the assumption that the demand model is adequate, movie distributors overweight seasonality and do release too many top movies at the same moments in time. Improvements can be made by releasing more evenly. Earlier research performed by Corts (2001) already established the imperfect releasing strategy, but relates it to the degree of movie structure integration. Frank (1994) is one of the earliest papers on that topic in the cluster. It provides a theoretical model that calculates the optimal period between movie release and video release in terms of minimization of opportunity costs.

The IO cluster also contains two articles by Arthur De Vany. It was mentioned before that cluster 2 offered a demand orientated explanation of the phenomena in the movie business, but at the same time, that theories were written with the aim to provide feedback to the supply side. In that sense, one could state that the

literature of DW bridges the demand references of partition 1 with the industrial organisation (IO) literature. The IO cluster contains an article by De Vany and Walls *"Hollywood Economics, how extreme uncertainty shapes the movie business"*, and a second titled *"The motion picture antitrust, the Paramount cases revisited"*, authored by De Vany and Eckert. The latter is important reading to understand some of the other surrounding papers. It treats in more detail the Paramount antitrust cases, consent decree and the impact of the Sherman act. Court decision in the 1940s and 1950s stated that the movie industry was characterized by a too high degree of vertical integration and that the complexity of practices was at odds with free competition. It did decree the separation of production and distribution from exhibition activities. At the same time, they ordered that pictures had to be licensed individually at the request of exhibitors, which did put an end to practices of block and blind booking. Block booking consists of the selling of licenses in block for an entire season. It was a form of forward booking and blind in the sense that yet unmade movies were contracted. The court also stated that, through separate contracts between the distributor and the exhibitor, a more competitive price structure would be installed. De Vany & Eckert (1991) argue that the former practices of the industry were reasonable ways for dealing with the particularities of the product such as its uniqueness, produced at high suck cost and offered under demand uncertain circumstances. Therefore time is needed to gradually build an audience. The Paramount decisions, targeting disintegration, while meant to improve efficiency resulted in lower output and higher prices. The vertically integrated structure, as it existed before, was a way to deal with the high risk typifying the movie business. Once distribution was prevented from contracting forward and assuring income in advance, they no longer contracted backward. The number of creative people under contract declined as did the number major studio productions. The emphasis shifted to competition for first run features and real admission prices rose, partly because the amount of late-runs at reduced prices diminished. De Vany & Eckert (1991) conclude that, because so little was known about the industry at the time, the Paramount decisions failed their objectives.

Other authors in the cluster also treat the consequences of the Paramount court

cases. Hanssen (2000) elaborates on the issue of block-booking. In line with De Vany & Eckert (1991), the author disagrees on the logic of the Paramount decisions. He starts by rejecting an earlier explanation by Kenney & Klein (1983) for the practices of block booking. They claim that block-booking prevents oversearching on behalf of the exhibitors who might reject movies that appear of lower value ex post compared to the ex ante value deal. Block booking prevented exhibitors of post-contractual rejection of unsuccessful movies. Hanssen (2000) illustrates that contracts where not all that rigid and the motive of block booking by the producer was to get the movies to the exhibitors in the right quantities at the lowest possible cost. The oversearching is as such not related to the block-booking and up until now quite a degree of early substitution of bad performing films takes place. Given the profit sharing nature of contracts, this is in the advantage of both parties. Moreover, nowadays movies are released on a large scale and multiplexes work to a large extent with the same producers, negotiating the same number of movies. Films are accepted as they are released and the length of play is dependent on the demand evolutions. In the end, the author concludes that the current situation resembles pretty much the old block-booking system.

Orbach & Einav (2007) further explore the elements of the Paramount decision, starting from the observation that while the legislation was meant to increase competition, almost everywhere a uniform pricing policy is applied. Most theatres hold on to a uniform ticket price all days of the year, with exception of a limited amount of discounts, despite the fact that the common economic rationale for fixed prices do not hold in the case of movies. Consumer preferences in favour of specific exhibition times or product heterogeneity are hardly reflected in the price structure. Explanations are sought in 1. the care of the exhibitors for perceived fairness 2. the fact that lower prices might signal low quality and make the already unstable demand even more unstable 3. the observation that a fixed price is a way of the exhibitors to deal with ex ante uncertain demand 4. that varying prices may involve large menu and monitoring costs 5. agency problems that relate to the fact that varying prices are less observable by the distributors in the light of a revenue sharing contract 6. uniform prices mitigate the problem of double marginalization. None of the rea-

sons mentioned are retained by the authors as full and adequate reasons to explain the phenomenon of a fix price structure. They conclude that the interests of the distributors outweigh those of the exhibitors who the authors claim would benefit from price differentiation. In contrast, a recent article by Courty (2011) published in Economics letters and citing the paper by Orbach & Einav (2007) demonstrates, using a dual model of optimal product line strategy under fixed prices, that uniform pricing might be optimal when high quality movies are likely to be consumed by viewers in the lower quality segment. It may therefore be a consciously chosen strategy. The result is enforced under circumstances where sellers give an absolute priority to advertising tools to signal the quality of a good. Sunada (2010) offers an additional perspective to the matter by focusing on the Japanese market and observes that the Paramount case had not such an effect because vertical integration was not prohibited. Major distributors can therefore still operate their own theatres. Their empirical results show that prices are higher when the level of integration is higher. However, this seem to be mainly related to the fact that they also offer higher quality. They are better liked by consumers and have higher attendance rates. The author concludes from this result that vertical integration seems to be rather in the benefit of the consumers.

Central in discussing current movie practices and their rationale are the diverse ways of contracting, the specific financing structures and their relation to the risk characterizing the sector. Filson et al. (2005) treat the nature of current profit sharing movie exhibition contracts. They argue that, where in economic theory asymmetric information is often the underlying argument for profit sharing, this cannot be the key argument in the movie sector. Here the uncertainty is situated at the demand side and all parties are equally affected by it. For the movies, it is more the elements of risk sharing and overcoming measurement costs that are the determining factors for this type of practices. Based on a constant risk aversion model, the authors calculate that the optimal sharing contract has a diminishing slope conform the practice of minimum floor/sliding contracts. If the attendance is expected to fall, then the exhibitors must get an increasing share of the revenue over the run and ex post adjustments are justified when attendance is considerable

higher or lower than foreseen. A study on a large set of exhibition contracts provides support to this view. An earlier paper by Weinstein (1998) deals with profit sharing contracts between the studios and the actors. They remark that this relationship cannot longer be placed in a principal agent frame. Given that the sector is getting riskier, risk averse managers do shift part of that risk to the stars who are financially capable to bear part of it. Goettler & Leslie (2005) reduce the role of risk sharing in explaining the large amount of co-financing activities in the movie sector. About one out of three movies are financed by multiple partners. Although risk management is considered a reason for this, the paper demonstrates that there is hardly any justification for it. Several types of potential risk diversification are tested that co-financing could bring about. First, co-financing could be applied for extremely risky movie projects, secondly, co-financing can bring diversification into the studios' portfolio and last, risk can be reduced through the law of large numbers. Goettler & Leslie (2005) show however that the return on investment does not depend on it being co-financed, that the covariances between revenues of movies of different genres is insignificant and that studios majorly finance bigger productions which would be suboptimal under large number statistics. The reasons for co-financing are sought in the fact that studios, by cooperating, can better coordinate their release dates.

Like the IO related group, the historical cluster incorporates supply and demand side aspects. Part of the vertices are located closer to the performance cluster, others closer to the IO cluster. The considered group of references contains books as well as articles and a relatively high number of works comes from outside the movie literature. The core of the non-movie items is on the topic of cultural goods and creative industries and contains some classical works in that area such as the articles authored by Hirsch (1972) and Lampel et al. (2000), both conceptualizing cultural goods and the referential work by Caves (2000) *"Creative industries, contracts between art and commerce"*. Besides that, the cluster consists of three books providing a description of the movie industry from a historical perspective namely: Bordwell et al. (1985), *"The classical Hollywood cinema: Film style and mode of production to 1960"*, Thompson (1985), *"Exporting entertainment: America in the world film market 1907-1934"* and Gomery (1992), *"Shared pleasures: A history of movie pre-*

sentations in the United States". Joining the key references makes it possible to label the group. The majority of studies investigate phenomena taking place in the movie sector from a historical perspective. To a degree, they find their origin in the accessibility of some important data archives by major film studios, offering detailed information on the genre of the movies, costs and performance. Two major subjects can be identified. First the study of the impact of the industrial development on the choice of resources used in the movie sector. Second, a historical based analysis of the overall performances of movie studios that bears a lot of similarity with the rest of the literature, treating elements such as stardom, box-office performance and organizational structure. However, the angle of analysis is different since it relies more heavily on creative industry literature to ask what is specific about movies as a managerial project. Therefore, this group of literature deserves its place as an autonomous cluster. By framing the analysis of movies in relation to its creative goods characteristics and by placing them into a historical perspective, a broader insight is offered into the rationale behind certain managerial decisions and why observed phenomena such as increased stardom and typecasting grew out of an industry faced with changing circumstances.

Faulkner & Anderson (1987) analyse the production of each movie as a single project where output is unique, the environment is uncertain, complex and the coordination takes place through feedback. In such a context, career lines are successions of temporary projects and each contract is an opportunity for all actors to demonstrate talent and capability. Sustained participation is dependent on past successful performances, on credits and ties. It leads to persistence in the observed organizational network and a sharp separation between the winners and the periphery players. Also DeFillippi & Arthur (1998) sketch the split between a core and periphery group and the importance on how "project participants develop their own memories by accumulating sets of credits earned and experiences involved". The intertemporal nature of networking is also portrayed in the article by Baker & Faulkner (1991) through filmmaker's adaptation and imitation behaviour when dealing with single project organizations. The industry adapts by holding on to role combinations that proved strong in solving technical and organizational problems and copy role combinations

of productions that turn out to be hits. Studios look for the right blockbuster ingredients, resulting in higher degrees of specialization. Miller & Shamsie (1996) make a distinction between propensity based resources and knowledge based resources. The former category is protected from imitation by property rights, the latter by knowledge barriers. The authors estimate the effects of both types of resources on movie productions by 5 major studios. The study is performed over the period 1936-1965, but the period is divided in a pre- and post- 1950 period, marking the change in integrative structure of the industry as well as increased uncertainty characterizing the market. Property right resources are proxied by stars under long term contract and theatre ownership, knowledge based resources by academy awards and the average production costs over the previous years. The latter have to reflect the respective presence of skill and complexity. The authors establish that property owned resources have a dominant impact at periods of environmental stability and predictability while knowledge based resources are more performance enhancing in a context of increasing uncertainty. Movie performance is measures by returns on sales, operating profits and market shares.

Sedgwick & Pokorny (1998) is one of the highest cited publications in the cluster. The paper leads a second group of literature offering an economical performance analysis of the sector from a historical perspective. Other important references include two articles by Glancy (1992), (1995) providing separate overviews of both MGM and Warner Brothers film grosses in the first half of the twenties century based on the William Schaefer ledger and the Eddy Mannix ledger. Those information lists record the history of the studios' films, their production costs, domestic earnings, foreign earnings as well as profits and losses. The study of Sedgwick & Pokorny (1998) is based on the Warner Brothers historical dataset. Historically, it is important to view the evolution in the movie industry against 5 demarcation points. The early 1920's where studios grew through a strategy of vertical integration, secondly the introduction of sound at the end of the 1920's, the period of the great depression between 1929-1931, the second world war and last the post-war period, where due to legislation introduced, disintegration took place and some common practices of the sector had to be altered. The early 1920's were an area

of expansion. The introduction of sound initially lead to an increased expansion, but this effect was stopped by the great depression between 1929-1931. The studio's performances were strongly hit by the recession and film attendance was reduces up to a third. Warner studios reacted to the first recession through a strategy of innovation and cost control, which resulted in a slow recovery in the late thirties, but the relations between performance indicators were substantially altered from that period on. Sedgwick & Pokorny (1998) show that large costs attached to high profile productions, unlike before, were not necessarily reflected any more in higher profits, while it still went hand in hand with increased variability. The return to risk ratio was generally estimated to be lower. Those findings can be better understood in the framework of portfolio theory, where the performance of a movie is not seen in isolation but in interrelation. Not only considering risk taking as a weighted average of risks, but also accounting for the covariances between the rates of return of different projects, the authors show how risk taken differs from experienced risk. Risk taking can from this perspective be interpreted as an intrinsic strategy and the risk stance taken by the studios were dynamic and depended on previous period's increment in the rates of returns.

Mezias & Mezias (2000), analyse a data set over the period 1911-1930 taken from the American film catalogue of motion pictures, to show that a lesser degree of integration and a higher form of specialization led to the introduction of new genres. They observe that Warner studios, even in an integrated structure, reacted to the depression of the early 1930's with product diversification consisting of artistic and genre innovations which appeared a successful strategy in turning losses into profits. Robins (1993) uses Warner Brother archive data over the later period 1945-1965, to study differences in performance between integrated productions and independent cooperation relations. The changes taking place in that period in history are argued to have led to different types of productions altogether with a high impact on economic performance. A complex system of contracting resulted in a situation of increased cost combined with more distinguishing products. Aiming at hits, the performances were at the same time more volatile, but independent productions outperform studio films in return, in return over cost and in cash flow. This high

focus strategy is also confirmed by a more recent paper by Sedgwick & Pokorny (2010), studying long term post-war movie performance.

The last cluster stands more isolated from the rest. It bundles literature related to trade in movies and is the smallest cluster, composed of only 10 items, 2 being books and three writings on trade in television programs. Only 2 papers on movie trade are represented in the 100 best cited articles, namely an article by Oh (2001) titled *"International trade in film and the self sufficiency ratio"* and *"The Economics of American Theatrical Movie Exports: An Empirical Analysis"* by Jayakar & Waterman (2000). The study by Oh (2001) investigates the determinants of the self-sufficiency ratio which is defined as the share of domestic films in the total box-office revenues of a country. The article includes an empirical test of some theoretical hypothesis' described in the work by Wildman & Siwek (1988), also represented in the cluster, suggesting that in case of free trade in movies among countries, the larger country will benefit. The reason is that linguistically, larger countries are able to make larger investments, leading to higher quality products. This will not only increase overall box office revenues but simultaneously result in a substitution effect which will influence the self-sufficiency ratio. The second factor into play is the cultural distance between countries. As the US is the major producer of movies, the position of certain countries are mainly measured against it. In line with this, the linguistic factor is narrowed into whether or not the domestic country is English speaking. The author establishes significant positive empirical relationships between the market size, as expressed by GDP, cultural distance and total box-office revenue. Cultural distance is measured using a four dimensional scale introduced by Hofstede (1980) reflecting power distance, individualism, masculinity and uncertainty avoidance combined with the Hofstede & Bond (1988) dimension of Confucian dynamism. The English language effect appears to be of less significance. The article by Jayakar & Waterman (2000) is very similar. They use total spending in terms of the sum of annual box-office, video and pay TV revenues and deny the cultural difference effects. However the conclusions are the same, a positive effect between overall spending and domestic market share as well as a limited effect of language. Also more recent articles follow grossly the same pattern. Fu & Govindaraju (2010) rely

dominantly on the same explanatory variables, but take taste similarity as dependent variable. Similarity of taste is measured as the strength of correlation between domestic and foreign box-office receipts and is shown to be significantly affected by the size of the market. The study of trade patterns is a still actual item in research covered by economics as well as media and communication literature. Although there is a clear relationship to strategic decisions by major studios, this group of literature hardly relies on it. The nature of the analysis is quite stereotypical and estimates self-sufficiency type ratios or other forms of indices, expressing the share of foreign over domestic demand to variables of market size and cultural difference.

1.3 Discussion

The cluster analysis based on co-citation elevates the review process, uncovering items that are considered as the core references. They point to a segment of research represented in most overviews on the specialty, while at the same time offering valuable insights into key topics and how they interrelate. It teaches that a more traditional division in terms of supply-demand, micro-macro economics does not always hold, because micro behaviour in movie economics is often tested using aggregated data and the actions observed at the supply side of the sector follow a logic in which demand uncertainty rules due to the experience good nature of film as a creative product. A large share of the movie economics literature is highly empirical and focuses on investigating the relation between various measures of performance such as box-office revenues or profits, placed in relation to factors which are considered as beneficial forces behind decision making. The dominance of this type of literature has two rationale. First, there is the data availability. The fact that certain organizations keep data on movie characteristics and movie performance provides researchers with large and interesting databases. The rise of all sorts of movie internet databases and communities makes this specialty even more interesting for empirical study. Secondly, the movie industry is one faced with high uncertainty. It makes movie projects risky where investments are high and predictive models are

thus welcome to support the sector in its decisions.

Although the econometric methods applied to study movie performance advanced highly over the last decade, in essence they consist of estimating a predictive model between box-office statistics and a number of explanatory factors that remained highly constant. The elements mainly investigated as potentially influencing movie success are the presence of stars, the influence of critics, reviews, ratings and other "quality certifiers" such as academy awards or nominations. While there is a high similarity in studies in terms of factors examined, there is overall little agreement nor on the significance of the variables, nor on magnitude or even direction of the effect. The role of stars and critics serves as important cases in point, where also conceptual differences makes results difficult to compare. The choice to opt for a variable is generally motivated and placed in the framework of movies as an experience good with uncertain quality. That said, there is overall little theoretical underpinning as to why the model is like it is. Despite the fact that the empirical analysis is very demand orientated, there is ample reference to consumer theory and little attention to the formation of preferences and taste heterogeneity.

The writings of De Vany and Walls fill that gap to a certain extent, offering a model of consumer choice under quality uncertainty. Starting from a setting of consumers sharing information, they demonstrate that the transmission of information leads to Bose-Einstein distributional dynamics for motion pictures revenues. The Bayesian choice model employed and the stochastic dynamics that follows explain the bifurcation process. Positive word of mouth can influence others in their choices, that way creating successes or hits or vice versa. Theatrical supply does follow demand in that respect. The literature of De Vany and Walls can be considered as an important contribution to the specialty. It provides an explanation for the observation of skewed distributions of movie performances which questions the validity of many linear regressions used in the movie literature assuming Gaussian distributions. Important also is the Bayesian intertemporal frame where consumption at a time point is conditioned on previous choices, hence introducing decision dynamics. Despite

the fact that De Vany and Walls provide an experience good based intertemporal consumer model, demand type empirical studies are as such not omnipresent in their work. The emphasis is on studies related to the shape of movie revenues and how they are affected by the factors that can in general be traced in the performance literature. The papers of De Vany and Walls are written with the aim of feedback to the supply side and their article that revises the Paramount case is very much rooted in industrial organization. Past and current practices such as block-booking and price policy are viewed in the light of those historical decisions and their rationale are put into question relying on models borrowed from industrial organization theory. The literature also relies on the creative products literature, like does the more historical cluster. The latter, while sometimes more narrative in nature, is also more in depth when thinking about movies as a cultural product and the film business as a temporal project. They are located close to the performance cluster, but the higher degree of conceptualization and the historical perspective gives a better understanding of why the current movie industry is shaped the way it is. The IO group is more inspired by game theory or by models of oligopolistic decision making under risk. The theoretical underpinning is stronger compared to performance studies, but the proposed models are generally not directly estimated nor simulated.

To achieve a better understanding of consumer decisions for creative goods in general, and movies in particular, it is important to reflect upon what typifies movies as a product. The majority of existing literature implicitly or explicitly refers to the innate uncertainty characterizing the sector and performance estimates bear on the notion of art as an experience good or information good. This is why information sources like critics and consumer reviews are considered as being of high importance in studying factors which determine box office results. They send the consumer a quality signal, filling the lack of information before making a choice. Similarly, the uniqueness of the product is the underlying determinant in the industrial organizational studies where the production of the movies is studied as a temporary project. The uniqueness of the product raises questions on what is meant by endogeneity of preference when in fact each time a new choice is made over an entirely new spectrum of products. One way to think about this is to consider the creative item as

an entity endowed with a bundle of characteristics. Consistency in taste can then potentially be expressed in terms of consumer classes which assemble products of similarity. Such an approach demands a different type of modelling, looking at decisions as a discrete act over a multifaceted item, away from consumer models defined in terms of standardized infinitely reproducible or divisible goods. Away also from the idea of a representative consumer. Despite the fact that most empirical models are meant to predict the market, aspects of consumer heterogeneity are largely ignored. However, if one wishes to fine tune marketing actions, questions such as how to define preference partitions for films are prominent. One option is to work with genre, but genre is not uniformly defined and not the sole factor shaping an item.

Yet, the notion of movies as an experience goods is one to be transmitted. It includes aspects of ex ante uncertainty as well as ex post evaluation. Agents inform themselves, but the question remains what type of factors are dominating their decisions. It is difficult to disentangle those elements by studying movie performance at aggregate levels. Moreover, the question of the nature of the experience remains open. A lot of information is available about ex post experience given that quite some viewers share their opinions on forums and give ratings to the movies. However, it remains unclear what shapes the experience and how the major elements that determined the experience relate to the final evaluation. Experience encompasses endogeneity or learning. Individuals are endowed with an information set, which can be labelled as cultural capital. Having had the experience means that they can learn and revise their preference structure. Given the uniqueness of the product or project, agents will not consume the same product at a later stage. The question than is how to model and test dynamics when dealing with similarity classes and heterogeneity between consumers. The main objective of this thesis is to select the most apt models to deal with the specific nature of creative products and investigate whether meaningful preference segments can be distinguished that reveal the main decision making factors for each such class at single or multiple time points.

In order to detect the factors underpinning consumer decision for movies, it is impor-

tant to work with individual level data where consumers register movie attendance over a longer period of time. Movie economics performance studies are mainly based on aggregated data, mostly box-office data. While acknowledging their merits for movie economics research, they are less suited for tracing heterogeneous consumer classes. Nowadays, the important source of individual level consumer data is the internet. Through multiple fora, consumers express their appreciation for films and rate them. That type of data sets have been used to study word of mouth effects in movie economics, but more importantly lead us to a large body of computer science literature dealing with recommendation of creative products. They allow to segment the market for creative goods and generate automatic recommendations to users based on the abstraction of his/her preference profile. These tools are used by e-commerce companies and rely heavily on online data of creative goods, and more specifically on movie ratings, to test and to predict consumer choice. Being aimed at prediction, they specify consumer preferences without real reference to them, nor do they study in depth the nature of the segments that are underneath the recommendation. The data used are big data that often come in an unstructured form. The statistical methods applied to them depart from those used in movie economics, consisting mainly of data mining techniques.

The aim of this thesis is to bridge the gap between consumer theory on artistic goods and computer science, more specifically, studies on recommender theory. This implies investigating the potential of rich but unstructured online data as well as the data mining techniques to segment consumer groups. In the next chapter, the concept of experience good and the multi-featured nature of creative products are unravelled, to be embedded into a selection of economic models that deal, at least partly, with the challenges this type of commodities offer to choice theory. Then, the main streams of recommender theory are investigated, resulting in the specification of a number of translational methods, concepts and data types that can be transferred to economic analysis. The next two chapters can be seen as an exploration in research on creative products in two branches of science, each characterized by their own methodologies and practices but with little interference between them. A selection of data and models will be used in the second half of the thesis to inves-

tigate the main features that consumers attach to movies and how those elements co-occur. The most representative elements will be used as independent variables in a consumer segmentation model. In a last stage, consumer dynamics will be introduced.

Chapter 2

Creative Goods Modelling in Economic Theory

2.1 Creative products as experience goods

As was argued in previous chapter, the majority of movie economics research, theoretical or empirical, either explicitly or implicitly, refers to the underlying idea that creative products, such as films, are experience goods. The concept of experience goods was developed by Nelson (1970) and Wilde (1980). It is generally defined as a class of goods for which consumers can evaluate quality only after they have been bought and consumed. Schmitt (2011), in a comprehensive introduction to experience marketing, detects two streams that correspond to the linguistic nuances between the German words erfahrung and erlebnis. The first is related to accumulated knowledge over time where the second bears on here and now direct observation or perception. Consumer experience literature builds strongly on the core article by Holbrook & Hirschman (1982) "The experiential aspects of consumption: consumer fantasies, feelings and fun". They oppose a rational choice information processing approach, which is the stance of micro economic theory, to an experiential view encompassing an immediate flow of sensory pleasure, aesthetic enjoyment and emotional response. Complex knowledge structures of beliefs and thoughts anchored

in psychological theory of cognition are placed against mental processes that are more subconscious and private in nature and thus more phenomenological. Learning is not only instrumental - where the past cognitively reinforces future behaviour - but also entails contiguity: the frequency with which neural events, feelings, pleasures, symbolic components are paired together to become evocative at a later stage. Their theory appeals to products where not so much utilitarian objective features steer customer's decision, but rather the symbolic meaning of in particular subjective characteristics, and therefore suited to address creative products such as novels, movies, fashion or music. The stimuli associated with them are primary non verbal multi-sensory descriptions, replacing the delineated attributes inherent in traditional consumer research paradigms. Opposing the experiential view to the traditional information processing consumer theory, Holbrook & Hirschman (1982) consider their theory as complementary rather than a replacement for it. Gentile et al. (2007), borrowing components from others, LaSalle & Britton (2003), Shaw & Ivens (2002) and Schmitt (1999), provide a comprehensive definition of what they gather as determinant elements of consumer experience:

"The customer experience originates from a set of interactions between a customer and a product, a company, or part of its organization, which provoke a reaction. This experience is strictly personal and implies the customer's involvement at different levels (rational, emotional, sensorial, physical and spiritual). Its evaluation depends on the comparison between a customer's expectations and the stimuli coming from the interaction with the company and its offering in correspondence of the different moments of contact or touch-points."

The schism of the cognitive versus the emotional approach was contrasted by Simonson et al. (2001) and rephrased as hot versus cold aspects of consumer behaviour. Cold refers to elements such as perception, learning, attribution and decision making while the hot elements are in the sphere of mood, arousal, regret and the more hedonic dimensions of consumption. The authors point to a fast decline in the treatment of cold aspects, going from 85 percent in the mid 1970s literature to around

64 percent in the 1990s. Indeed, marketing research, inspired by some influential books on consumer experience, Caru & Cova (2007), and the experience economy, Pine & Gilmore (1999), performed a great number of research on the nature of the experience, mainly pinning down the core dimensions that constitute experiences in terms of sensorial, emotional, cognitive, pragmatic, lifestyle and relational components, Schmitt (2011), Gentile et al. (2007). Pine and Gilmore (1999) distinguish four types of experiences, namely entertainment, educational, escapist and aesthetic, each placed along different lines of customer involvement, the desire to connect, active versus passive and absorption versus immersion, Hosany & Witham (2010). Oh et al. (2007), in an attempt to capture the nature of tourism experiences, appeal on replication to strengthen results with an emphasis on the relation with post-consumption evaluation. Where rational consumer theory based on information processing has long been the standard, dominating both theoretical and empirical literature, the treatment of experience goods appeared difficult from that perspective. Looking at it as "erfahrung", as the interaction between prior expectation and the judgement made afterwards, got far less attention. It requires pinning down to some key factors, namely time, expectation, intangibility and the multi-featured nature of items, all making that cognitive views on experience goods are difficult to conceptualize. The basic question is what elements are judged prior or posterior to the choices made, how to discover them, how to catch them into an empirical setting, into a dynamic setting. Moreover, what type of data are suited to have the belief and knowledge evolution tested?

To understand the concept of creative goods, one needs to focus in on those elements. The multi-featured and re-featuring nature of the good, the importance of the time aspect and the uncertainty involved are interrelated to each other. Creative goods are characterized by the fact that each creation is different, and to that extent one could say that experience goods, contrary to most goods, do not allow standardization. For example, each film is novel in that its constitutive images and the renewed ordering of it makes the film meaningful, Sedgwick (2007). Bianchi (2002) redefines a class of novelty goods where not functionality or efficiency are at stake, but rather their variability, the offer of new variants, new combinations of characteristics. As

stated by the author: "individuals do not choose different varieties among goods but the same good in different variants, consumers change locations in the characteristic space". What makes a creative product is its uniqueness. Lancaster (1966) already pointed to the challenges facing traditional consumer theory, arguing it is impossible to foresee the potential utility of a good that is yet to be invented. Each creative good is a highly differentiated composition or re-composition of already known or innovative features. Yet, along the dimensions that establish the experience, the conviction exist they form a holistic evaluation that surpasses its constituent parts. One speaks of experiencing that performance or that movie. As noted in Holbrook (1987) "one perceives the components of a performance together as a gestalt in which all stimuli interact". When preference judgements are made, a product is opted for as an entity with the underlying choice consideration based on feature comparison.

A well known feature matching model was initiated by Tversky (1977). Borrowing on the insights of Lancaster that a commodity corresponds to a particular profile of attribute values the consumer chooses from, he presents a theoretical model of similarity where objects are represented as collections of features and similarity is described as a feature-matching process of comparing linear combinations of common and distinctive elements respectively. It is called the contrast model. By conceptualizing similarity/dissimilarity as a feature matching act he confronts previous geometric approaches to similarity. Other authors, building on those insights, show that some key features are more salient and influence preference stronger, Dhar & Simonson (1992), Dhar et al. (2000). Others compare objects to make judgements of similarity that are based on categorization, generalization and discrimination, Nosofsky (1986), Dhar et al. (2000). Tversky & Kahneman (1974) describe the process as one of representative heuristics, weighing the degree to which an object is similar or stereotypical for its category. They could be thought of abstract goods representative for that class. Conversely, two products can be considered as equivalent or belong to the same class when their attributes range around those of a typical product. Bernardo & Blin (1977), represent goods by means of a matrix of features and segments. The consumer decision process translates into four factors: 1. A set of silent attributes typifying the objects. They are taken as data inputs in the broadest

sense 2. Attribute weights 3. A ranking over attribute weights which implies taking that good for which that attribute is most important, not to be interpreted in a qualitative sense. 4. A preference over items. The theory latently implies a number of mental actions: that the structure of importance of attribute attachment comes across through training and that, once faced with an unknown item, the user is able to categorize it by inferring if its constituent components fall in the neighbourhood of the prototype in mind. The multitude of attributes makes judgement somehow fuzzy, as individuals do not always hold perfect information. An allocation to a category is therefore based on beliefs about the likelihood of outcomes. Tversky & Kahneman (1974) identify three heuristics that are applied to assess probabilities in judging this type of multi-features items: 1. Representativeness, judging the degree to which the object is typical for the category 2. Availability, the ease with which the frequency of an event or class can be brought in mind 3. Adjustment or anchoring, referring to the dependency of the process to how and where you initiate and the way it is adjusted over time. All heuristics are shown to suffer from cognitive biases.

Each time a consumer is faced with a decision to view a movie or not, he/she has to deal with a different creation/production, a product within the same category, yet different and expected to provide you with a new experience. The time aspect can therefore be considered as fundamental, since experience has to do with undergoing the modalities of a commodity or with living the effects it exhibits, be it cognitive or sensitive. The characteristics are of both a tangible and intangible nature and can only be judged upon by going through them. Kahneman & Snell (1992) translate it in terms of distinction between experienced utility, defined by the quality of the hedonic experience and the predicted utility, individual's beliefs about the experienced utility of that outcome sometime in the future. In their decision to choose one movie above the other, individuals form expectations on various features through the interpretation of signals that are provided to them by producers such as genre, actor. The signals are added to a consumer's mental framework of evaluation, which in turn is a substructure within that person's world-and-life-view. They are both consonant or dissonant to the ones that shaped their current taste. Actual views can be considered as dynamic systems influenced by exogenous factors as well by en-

dogenous learning based on going through experiences, Sedgwick (2007). The class of creative goods generates a different sort of dynamics, including stages of perception, pattern recognition and pattern comprehension. Important in the philosophy of aesthetic experience are the writings of Dewey, claiming that the interaction aspect of the experience is fundamental, those between a person and the environment, where one takes up something from past experiences which modifies in some way the quality of those which come after. It occurs when information from the artwork interacts with that already in the person's mind, Dewey (1934). Dewey pictures how qualities are constructed and reconstructed of what we learn and know, Uhrmacher (2009). With that, he acknowledges past, present and future of an experience and partly lifts the distinction between the cognitive as opposed to the sensory, or the here and now versus the knowledge building. In their interaction with the product, consumers learn to map the relationship between the characteristics and the objects and the potential utility provided. Judgements and decisions are hereby affected by previous judgements. Dhar et al. (2000), taking the Tversky framework as starting point, sketch feature comparison effects on preference formation in a dynamic context and demonstrate how the type and direction of the initial judgement influence the weight attached to common and unique features in subsequent choices.

As a summary, a suited consumer modelling for movies will treat creative products as holistic entities, comprising singular features that are both intangible and tangible. This can refer to characteristics attached to the product and provided for by producers/distributors as well elements signalling the information value of the commodity. Where each product is an innovative creation, judgement does not stand on its own, but compares to categories and concepts already known. It is assumed that a new movie can be placed against categorical structures, classes which are potentially of a probabilistic nature. The probabilistic nature transcends observable attributes and encompasses latent class structures that can only be inferred. Discrete choice theories, bearing on Tversky's contrast model, are a good starting point to think about the choice procedures underpinning movie decision making, treating the choice process in relation to a holistic yet composite view of commodities. However, more is needed to capture the complexity of movies as experience goods.

While the influential work of Holbrook & Hirschman (1982) on the experiential introduces the non-cognitive processes that steer choices for commodities with high symbolic and aesthetic value, one cannot by-pass the intertemporal nature of the decision process. The value of uncertain attributes are judged, placed against potential similarities and evaluated again. One can strive for a comprehensive theoretical model, fully describing the complexities of the choice of creative goods, if anyhow possible, but empirical research is likely to be limited to distill only a fraction of the relevant features and categories. Here also, driven by data availability, the exercise presented will be one of stimuli against decision, acknowledging the latent categorical structure. The assumption of the representative consumer is thereby replaced with taste heterogeneity or differences between individuals. The latent structures, the features that make the categories, as well as the relation of individuals vis a vis the categories cannot be observed, but are indirectly statistically inferred from observing behaviour. The next section will give an overview of the economic models used in literature to model consumer behaviour for experience goods. A following part will present a random utility model, matching a probabilistic choice model that will serve as building block for the analysis performed in this thesis.

2.2 Formal models of experience goods in economic theory

Economic theory formalizes consumers as active maximizers over the choice sets available to them. Goods are often unidimensional and infinitely divisible and the utility function defined over goods is homogeneous. More of a good comes with greater utility. Gorman (1953), Becker (1965) and Lancaster (1966) opened up the definition of commodities in a way that an intermediate stage was introduced by defining them as composites generating a stream of services. A movie can be thought of as an item producing immaterial commodities such as recreation and enjoyment, Bianchi (2002). With time and improved skills, cultural capital enforces, and individuals become more efficient in their attribution of item to commodity. In the same

setting, Stigler & Becker (1977), in one of their key writings "degustibus non est disputandum" install intertemporal decision dynamics. Utility is made dependent on commodities, made functional on the quantity of goods, time spent and a variable representing human capital formation. Preference shifts are rationalized by modelling them as state contingencies. An example is provided where the utility function includes a factor reflecting the amount of music appreciation. Music appreciation is produced by a function that depends on the time allocated to music and training therein and of other human capital conducive to music appreciation. The human capital factor depends on previous amounts of the commodity consumed, hence installing dynamics. The addiction effect lowers the price of music appreciation at younger age and increase the time spent on it, which can be seen as investment in cultural capital formation. Despite addiction, time spent does not necessarily keep growing as the consumer becomes a "more efficient" generator of music enjoyment.

The rational addiction model of Stigler and Becker caught the interest of applied cultural economics for its tractable treatment of cultural commodities. Cameron (1999) tests the model on cinema demand in the UK over the period 1965-1983, the dependent being per capita attendance. Based on a t-test, the study rejects the myopic model in favour of the rational addiction model, both past and forward indicator being significant. The author places a footnote however, admitting results cannot be repeated using estimation procedures suggested by Becker et al. (1994). Sisto & Zanola (2010) empirically investigate the addictive component of cinema consumption based on pooled cross-section and time-series data on 12 European countries over the period 1989 to 2004. Results also provide evidence that cinema consumption conforms to a rational addiction hypothesis. Both the past and forward coefficient prove positive and significant, with the forward indicator being smaller. Here also, the authors add a word of caution based on the fact that aggregated data are used, potentially causing bias.

As pointed out by Bianchi (2002), it remains difficult to think of this type of models in terms of novelty goods. Where it is easy to understand functional improvements,

it is altogether more difficult to see it in the light of say a new book or movie. Moreover, buying more of the same does not add to utility or experience. Another fundamental shortcoming of such models is that they don't deal with the fact that, prior to consumption, persons might face uncertainty with respect to how useful an item will be in relation to the characteristics produced. Inspired by the "experimental consumer model" such as introduced by Kihlstrom et al. (1984), Amez (2003) suggests a model where consumers face prior uncertainty about the quality level of aesthetic attributes. They form ex ante beliefs and decisions are taken rationally, maximizing expected utility over the uncertain attributes. After going through the experience, agents make a judgement and update their beliefs, through a Bayesian procedure. As the experience is only an imperfect signal, exposure to arts affects the information set in the sense he/she gradually learns about the "true" quality level. It is shown that in a fully rational model of forward looking behaviour, consumers experiment with arts at early age because departing from the myopic action contributes to the store of information from which to make more refined choices in the future. Based on similar theoretical foundations, Garboua & Montmarquette (1996) estimate a model for theatre demand using French survey data taken from 1000 participants. They show that demand reinforces in subsequent periods. Their results do underline the experiential learning model by showing how pleasant experiences encourage us to repeat our choices, and vice versa, and by establishing the effect of prior information. A Bayesian, limited rationality version of the model has been presented by Armantier et al. (2015), examining its validity by means of a laboratory experiment on preference learning for music. Their results provide evidence for Bayesian learning in combination with expected utility maximization, albeit imperfect in the sense that individuals overreact to information about unfamiliar goods and under-react on matters that are more familiar.

The path breaking work by De Vany (2004) on box office dynamics and herding also relies on a bounded rationality Bayesian demand model. Given a particular information structure, agents are faced with a sequential decision problem between alternatives. The information set consists of private quality signals on the goods, the sequence of past behaviour and public signals indicating the observed choices of

others. If the dynamics of demand takes the form of a mapping where choices are dependent on the history of previous choice from self and others, then under well defined circumstances, a bifurcation can take place where the mapping is contractive for some movies and expensive for others, hereby explaining why some movies capture the whole market and become hits while others fail. Sedgwick (2007) also deploys a bounded rationality model for movie demand, acknowledging that the novelty character of a movie induces prior uncertainty to the consumer. Prior to the consumption of a film, individuals do not have a complete idea of the visual and aural cinematic utility, but interpret information signals, each of which and in combination, become a factor within a consumer's mental framework of evaluation. To make a decision, consumers call on their personal history of filmgoing and to a full range of sensory experiences entailed, both imperfect guides of future film quality. Based on their experiences, agents develop heuristic for filtering films from their choice set, in order to arrive at a decision set. The decision making process within that set is modelled through a Shacklean approach, expected utility promised by a film is mapped against the uncertainty of that utility being realized. People make errors that are manifested in differences between expected and realized, but those are not considered to be systematic and people learn from experience in a gradual manner. At decision stage, the expected enhanced utility is weight against the degree of potential surprise, hence intrinsically dependent on the filmgoer's attitude to loss and gain of cinematic utility. The theoretical processes are empirically linked to release patterns of big screen versus low budget "sleeper" productions, where the latter makes consumer's consideration set gradually shift.

The model of Sedgwick (2007) on movie demand integrates thoughts on uncertainty of movie quality and expectations with a discrete choice model that takes into consideration how individuals build their consideration set, hence supporting to a larger extent the idea of movies as novelty goods. By analogy, this thesis suggest a discrete modelling approach as the most appropriate way to deal with the nature of creative goods. In this type of model, choices are made between a finite set of discrete alternatives that can be specified by their constituent elements, choices that are stipulated in probabilistic terms. In the next section, a short overview will be given

on random utilities models and how they give rise to the logistic and latent class models that will be used as empirical entities in this research. Moreover, through a Markov sequence, a dynamic layer can be placed on this type of models allowing us to test the intertemporality of the decision making process.

2.3 Probabilistic discrete choice models as core building blocks

Discrete models are predominantly used to deal with choice experiments data. They model exclusive outcomes and unite a vision that utility is constructed from the attributes of products, in line with the early Lancaster (1966) writings and with random utility theory based upon McFadden (1973). Movies cannot be placed in a continuum and decision models assuming infinite divisibility of quantities fall short in designing the process of choice. The central question is, which movie an agent will choose, given an array of alternatives, conditional upon the portfolio of items supplied at a specific moment of time. Random Utility Models (RUM), which date back to Marschak (1960), maintain the economic principles of utility maximization. They work on the assumption that, albeit imperfectly, observed choices reveal something about underlying utility. Utility is decomposed linearly into a systematic observable component and a degree of unobservable randomness. The deterministic component integrates the measurable attributes of the alternatives/individuals, Baltas & Doyle (2001), the randomness reflects the latent personal and uncertain structure inherent to utility formation. This reflects the heterogeneity in taste among agents as well as measurement errors and effects of ill-specifications, Manski (1977). It means that choice is expressed in probabilistic terms and that the distributions attributed to the error terms strongly determine the shape of the experimental choice model.

Let U_{ij} note the utility of individual i for alternative j composed of a deterministic component V_{ij}, also called representative utility, and a random component ϵ_{ij}.

$$U_{ij} = V_{ij} + \epsilon_{ij} = z_{ij}\beta + \epsilon_{ij} \tag{2.1}$$

where z_{ij} represents the observable attributes to the consumer related to alternative x_j, and β is the vector of coefficients. When faced with a set ξ of X mutually exclusive alternatives, utility maximizing behaviour will make the consumer opt for alternative j over alternative k if:

$$\begin{gathered} P(U_{ij} > U_{ik}) \\ P\ ((\mathrm{V}_{ij} + \epsilon_{ij}) > (V_{ik} + \epsilon_{ik})) \\ P\ ((\mathrm{z}_{ij}\beta + \epsilon_{ij}) > (z_{ik}\beta + \epsilon_{ik})) \\ P\ ((\mathrm{z}_{ij}\beta - z_{ik}\beta) > (\epsilon_{ik} - \epsilon_{ij})) \end{gathered}$$

The probability that an individual opts for an alternative is determined by the shape of the distribution of the random variables and the integration over the possible values the variable ϵ can take. Luce (1959) and later McFadden (1980), establish the equivalence between random utility and logit models. When assuming that the error terms are identically and independently distributed and follow an extreme value distribution, the choice can take the compact form, Baltas & Doyle (2001):

$$P_j = \frac{exp(V_j)}{\sum\limits_{x \in \xi} exp(V_x)} \tag{2.2}$$

The independence of irrelevant alternatives assumption (IIA), or the choice axiom is important in establishing the equivalence between RUM and logit. It implies that the choice between two alternatives is not affected by the presence of an external

option. The basic logistic model is the core of the empirical analysis pursued in this thesis and will be discussed more in depth in chapter 6. It is the kernel that can be enhanced to embed to a larger degree information processing, state dependencies and heterogeneous classes across individuals in what can be defined as hybrid choice models, Ben-Akiva et al. (2002). One way to capture the differences in individual behaviour is by introducing latent classes. It assumes heterogeneity by allocating it to a finite number of segments or classes, resting on the assumption that individual behaviour depends on observable attributes attached to latent factors that can be traced, Greene & Hensher (2003). The interaction with the characteristics are taken similar in a group or segment, the choice of an item becomes dependent on the class. Assume a finite set of classes $c \in \chi = \{c_1, ..., c_K\}$:

$$P(j|c \in \chi) = \frac{exp(z_{ij}\beta_c)}{\sum\limits_{x \in \xi} exp(z_{ik}\beta_c)} \tag{2.3}$$

Starting from a latent class framework, dynamics can be introduced by making the classes in one period dependent on the outcome at a previous time point. This can be done by imposing a latent Markov or latent transition model. Starting from the latent segments introduced, a time dimension can be added by considering a sequence of $t : 0, ..., T$ measurement occasions. Then, the probability of an individual belonging to a particular latent class can be made conditional on the outcome of previous period: $P(c_{ik,t}|c_{ik,t-1})$. That way, heterogeneity between individuals is captured by the diversity of classes, the potential shift of taste is modelled through class switches. A full model estimates the probability of initial classes, the probability of staying or moving between classes and the probability of making a choice for a product, given the endowment of features.

2.4 Summary

Creative goods are characterized by a low degree of standardization. It implies that classical economic models assuming reproducible products are not suited to serve as a theoretical base. As each new movie is a reshaping of features, other types of decision models will have to inspire the empirical design of the consumer choice process. The feature matching model by Tversky (1977) and the discrete choice theories suggested by McFadden (1973), (1980), offer more valuable perspectives to deal with the multifaceted nature of creative commodities. Agents do not always judge an item in isolation, but compare with past experience and set their judgement in similarity or contrast to categories already installed, against prototypes or exemplars. Accordingly, it is appropriate to model movie decision making using discrete choice models that are probabilistic, thereby reflecting the inherent uncertainty of creative goods, adding latent segments to deal with categorization and consumer heterogeneity.

From the experiential literature, one can borrow the concept of "contiguity", the frequency with which neural events, feelings, pleasures, symbolic components are paired together to become evocative at a later stage. Translated in statistical terms, it calls for an approach in which the frequent joint co-appearance of features leads to the formation of categories and where the heterogeneity of consumers can be expressed in terms of their distributions over alternative consideration sets. Here also, clustering or latent probabilistic models appear valuable candidates. Experience also implies a time aspect and installs dynamics. A person goes through an experience, enriches her/his knowledge base and places novel products against known prototypes. Building an empirical framework to support the presence of dynamics in consumer behaviour has always been difficult and this is even more so for creative products as it is an open question what persistence in preference means when dealing with products that are always new. However, when thinking of it in terms of an ongoing reshaping of features, placed against categories, stability in taste can be viewed as staying within a certain segment. In the realm of probabilistic choice

models, extending to a Markov transition model is a logical step.

As was argued in the literature overview, movie performance studies are dominated by linear regression models. Bayesian clustering techniques are rather found in the domain of big data analysis. Computer science, and more specifically recommender theory, provides a substantial contribution to predicting user's potential choices for movies, based on various machine learning techniques. This thesis is therefore approached as a multidisciplinary investigation, screening the best statistical techniques to support the ideas borrowed from movies economics. Moreover, the aim is to learn what type of models can bridge between both worlds. The datasets used in the recommender literature are interesting, as they can provide that type micro-level data suitable to test the experience and categorical nature of movies. The movie economics literature puts forward some driving forces of decision making and question is if indeed these are the elements at stake. Internet data gathered through fora where users can freely express their opinions are loose and unstructured, but might provide insight in some of the main facets underlying consumer's motives. While testing the intertemporal categorical nature of decisions in a bottom-up way, using online data freely added, is bound to be rough and tentative, it will be presented in this thesis as a potential way forward in thinking about consumer models for creative products.

Chapter 3

Recommender Systems: An Overview

3.1 Introduction

The internet offers networked individuals wide opportunities to share information. Web 2.0 broadened the user's options to contribute information and share opinions through enhanced interfaces and new types of public fora made available. The upcoming of an array of social networks made connected individuals while at the same time partitioned them into niches. In parallel, the actions of the participating citizens proved valuable input for those wishing to monitor or analyse user behaviour, be it to improve applications or search engines, to integrate contextual information or to steer people faced with an abundance of information and options, in their search of what product is worth looking at. Indeed, the increased accessibility and the open access nature of the web brings new opportunities, the backside being a potential danger of information overload. Recommender systems are software applications that convert user information into item recommendations. First developed to guide users through internet searches, they are used more and more for contextual advertising or to suggest sales items that potential buyers might be interested in. The automated systems help people to deal with information overload by building mod-

els that predict the extent to which they will like an item and then recommending it, Ekstrand et al. (2011).

In a reference definition, Resnick & Varian (1997), describe recommenders in terms of people providing recommendations as inputs which the system aggregates and directs to appropriate recipients. Burke (2002) systemises this by specifying three information or action sources 1. background data that are in the system 2. input data generated by the interaction between user and system 3. an algorithm connecting background and input to create a suggestion to the user. The author speaks of "*a system that produces individualized recommendation as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options*". The difference with information systems lays in the personalized. Like information systems, recommmender theory borrows techniques from various disciplines such as machine learning, data mining, artificial intelligence and is inspired by cognitive science, and forecasting theory. Being designed to recommend new items to users, algorithms are often valued in terms of prediction accuracy, using statistical measures such as the root mean squared error (RMSE), in terms of precision: the number of true positives over all recommendations, or recall: the number of true positives over the preferred items.

Historically, the forerunners in recommender theory are placed in the 1970s, with methods to group library users into stereotypes. The upcoming of IT related recommender systems is placed at the beginning of the 1990s. Early examples include Tapestry, developed by Xerox, supporting users by reducing the quantity of incoming emails, and the BellCore Video Recommender. From the mid 90-ties, the topic received a growing academic involvement with important study groups building recommenders and sharing both findings and results to the research communities. Grouplens, a research lab of the Department of Computer Science and Engineering at the University of Minnesota, developed an early engine to identify UseNet articles. Jester, a joke suggestion application was developed by UC Berkeley and Ringo, a music recommender was the work of Firefly Network, a joint venture of MIT Media

Lab and Harvard Business School associates. Since the mid 1990s, the topic grew into a more independent research area, characterized by a steep increase in literature on the subject. In academia, recommender systems became an important area, with courses being integrated in curricula and conferences or workshops being devoted to the subject, the most known being the ACM Recommender systems (RecSys). Also special interest groups were formed in already established gatherings, Ricci, Rokach & Shapira (2011). The fact that some of the big research labs opened their data, following their findings, which was the case for both Jester and MovieLens, boosted research activity and debate in this field. Nowadays, the algorithms are widely applied and form the engine for numerous item suggestion mechanisms in commercial sites such as Amazon or YouTube.

In terms of literature, the contributions to academic research output is substantial. A google scholar search on the tag "recommender system", discarding other relevant terms such as recommendation or prediction, provides over 40.900 hits, Web of Science generates over 2.700 scientific contributions upon entry of the keyword. Multiple manuals, surveys and books are available. Here, the book of Ricci, Rokach, Shapira & Kantor (2011) stands out, as do the reviews by Adomavicius & Tuzhilin (2011) and Burke (2002). Traditionally, the techniques underneath recommender systems are divided into 3 segments: content based algorithms, collaborative filtering techniques and hybrid systems, the latter mainly combining elements of the two first. They will be reviewed in this chapter along with social recommenders. These are as such not a technically separate type but rather an enhanced form enriched with an additional social dimension. Overall, the recommender theory literature is typified by the fact that the nature of the study objects are mainly creative products such as scientific papers, books, music or movies. For that reason, a thorough study of the topic is particular inspirational to understand how, from a informational technical point of view, this research branch incorporates items with continuously varying features and makes predictions on them. Movies in particular have been given a prominent status in the recommender literature. As will be shown, quite a number of techniques were developed, supported by big sets of movie data, coming from Netflix, EachMovie, MovieLens or IMDB.

3.2 Content based systems

Content based recommender systems generate item suggestions starting mainly from the person's individual preference pattern in relation to the items and features attached to them. The predicted preference is based on similarity between the object's attributes. Ekstrand et al. (2011) refer to this as item-to-item correlation. Content based systems have their origins in information retrieval. It accords to finding ways of personalizing a person's content and keeping a history of it. The user profile is a structured representation of his/her interests. The features of "new" objects are matched against attributes of the user profile in order to build a judgement on the individual's potential interest. Technically, Lops et al. (2011) distinguish three components in a content based recommender: a content analyser, a profiler learner and a filtering component.

Content analysis is grounded in information theory. It relates to connecting the proper features with the items. Both are transformed and connected in a structural form, mostly in vector space models. Content based systems require domain knowledge. In some cases, like movies, objective and publicly available features are extracted such as actors, directors and/or genre characteristics. Still, the processing can require big data manipulations in order to structure and connect items to features in the data space. In other instances, the content base can require more intensive acts of information retrieval. This happens in cases where the content is textual, such as film reviews or tags, forming an unstructured linguistic corpus. Then, the connection is made after semantic analysis and the selection of the main features is based on measures such as term frequencies.

The second step is learning the user profile. This is usually done through machine learning techniques, using data on consumer behaviour exposed in the past. A user profile is a representation of what is of interest to that person. The profile weights the importance of each feature to each user, which is learned by collecting feedback, either implicit or explicit. Implicit methods collect opinions indirectly by

monitoring certain actions such as search or click behaviour, bookmarking, buying or downloading. In case of explicit feedback, the user identifies how relevant an item is. Consumers can reveal relevance or irrelevance by means of a like/dislike button, translated into a binary scale, or they give explicit ratings, which are mapped into discrete numerical profiles. Also textual comments are often processed into positive or negative classification. Movie recommender systems, such as MovieLens, tend to work with a five degree ratings system. Starting from a training set of item/rating pairs, a profile is created applying learning algorithms. The estimated function generates for each item the likelihood of a person being interested in it. As tastes change, the profile needs a periodical upgrade.

Finally, the filtering component refers to separating useful from nuisance data and installing a strategy to rank new items according to their relevance when balanced against the estimated user profile. Recommendation can be computed in ways, similar to collaborative filtering discussed in next paragraph, using techniques such as k-nearest neighbour, where similarity between items is based on the resemblance of features according to distance measures such as cosine values. The rating score given to a user for a new item is the weighted average of items where the feature set is nearest, or most corresponding to the one under consideration. Sometimes, more sophisticated data mining techniques are employed to filter relevant items, such as naive Bayesian classifiers or decision trees. Bayesian classifiers, are probabilistic techniques, that allow to make a division into different liking sets, mostly two. They are estimated on a training set including features and past decisions by users, where then the posterior probabilities are calculated that are reused as suggestion rates. While the k-nearest is well suited to evoke short run suggestions, classifier better capture long-run interests.

Content based recommender systems have some advantages compared to collaborative methods discussed in the next section. As explained, they rely mainly on information about the targeted user and the item, without having to identify users with similar tastes. The systems are transparent, being a manifestation of the im-

portance of features for items and of items for preference. The features isolated to predict are easy to explicit. The cold start problem, the issue of not being able to provide a recommendation due to insufficient likes, which characterizes some other techniques, is of less importance. Previously unrated items can be recommended based on feature similarity only. However, a sufficient amount of ratings has to be at hand in order to perform adequate profiling. Moreover, the fitting of a profile often makes it over-specialized, favouring the habitual and discarding the unexpected. Persons are pointed to similar items rather than being suggested innovative goods. This is known as the so-called serendipity problem. Moreover, content analysis can be complex and a substantial knowledge base is needed. Making a liaison between objects and features may be done both manually or automatically. The first asks for prior expertise on the field, automated methods, such as word frequency extraction, do not always offer sufficient diversification. This can be the case when feature extraction is used to recommend textual representations such as poems or web pages. For movies, feature information is often commonly available and agreed upon. They include descriptive characteristics such as actors, directors, genre.

3.3 Collaborative Filtering algorithms

Content based methods are individually orientated and rely on the content or features of the creative product. In contrast, collaborative filtering techniques are intrinsically social and make predictions using minimal information about the item. Recommendations come from looking at the behaviour of other persons taking part in the system. The idea is to match together people expressing similar tastes, which can be thought of as a peer group or taste segment. Calculating the predicted rating about a new item starts from expressed likings by similar users or similar items by other users, Herlocker et al. (1999), Adomavicius & Tuzhilin (2005). The methods are founded on ideas of common expression of judgement or opinion, thus leaning on a joint behavioural component. Similar behaviour in the past is expected to hold in the future and agreeing on quality regarding a number of items implies agreeing

on others, Ekstrand et al. (2011).

Consider a set of persons $\iota = \{i_1, ..., i_I\}$, a set of items $\chi = \{x_1, ..., x_X\}$ and a rating function $\rho : \iota \times \chi$. To predict, one wants to learn a mapping $S : \iota \times \chi \rightarrow \rho$ where for each user $i' \in \iota$ the expected reward is maximized.

Collaborative filtering, first described as a methodology by Goldberg et al. (1992), is traditionally split into two approaches, namely neighbourhood and method based systems. In both instances, the collective user-item-rating triplet, users i gave rating value r to item x, is stored in a training set. The neighbourhood mechanisms make an immediate prediction from that training set based on similarity between users or items. The model based systems will use the information in that set to train a predictive model which is used to generate a recommendation. The neighbourhood approach is often defined as memory based, because they use the entire user-item rating matrix each time a recommendation for a new item is made. For the model based recommenders, the matrix is only used for the purpose of model fitting, not at the prediction stage, Kagie et al. (2009). In terms of mathematics and algorithms, the first approach relies more on clustering techniques, the second, as will be shown, has a closer association with factorization techniques.

Neighbourhood based recommendation entails the researcher uses the information of someone/something closely related. A detailed overview is offered by Desrosiers & Karypis (2011). User based collaborative filtering predicts the rating for a new item by a user as a function of the ratings of like minded users, so called peers, or similar items others have opted for. Similarity or closeness is expressed in terms of distance, using cosine measures or a Pearson correlation coefficient. It can start from either the user or the item. A K-nearest-neighbourhood based rating prediction delineates the set $\nu(i)$, the K-nearest neighbours of i or the k users with the highest similarity to i. Within that set, one only considers the persons having rated item x, which defines the subset $\nu_x(i)$ to define the prediction function as:

$$r(i,x) = \frac{\sum\limits_{j \in \nu_x(i)} similarityweight(i,j) r(j,x)}{\sum\limits_{j \in \nu_x(i)} |similarityweight(i,j)|)} \tag{3.1}$$

One of the earliest and prominent examples was the work issued by Konstan et al. (1997) of the Grouplenz team with an application on usenet news item recommendation. The item based recommendation works opposite and looks at the ratings of similar items. Take the set $\nu_i(x)$ consisting of items rated that are most similar to x. One potential formula to calculate the expected rating:

$$r(i,x) = \frac{\sum\limits_{q \in \nu_i(x)} similarityweight(x,q) r(i,q)}{\sum\limits_{q \in \nu_i(x)} |similarityweight(x,q)|} \tag{3.2}$$

Nearest-neighbour mechanisms are popular in use because they are fairly easy to implement while still generating efficient predictions. Moreover, the applications are easily scalable and are stable in the sense that addition of a user or item will not highly alter the predictors. Unlike the content based methods, they do not need specific information on the characteristics attached to the product but are based on a minimum of user-item-rating input. More than content based models described above and the model based collaborative filtering discussed in the next paragraphs, this type of recommender system proves able to face the serendipity problem as it might suggest items on the grounds that someone close to the user's profile esteemed it high, even if the item does not resemble past choices of the person. The method is therefore better in generating "out of the box" or "out of genre" predictions. This holds stronger for user based than for item based collaborative filtering.

One of the flaws of the method is that rating information might be insufficient as a determinant for equality in taste. Certainly when the number of inputs of a user is small, distance measures are potentially biased. In aggregated form, dealing with a large number of users that rate only a limited number of items means that the co-occurrence matrices become very sparse. A potential solution is to fill the

missing data, using averages, Breese et al. (1998), or by using content information, Good et al. (1999), Konstan et al. (1997), Melville et al. (2002). Moreover, it is questionable whether the same rating behaviour really expresses similar interest. The allocated rating is dependent on habit, tolerance or strategy. This issue can be technically addressed, through heuristic normalization, Jin et al. (2004). Yet the final outcome stays dependent on limited rating information. Finally, one major drawback is that, in order to recommend an item, at least one person must have shared an opinion. This is known as the "cold start problem", which has proven a major issue for collaborative filtering algorithms and one of the main reasons to re-include content or to move to more hybrid techniques.

3.4 Model based recommender systems

Item-user vector spaces are often high dimensional. Because of limited rating activity by some users, rating matrices are sparse. One could state there is a level of redundancy in the data because it might be possible to reduce dimensionality by making divisions into classes of similar profile, Ekstrand et al. (2011). If so, it makes sense to try to identify sets or topics over which persons express interest and attach items to topics in accordance to relevance. Latent factor models do aim at uncovering latent features trying to explain underlying patterns driving the observed variable, which in case of recommender systems are often user ratings. They explain the observed patterns by a smaller number of typical patterns expected to be underlying to the observed data, Hofmann (1999). Examples include singular value decomposition, techniques inspired by probabilistic latent semantic analysis and Latent Dirichlet Allocation, Koren & Bell (2011).

3.4.1 Singular Value Decomposition

The SVD method was adopted in recommender theory by Billsus & Pazzani (1998) and Sarwar et al. (2000), but is irrefutably associated with Koren & Bell (2011), as matrix factorization techniques were introduced into their approach that settled the Netflix competition. The method leans on the rule that a $m \times n$ dimensional real matrix R can be factorized into 3 parts as $R = U\Sigma V$, where Σ is a $m \times n$ rectangular diagonal matrix with the n elements of the diagonal being non-negative real numbers, called singular values. The matrix U is of dimension m and V of dimension n. The dimensionality reduction is installed by using a truncated representation of Σ, retaining the k largest singular values, denoted by Σ_k. Represented in its rank-k approximation, it can be read as a model of topic preference relevance, where the rows of U represent the person's interest in the k topics and the rows of V mirror the relevance of the items in the topics, Ekstrand et al. (2011). The diagonal elements are considered as weights, reflecting the individual's influence of a topic on user-item preference. The predicted rating of an item is the weighted dot product of both the user-topic interest vector and the item relevance vector. Apart from dimensionality reduction, the model becomes more stable. By dropping the smallest singular values, small noise is eliminated.

The latent class models are generally shown to outperform nearest-neighbours algorithms and to deal with some of its inadequacies. However, the SVD decomposition is undefined when rating information is incomplete, which implies that methods of imputation need to be used to fill the missing data. Dependent on the method used, this can cause over-fitting. Moreover, the expansion of the data volume impacts computing use. In later work, Koren and Bell improved the SVD method, into what they called SVD++, integrating user feedback and time changing factor models, Koren & Bell (2011). While expensive in terms of offline computation steps, the methods showed high potential for prediction. Also the latent grouping idea proved valuable in dealing with large dimensional movie datasets. However, whilst SVD partly deals with the sparsity problem by reducing dimensionality, it remains an

issue when the number of items is high. Collaborative recommenders need that the density of users is relative high compared to the number of items. Also, it is important that the amount of items stays relatively stable since older items might lose relevance. Collaborative filtering works best for users fitting into a segment and holding strong neighbours, peers with high resemblance. They function less well for cases situated at the border of user groups, the so called 'grey sheep', Burke (2002).

3.4.2 Probabilistic Latent Class Model

The theory of Probabilistic Latent Class Collaborative Filtering (PLC-CF) is connected to the names of Hofmann and Puzicha, who were inspired by theories on latent semantic analysis (LSA), Deerwester et al. (1990). LSA is an indexing and retrieval method to analyse patterns of similarity between documents based on the words they contain. The analysis starts from making a term-document matrix. Hofmann & Puzicha (1999) transform the probabilistic semantic model into a recommender model applied on movie data. In the so called Aspect Model, they start from a dyadic data setting where the observation is the joint occurrence of item and person and a latent class variable $c \in \chi = \{c_1, ..., c_K\}$. The user-item pair (i, x) is assumed to be generated independently, conditioned on z. Then, the model can be defined as:

$$P(i, x) = \sum_{c=1}^{K} P(c)P(i|c)P(x|c) \tag{3.3}$$

where $P(c)$ are the class prior probabilities, and $P(i|c)$, $P(x|c)$ are the class conditional multinominal distributions. Using the identity $P(c)P(i|c) = P(i, c) = P(i)P(c|i)$, equation 3.3 can be reparameterized to:

$$P(i, x) = P(i)P(x|i) \tag{3.4}$$

where

$$P(x|i) = \sum_{c=1}^{K} P(c|i)P(x|c) \tag{3.5}$$

Equation 3.5 expresses the conditional probability of an item given a user as a convex combination of the aspects or factors $P(x|c)$. It suggests that the behaviour of an individual can be represented by a combination of "typical preference patters". Hofmann & Puzicha (1999) explicitly state that this does not imply that a user belongs to a group or cluster. People might be probabilistically attached to different segments. The model is contrasted with two sided clustering where each person is denoted to exactly one group and an object to one segment.

In a second stage, the authors add the person's valuation r, in a binary way by means of a positive and negative rate, $r \in (-1, +1)$. They present multiple dependency scenario's between rating, user, item and class. Assuming that r is conditionally dependent on x and c, translates into the following equation:

$$P(i, x, r) = \sum_{c=1}^{K} P(c)P(i|c)P(r|x, c) \tag{3.6}$$

Hofmann & Puzicha (1999) introduce the EM-algorithm to fit the parameters by means of maximum likelihood estimation, which generates estimated class dependent probability functions for users and items. The testing of the model is performed on a user rating set of EachMovie, with around 3 million preference inputs that were converted into positive-negative evaluations. Compared to two sided clustering, the aspect model performs substantially better in terms of the perplexity criterion used as comparison by the authors. More importantly, the results show that movies are not allocated with probability one to one segment, but reappear with altered probabilities in each latent segment. It makes this type of model more interesting when thinking about how users shape their decisions. In a later paper, Hofmann (2004) uses prediction statistics such as the Mean Absolute Error (MAE) and the root mean square (RMSE) to show that probabilistic semantic analysis achieves

accuracy gain over a baseline predictor of almost 18 percent and about 13 percent over Pearson correlation methods.

While the interpretation is very similar to that of SVD - the left and right eigenvectors can be linked to the factors $P(i|c)$ and $P(x|c)$ - Hofmann (1999) highlights the differences. Probabilistic latent semantic analysis, bearing on the likelihood function of multinomial sampling aims at maximizing the predictive power, contrary to SVD which is defined in terms of minimizing errors. The estimated functions provide probabilistic interpretations of preferences over topics and items in topics. Users are represented by mixtures of preference profiles or feature preferences. It assumes that the preference of an item is mainly determined by the latent class he/she is connected to, an item is of similar importance to all users who expressed their liking of the profile. The attraction lays in defining prototypical profiles that are not deterministic. Moreover, the probabilistic structure of the theory makes it practically and conceptually more straightforward to separate user decision from rating, both defined in terms of conditional dependencies on items/classes.

3.5 Hybrid algorithms

Triggered by some shortcomings of content based and collaborative filtering systems, a substantial amount of research went into combining elements into what is referred to as hybrid methods. Burke (2002) provides a comprehensive overview of hybrid recommenders, mainly grouped by the way they incorporate the different approaches. Weighted and mixed recommenders join scores to calculate one new value (weighted) or present several (mixed). Switching systems apply the algorithm expected to provide the best results. If a user has not sufficient entries to generate a reliable outcome under collaborative filtering, then the optimal rating is taken from content based calculations. Cascading systems use the outcome of one system as input for another algorithm. An example of a switching system is the MoRe recommender, authored by Lekakos & Caravelas (2008). They use a version of

the MovieLens dataset which they augment with content based features taken from IMDB, including genre, cast, director, writing credits, producer and plot keyword. If collaborative filtering works ineffective because the number of ratings falls under a threshold, the mechanism triggers a content-based prediction. CinemaScreen by Salter & Antonopoulos (2006) is an example of a cascading system. Starting content based, links with other movies are captured through collaborative techniques. Features are collected of the predicted item. Then the collaborative rating is calculated, which value is transposed to all features. After several ratings, the average appreciation for each feature is learned, then used for new items.

Feature combination systems use collaborative information as an additional feature in content based techniques. Basu et al. (1998) match rating information with content information. They add genre (comedy, action and drama) to the rating-user pair as an additional collaborative feature. Their hybrid approach, applied on a movie dataset of 45.000 ratings by 250 users, improved precision compared to the collaborative method. Melville et al. (2002) create a pseudo rating user vector based on content based methods. Then they perform collaborative filtering on those vectors. Because the pseudo matrix is a full matrix, the authors deal with the sparsity issue and tackle the first rater problem. Using data from EachMovie, they establish that the hybrid recommender performs better than the content based, collaborative and a naive hybrid. The naive hybrid approach takes the average of the ratings generated by the content-based and the collaborative predictor. Several research papers propose hybrid algorithms applied on the MovieLens data set. Christakou et al. (2005) suggests a semi-supervised clustering method. They automatically retrieve the synopsis of the movie they complete with director, actors and script writers. Movie summary and title are textmined and used as content. Another application is provided by Gunawardana & Meek (2009), using tied Boltzmann machines, which are a form of stochastic neural networks. It allows to model the joint distribution of binary variables, in this case, whether or not a user has acted on an item of interest. The parameters are tied with a feature vector, representing genre, actors and the actor/genre combination. The estimated parameters reflect how actor/genre contribute to the overall popularity of a movie as well as how much

the co-presence of certain features determine movie choice. However, while untied Bolzmann approaches perform clearly better than Pearson correlation and item-item based collaborative filtering techniques, the results are less pronounced for the tied solution. Here, the model performs slightly worse when genre or actor information is used in separate and better when joined. The two described MovieLens papers prove, as was established by Basu et al. (1998), that the hybrid approach improves precision compared to item-item and Pearson correlation. This is an overall constant in the conclusion on testing hybrid movie recommenders, that they generally outperform collaborative filtering or content based methods used in isolation. On the downside, they are more complex to implement at a larger scale. Moreover, the minimal information setting that was so typical for collaborative filtering gets lost. That is why, in practice, the pure collaborative filtering algorithms remain popular, despite research providing better hybrid alternatives.

Most hybrid systems combine elements of content with nearest neighbour algorithms. Few start from a latent class approach. An exception is the paper by Symeonidis (2008). The author constructs a hybrid profile of a user, bearing on Latent Semantic Indexing to reveal the dominant features. Also in the work by Gantner et al. (2010), a matrix factorization method is optimized to generate Bayesian Personalised Ranking. One of the aims is to deal with the cold start problem. The authors introduce a mapping of item and user attributes to the latent classes of a matrix factorization model. Movies are ranked according to their probability of being viewed/purchased. The attribute space is mapped into the factor space according to different methods, a K- nearest neighbour mapping, a linear mapping and an optimized regression mapping. The attributes set consist of genre, included in the MovieLens datasets, and directors/actors connected via the IMDB database. There are quite some outcome variations induced by difference in mapping methods and results are ambiguous. Overall, the feature mapping factorization model seems to perform better than does the plain collaborative variant based on cosine similarity. Including genre and director generates comparable results. Again adding both performs best. When considering a high dimensional feature set, including actors, the results are less in favour of the mapping alternative. Including actors is inferior

to user-user cosine similarity while the mixed genre/actor outperforms the collaborative filtering solution. This is explained by the higher sparsity. The sparsity of the MovieLens genre data is around 90 percent where that of IMDB actor/director sets reach figures close to 100 percent.

An important hybrid probabilistic latent class contribution is proposed by Kagie et al. (2009). The paper builds on the probabilistic latent class collaborative filtering approach of Hofmann & Puzicha (1999), Hofmann (2004). Additional feature characteristics are included by converting the probabilistic model into a latent class regression recommender system. Inspired by equation 3.6, the authors endow $P(r|c_k, x)$ with a Gaussian distribution:

$$P(r|c_k, x) \sim P(r|\mu_{k,x}, \sigma_{k,x}) \sim \mathcal{N}(r|\mu_{k,x}, \sigma_{k,x}) \sim \frac{1}{\sqrt{2\pi}\sigma_{k,x}} exp[-\frac{(r-\mu_{k,x})^2)}{2(\sigma_{k,x})^2}] \quad (3.7)$$

There is a class specific mean and standard deviation, shaping the probability of a rating given the user's choice of item/class. The marginal density, denoting θ as the estimation parameters, then is:

$$P(i, x, r|\theta) \propto \sum_{k=1}^{K} P(c_k|i)(r|\mu_{k,x}, \sigma_{k,x}) \quad (3.8)$$

In the latent class regression recommender system, the means $\mu_{k,x}$ for each latent class c_k is represented by one vector of regression coefficients b_k. Also, the item-specific standard deviations $\sigma_{k,x}$ are replaced by one class specific standard deviation σ_k. This substantially reduces the number of estimation parameters, that become independent of the number of items. The new rating functions now look as follows:

$$P(r|c_k, x) \sim P(r|z'_x b_k, \sigma_k) \sim \mathcal{N}(r|z'_x b_k, \sigma_k) \quad (3.9)$$

where z_x is the vector containing the characteristics of item x. The paper offers a likelihood estimation, fitted with EM. The experiment is performed on Netflix data connected to IMDB characteristics. They reduce the Netflix set to users with more than 50 ratings and take a random sample of .25 percent. As features, they include the genre variable and the 100 most used keywords, such as provided by IMDB. The resulting test set consists of 19.105 ratings, inputed by 741 users on 6.305 singular movies. The results show that PLC-CF scores marginally better than a Pearson based collaborative method. The latent class regression model does well, improving baseline and content based substantially, but does about 3 percent worse in terms of MAE than collaborative filtering. However, the latent class regression model has the advantage that new items can be recommended since the estimated regression coefficient can be used on each item and user. Moreover, the method makes a division into several user segments, with clear indication what characteristics drive the ratings within that group. For targeted commercial purposes, it is more clear who to address.

Finally, Agarwal & Chen (2010) use a Latent Dirichlet Allocation Model (LDA), comparable to the one that will be used in the next chapter. LDA is in many ways similar to PLC-CF, founded on the connection of individuals to latent topics in a probabilistic way, and of items to topics. LDA is however more general than PLC-CF in the sense that it also provides a generative probabilistic model at the level of the topic distribution. The mixture proportions are endowed with a Dirichlet prior, Blei et al. (2003). Topic multiplicity is an interesting feature of both those models because individual's preferences are expressed in terms of their probability of being attached to one of the estimated layers or topics, unlike being attached to one particular group or cluster. At the same time, the items are conditional on the topic. Using the genre data of MovieLens and a number of user features like gender, available in the 1M set, Agarwal & Chen (2010) show that the feature LDA method outperforms other methods, including mere factorization, in terms of RMSE, however, no further details are given on the nature of the topics that appear.

3.6 Social recommendation systems

Methodologically, this group of recommender systems cannot be considered as really distinct. As will show, the modelling underneath is similar to the options presented in previous paragraphs. However, the social web has enhanced the ability of users to provide content and form networks. Knowledge and opinions are shared through wiki's and blogs and interactive interfaces are developed to annotate items. This pool of additional information is then re-integrated into recommendation systems to improve performance or achieve better profiled services. One type of user content that has been the source of many investigations is keyword addition or tagging. Social tagging systems lower the barrier to create and share light weighted metadata schemes. This has proven especially useful for annotating creative content such as literature, music, photos or movies. Later in this chapter, the potentials, advantages and drawbacks of tagging will be discussed in more detail. Now, a short overview will be given on how, in research, tags are used as an additional dimension to improve the recommendations of items.

By means of a tagging system, a user can annotate a particular object. Each action creates a three dimensional context with a link <user,tag>, a link <item, tag> and <user, item> . However, since there is an implicit involvement of a user tagging an object, this can be represented in terms of a three-partite graph, Milicevic et al. (2010), Tso-Sutter et al. (2008). When used for research, the common relations that are considered are all tags related to an object or to an individual. Define the set of tags $\omega = \{w_1, ..., w_W\}$. When tags can be shared and reused, they form a folksonomy, Vander Wal (2007), which in formal terms can be described as a tuple $(\iota, \omega, \chi, \Upsilon)$, Balby Marinho et al. (2011), where:

- ι, ω, χ are finite sets of persons, tags and items
- $\Upsilon \subseteq \iota \times \omega \times \chi$ is a ternary relation between them representing the tag assignments

From this

- $D_x = \{(w, x) \in \omega, \chi | (i, w, x) \in \Upsilon\}$ are the tag-item relations
- $D_i = \{(w, i) \in \omega, \iota | (i, w, x) \in \Upsilon\}$ are the tag-individual relations

Tag based recommendation literature isn't very extended, certainly not when it comes to movie recommenders. Yet, a number of tag based systems are worth mentioning in the light of the study performed in chapter 5 of the thesis. Conceptually, Pitsilis & Wang (2015) first separate between tag recommendation algorithms and tag-oriented resource recommenders. The first aims to ease the process of making annotations while for the second group, tags are functional to improve item prediction. Here, a further distinction is made between tag based and tag assisted algorithms. The first merely needs tags to make predictions while the latter involves a combination of tags and other variables such as rating.

An example of a tag-oriented resource recommender is presented by Szomszor et al. (2007) who study tagging data derived from IMDB, allowing users to add keywords to titles and thereby improve their search of movies. It is a free-for all system that is monitored only to prevent spam being added. Those tags are connected by the researchers to the rating/movie pairs listed in the Netflix database. As there is no direct relationship between the user and the tag, the authors define alternative user-tag clouds for various rating levels. Then, the expected rating of user i for movie x is calculated by comparing the set of tags associated with the new movie against the rating of the cloud where the number of keywords shared are the highest. Other important tag-based and tag-assisted examples are provided by Grouplens, Sen et al. (2009*b*), in what they refer to as tagommenders. This will be further discussed in the next chapter.

Studies related to other resource types include the paper by Stoyanovich et al. (2008) on tagging for del.icio.us, an online social bookmarking service. Here, similarity in

tags is a manifestation of similarity in the objects and is as such an additional feature of the movie. Also Liang et al. (2010) use K-nearest neighbourhood theory to determine a set of peers which are determined by their similarity in tag profile. They apply it to CiteULike and Amazon data. Tso-Sutter et al. (2008) incorporate tags in a collaborative filtering setting, but do so by augmenting both user and item matrices such that the item matrix is the horizontal augmentation with tags added to items and the user matrix the vertical augmentation with tags added to users.

Distinct and important in the light of this study are the probabilistic approaches. Inspired by the work on probabilistic latent semantic analysis by Cohn & Hofmann (2000), Wetzker et al. (2009) work out a hybrid PLC based approach where latent class probabilities for items and tags are estimated in parallel, also applied on bookmarking for del.icio.us. Conform the Aspect Model described above, co-occurrences of observations are associated with hidden classes. The relationships are specified twice, for both users and tags:

$$P(x|i) = \sum_{c=1}^{K} P(x|c)P(c|i) \tag{3.10}$$

$$P(x|w) = \sum_{c=1}^{K} P(x|c)P(c|w) \tag{3.11}$$

Then both models are combined based on a common factor $P(x|c)$ in a maximum likelihood function, where f represents the co-occurence counts and α is a predefined weight:

$$L = \sum_{x}[\alpha \sum_{i} f(x_x, i_i) log P(x_x|i_i) + (1-\alpha) \sum_{w} f(x_x, w_w) log P(x_x|w_w)] \tag{3.12}$$

A comparable approach is Latent Dirichlet Allocation, which will be discussed fur-

ther in chapter 5. Like in PLSA, the tags are made dependent on topics or hidden classes. Pitsilis & Wang (2015) elaborate on the ideas of topic based clustering founded on semantic similarity. In their study, the semantic distance is determined by a metric introduced by Resnik (1995), based on the notion of information consent and looking at the place of the nodes in the hierarchy of the taxonomy tree. This is derived from the lexical database wordnet. To cluster the terms, the authors use an affinity propagation algorithm. The similarity of users, needed to recommend, is based on the common clusters the tags of two users belong to as well as on their annotation contributions. It is beneficial to users putting more effort in tagging objects. The authors present the topic orientated approach as the way forward, stating that they believe that clustering using semantic similarity offers great potential as it is sufficient to cluster only once. They argue that if a person uses a tag many times or the tag is highly connected to an item, this might be a good ground for recommendation. More importantly, two persons with different tags might find themselves connected through their subject. Till now, LDA has been used more to improve direct tag recommendation rather than to recommend items. Krestel et al. (2009) serve as a counter example based on tagging for bibsonomy.

3.7 The decision making factors in recommender theory

The main objective of recommender systems is accuracy and the majority of scientific literature on the subject presents predictive improvement as established result. Underlying views on human behaviour or human cognition remain rather unspecified, but from the different algorithms presented in previous sections, it is possible to distinguish two main dimensions, namely peers and technical features. Collaborative theories are based on the assumption that for each individual a number of peers or like minded individuals can be identified. Once assigned, the past behaviour of that peer group is to some extent predictive for the future behaviour of the targeted individual. On the other hand, content based and hybrid recommenders work with features which are naive in the sense that they are made available in an automated

way from structured online repositories, Basu et al. (1998). They consist mainly of technical, objective characteristics such as cast or director of the movie. Another common feature is genre, which cannot be seen as merely technical nor objective. As will be further discussed, it is the result of human categorization.

The recommender theory overview made clear that the nature of the features' impact is not unidirectional. Examples of hybrid systems showed that including various features in combination improves performance, however adding a single variable offered less conclusive results. Gunawardana & Meek (2009) successively added genre and director and concluded that only the joint integration was effective. The same holds for the study of Gantner et al. (2010), be it here for the combination genre and actor. In a feature comparing study, Alspector et al. (1998) apply a decision tree approach to extract the most relevant movie characteristics, relevant here defined as being given the highest rating by users. The factors selected were category, MPAA rating, Maltin rating, Academy Award, length, origin and director. Distinct from recommender algorithm literature, the study adds expert ratings and awards. Director appears to be of the highest influence, but the variable performs weak when applied on the out of training set as it cannot be connected to just any movie. This relates to a major issue of the feature based approach; namely that matrices get sparse due to a very long list of potential elements, actors or directors, connecting only a few times to the sample of movies. Nasery et al. (2015) survey more directly the preferences over various movie factors, concluding that cast and external ratings are the most influential factors pointed out directly by movie consumers, followed by director and genre.

Social recommenders provide explanatory variables of a more subjective and diverse nature. Opinions posted or tags added are as such not decision making factors to buy creative goods. The motives of their creation are primarily annotation and ex post sharing of meaning. That said, extraction of the main keywords reveals the elements users labelled as important in their movie choices and why they liked or disliked watching it. Different dimensions can be discriminated, technical features such as

actor or director, semi objective variables such as genre, and elements that signal the experience good nature of creative commodities. The latter can be represented by users expressing the influence of external reviews, expert ratings or awards given to a movie. In the light of designing an explanatory decision model for creative goods, in casu movies, tag and opinion mining seem promising tools. They represent the direct description of the nodes that, when tight together, give a first glimpse of the motives behind the decision to watch that particular movie or not. The potential of tags as explanatory variables in a decision making model will therefore be further explored in this thesis. Results will be mainly compared to a baseline result using genre. Genre, while not unproblematic as a variable, is selected because it is a variable returning in most of the content based models as well as in some of the box office studies mentioned in the introductory chapters. Both variables will be discussed in more detail in the next two paragraphs.

3.7.1 Genre as a movie classifier

Genre is one of main denominators to devise and diversify tastes for movies. It is one way of classifying films into items of type similarity, types that audiences and filmmakers recognize by their recurring conventions, Bordwell & Thompson (2008). Occasionally, producers label genres, but some film genre classification systems dominate the sector, such as that of the Internet Movie DataBase (IMDB).

The genre classifier is as such not unproblematic. It is vague and abstract as a concept. Definitions tend to refer to similarity between items which can run over a mixture of generic elements that are part of a tacit knowledge base, not distinct or specified. The determining elements can be story content, intention, media type, cycle, technical process, location, series, purpose or many more. Features are not unique to label the genre type. Then, the prominence of a particular characteristic plays a role. Not only does a typical genre covers a mixture of characteristics, a genre type appears indefinitely divisible into many lower order subcategories and

the conceptual borders are ever shifting, Abercrombie (1996). In the light of the infinite genre division and lack of contours, the question can be raised if the genre concept is anything more than an analyst construction, Stam (2000). Genre is the result of conventions and boundaries changing over time with evolving collective experience. Even under the same constellation of constitutive elements, the genre categories can be interpreted in different ways under different cultures, countries or time periods.

Allocating an item to a genre is an act of recognizing resemblance, acknowledging that a predefined idea or prototype constitutes the nature of a genre category. In that sense, genre links consumers and forms a base for consumer targeting, Hodge & Kress (1988). The elements that determine a genre type can be seen as a knowledge base. It is a reference frame reducing complexity in identification, sensemaking and selection, where the commonly shared features are gradually learned. Experiencing examples is one way of learning the components that constitute the genre type and acquiring the ability to make the identification. In that respect one could speak of cultural capital formation. Neale (1980) defines genre as a set of expectations which helps enabling judgements and formulate prediction. The attachment of a movie to a category induces prior expectations and can be seen as a kind of tacit contract between producers and consumers, a shared code between makers and interpreters, Chandler (1997). It increases the efficiency of communication because some knowledge is already inherent in the expectations within the genre. Tolson (1996) states that genre is a way of categorizing that mediates between industry and audience, a practical devise to link the production to the expectation of the customers. It can be exploited in the creation of an audience staying loyal to a particular genre or subgenre, a way of getting grip on demand, Neale (1980). Buyers hold a default expectation from which to start and which might be met or challenged. Genre typification induces the hypothesis that audiences, in their interaction with the film and during the viewing process, attach meaning related to what is shown. Provided also the elements of common knowledge, these form the main reference for advertising and review. Producers can use the prominent features that made the category successful to pass them on to new creations and use them as a base for promotional

activities.

One genre classification used often in scientific literature is the one initiated by the Internet Movie DataBase (IMDB). This is an online database founded in 1990, providing information on movies, games and television series. The movie information is enriched with data on plot, cast, bio's and financials. In 1998, it was taken over by Amazon. The site has an important rating service and message board, but is as such no recommendation service. IMDB data did however serve as the main source of information to include features into studies, be it genre, actors or directors. The content is largely provided by volunteers, however, the addition, deletion or modification of the data occurs by the site's staff only and goes through a number of consistency checks. The database maintains a classification of movies into 28 genre labels: Action, Adult, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film Noir, Game-Show, History, Horror, Musical, Music, Mystery, News, Reality-TV, Romance, Sci-Fi, Short, Sport, Talk-Show, Thriller, War, Western. Because of the popularity of the site, the leverage of the data harvested by Amazon and the extensive use of the data series for analytical purposes, this genre classification is often the norm. It has to be argued however that the perspective of the IMDB labelling is mainly story content based, discarding other dimensions such as time period (thirties movies), technical process (3D movies), series (Bond movies). The audience line is only partly represented with family/adult movies, while kids, teens, man/women movies are ignored as separate categories. It is but one way of labelling the movies, which has been highly influential on the empirical results concerning the impact of genre on consumer behaviour, box office or branding.

3.7.2 The potential of tagging information

Tags are user generated keywords or phrases that enhance or label (mostly online) resources. They are a sort of soft metadata, namely data typifying items and serv-

ing as intermediate between individuals and the resource or system. Tagging came into practice in a Web 2.0. user driven landscape. Items can refer to any object like web pages, photos, music or movies. Tag systems refer to the applications or interfaces that allow agents to annotate resources. Well known tag based systems are del.icio.us (bookmarks), Lastfm (music), Flickr (photo sharing device) and CiteULike/Connotea (sharing of bibliographic references). Users can add tags, thereby annotating content and improve future navigation. When tagging is used as an individual annotation device, one speaks of personal tagging. The main motive here is structuring your own interest and information to improve future retrieval. Else, tagging behaviour can be driven by social motives, where users want to express themselves in a social context, Marlow et al. (2006). When tagging is performed in a social environment, one speaks of social tagging. Collaboration by annotating resources in the benefit of all improves search and tagging facilities. This can be done by suggesting already imputed keywords or by visualizing popularity under the form of tag clouds or tagometers. The ensemble of tags added by the users gradually evolves towards a vocabulary, made explicit to all, named a folksonomy, a term introduced by Vander Wal (2007) to express the organic development of the system.

Folksonomies offset taxonomies or controlled vocabularies that are maintained by experts. Contrary to formal taxonomies, they consist of terms in a flat space and lack an imposed hierarchy and lexical control. Collaborative tagging however has the advantage of being more inclusive and bottom up. It therefore more accurately reflects the population's conceptual model of the information, Quintarelli (2005). At the same time, it allows more flexibility in associating different terms or meanings to an object. Compared to taxonomies, Shirky (2005) argues that users move away from binary categorization into a probabilistic world where tags can be attached to several items and where you can say that a percentage of people provided the same tag to different items. They are flexible, given that, unlike experts, users do not necessarily need to know the ex ante structure of a taxonomy and can intuitively annotate new and unexpected contents. Moreover, tag systems can be installed at smaller cost as they are of a decentralized nature, thus more efficient than the labour

intensive controlled taxonomies.

Opposed to the opportunities tagging systems offer are the doubts expressed on the quality of the unsupervised entry of concepts. The act of tagging is a free expression by the user. It makes that questions can be asked on the nature of tags. A substantial amount of research has gone into analysing tagging behaviour and developing methods to improve interfaces aimed at augmenting the quality of the user's input or allocating relevance values to tags. As pointed out by Milicevic et al. (2010), tags possess as such little semantics and consist out of many variations. Different systems might use tags in different ways. Merging them does not necessarily allow us to attach similar meaning to terms.

Macgregor & McCulloch (2006) provide a comprehensive overview of the issues at stake, making suggestions on how to control for them. First, in terms of use and linguistics, the most prominent issues are those of synonyms and homonyms, pointing respectively to different words endowed with similar meaning and a singular word having multiple meanings attached to it. Also other lexicographic anomalies, plurals, multiple- or misspelling, multi-word-concepts, characterize tags databases and impose difficulties on their common use. Sinha (2005) connects tagging behaviour to the cognitive as a consecutive act of related category activation and decision what category is the right one. Resulting from a personal cognitive process, tag diversity can be aggravated by the lack of cultural consensus and by the will to make the keyword functional in the future. Additionally, terms connected to an item move along a continuum going from specific to general.

Golder & Huberman (2006) study the structure of tagging systems using data from del.icio.us. They discovered big differences in the frequency and nature of individual's use of the system as well as great variety in the growth rate of tags. Keywords might suddenly emerge, potentially signalling a trend. However, the collective behaviour of tags does not show a chaotic path, but converges to a stable pattern in which the proportion of each tag evolves to a fixed rate. As more bookmarks

are added, the proportion of a particular tag versus the total frequency of tags is nearly constant. The converging behaviour is explained by the authors in terms of the dynamic urn model by Eggenberger & Polya (1923). Underneath the behaviour might be imitation or shared knowledge. Indeed, tagging systems tend to suggest those keywords that are often used. It is considered to be social proof for individuals to make a choice based on what they think others have approved. The stabilization process was reconfirmed in a study by Millen et al. (2007), illustrating a gradual decline of new tags combined with a manifest tag reuse.

Apart from mere social mimicking behaviour, a second reason for stable patterns to occur is that the ideas represented in the tags are somehow stable and are supported by a common denominator. Kipp & Campbell (2006) illustrate that the distribution of tag occurrence is long tailed, with 30 percent of tags occurring only once. The long tail characteristics are not necessarily seen as negative, since they might represent emerging and innovative ideas within a small fraction of the population. Tagging is more than discovering information, but is a process of social learning by individuals who put effort in picking up terms that might be in the lower tail, Quintarelli (2005). At the same time, some terms do show obvious and frequent occurrence and common behaviour can be detected in them. Golder & Huberman (2006) explicit them as 1. Identifying what or who the item is about, mostly by using nouns at different degrees of specificity 2. Identifying what kind of item one is dealing with 3. Ownership identification 4. Refining already existing categories which can be by using numbers 5. Identifying qualities of characteristics, which is mostly by adding adjectives 6. Obvious terms of self organization and cataloging or task organization. Tags are personal and diverse but by its social component do also converge and some overall trends can be discovered in analysing them.

There are different approaches to deal with some of the shortcomings of tagging systems, Milicevic et al. (2010). Some are oriented towards improving tag literacy by users or by working through tag suggestions, in much the same way as item recommenders, Guy & Tonkin (2006). Others involve the more advanced use of

natural language processing and data mining techniques, Shepitsen et al. (2008). Clustering, as a case in point, can partition tags based on their co-occurrence, that way aggregating redundant tags while simultaneously detecting the combined trend. Methods to improve tag quality for movies have been intensively studied by the Grouplens team. Their findings are treated in the next chapter.

3.8 How recommender theory can inspire consumer modelling

This chapter gave an overview of an important current research topic in computer science dealing with the question of how the stream of information provided by users can be used to improve service levels or to recommend an item that is likely to be of interest. The subject of recommender systems has been widely investigated in research literature, has now become part of academic curricula and specialized conferences are held on the subject. Moreover, the algorithms developed are applied world wide on real life applications.

There are several reasons why research of recommender systems is of potential interest for researchers studying economic modelling of consumer choice for creative goods and more specifically for the study of movie demand. First, most recommender systems, while often acting on expected ratings, present themselves as predicting users' preferences for goods. The items of interest are dominantly creative products, be it music, literature or films. As was argued, movies take a prominent position in the study of recommendation systems. The improvement of collaborative filtering systems was highly driven by two big initiatives, the GroupLens team that initiated the MovieLens recommender interface, followed by a stream of algorithm studies produced by GroupLens and many others. Secondly, the Netflix competition led to the collaboration of several academic and private teams, who joined expertise and improved techniques on dimensionality reduction. The probabilistic latent variant, driven forward by Hofmann and inspired by Latent Semantic Indexing, was initially

applied on a set of Moviedata coming from EachMovie.

The fact that movies are a highly inspirational commodity is partly due to its nature, being a multi-featured popular commodity where people show little reluctance to share their opinions and ratings. Their choice of one movie over another seems an act of short-run portfolio optimization. The high escalation of movie recommender research was also driven by the fact that both MovieLens and Netflix opened up their data, making it a free source for investigation to many. The combination with other open initiatives such as IMDB, a source of information gathered on a crowd input basis, made that choices, ratings and movie characteristics could be merged into rich augmented datasets. The data did not only elevate research in collaborative filtering algorithms they were initially intended for, but also the study of content based and hybrid systems. However, while those seemingly outperformed the collaborative techniques, they found more resonance in scientific literature than in real world applications.

Despite the common nature of their terminology, at first glance, economic theory and recommender system studies bear little resemblance. First, the datasets used are highly different. The data used in economic theory are mainly aggregated data explaining box office evolutions. In contrast, research in recommender theory leans upon user generated micro data, coming from individual's input of ratings, opinions or tags. At the methodological level, empirical studies on consumer demand for movies are based on econometric techniques, often single equation modelling. The recommender systems approach is largely based on big data analysis, where data mining and machine learning techniques are underpinning theory. Looking at the statistical methods applied, clustering and latent class techniques stand out, to group persons or items that are esteemed alike. Empirical economics works top down, starting with a testable hypothesis verified by the data. Recommender systems work bottom-up, by looking at the data first. Their approach is performance driven, aimed at achieving optimal prediction, not revealing the drivers that are underneath, discarding the premise that optimizing is at the core of consumer theory.

What is of interest in the recommender approach is that techniques are highly based on individual profiling, where economic theory leans upon aggregated consumer data, thereby holding the strong assumption of a representative consumer. This involves a typical decision maker who's behaviour is representative, or in a weaker version, that the difference in agents is not manifested in the sum of choices. Hence the aggregate is conclusive for the individual. Unlike the economic theory of movie consumption, recommender theory starts from the idea that individuals differ from each other, even when their behaviour is observed as similar to that of others and therefore can be grouped. Economic theory has always faced difficulties dealing with heterogeneity in preferences, which was partly a motivation for the development agent based theories which are in some ways more profound, as they integrate optimizing behaviour as well as the influence of the few on the collective. Information scientist deal with creative goods in a less constrained way, incorporating their characteristics as well as their co-use by agents, acknowledging heterogeneity. This doesn't mean that no pre-assumptions are imposed on consumer behaviour. One is that similarity can be represented by certain types of distance measures and that the degree of similarity can be used for the future choice a consumer will make. When based on the similarity of items, there is an innate assumption that intertemporal persistence can be observed in consumer behaviour: the type of creative good a person used to like is a predictor of what he/she will opt for next. This is not only the case for item based collaborative filtering, but even more so for content based system, that are founded on the idea of similarity between the content features of items consumed in the past and those currently recommended.

Decision making factors in recommender theory are dominantly product specific. The decision to consume a creative product not only presumes intertemportal persistence, the transfer between past en presence is modelled through the composing features of the product. This vision offers much potential. It perceives a product as a bundle of characteristics on which a person acts and it recognizes that the cognitive process of decision making relates to those features. Segmentation and dynamics can be analysed based on feature similarity rather than on the item itself. Since each creative product is distinct, persistence of consumption defined in terms

of the good itself is rather empty. It opens up when speaking of intertemporality in terms of future feature similarity. However, is it necessary to restrain the features to merely technical ones? The behavioural impulses put forward by economic theory might be equally influential than object characteristics. When accepting the notion of experience good, external information, provided by experts, awards or ratings, are important explanations of why a film will be opted for by a person. So while offering a more comprehensive model to deal with creative products, recommender theories ignore another dimension that might be of importance.

When it comes to discovering different motivational dimensions, tag based approaches hold a lot of potential. Tags are a direct expression of the keywords that an individual connects to an item. They can reflect object specific features such as actor or director, semi objective characteristics such as genre, as well as information variables such as expert advise or awards. Moreover, they might reveal decision making factors that were otherwise hidden. As explained, being a free expression, linguistically covering a broad spectrum, tags are more difficult to handle. Textmining techniques however offer ways to reduce their dimensionality. Given that convergence has been established in the tagging behaviour of individuals, it seems justified to work with features of the highest relevance. Tag relevance, from its side, is a concept with many angles, as will be further argued in the next chapter.

A number of insights from computer science will be transposed to build an economic choice model for movies. For one, the recommender approach appears to offer an answer, at least partially, to the insufficient conceptualization of creative commodities, thinking about them in terms of the key descriptions provided by individuals. Here, the tag based approach is opted for, despite its challenges in terms of linguistic variety and subjectivity. The choice function is meant to cover a broad range of elements that shape the decision making for a movie, be it technical features, external information or others. In conjecture, data mining techniques will be adopted, since clustering and latent class methods allow to segment the multitude of movies based on their main tag characteristics. Following the ideas of Hofmann & Puzicha

(1999), individuals are represented by their typical preference pattern, or in case of latent class methods, by a distribution over segments. These methods agree with the notion of contiguity - when symbolic events are paired up, their concurrence gives significance to agents who evaluate new items based on their similarity. The aim of recommender systems is to predict. Therefore, their results are almost exclusively expressed in terms of prediction statistics. To date, research on the subject shows little interest in how this prediction came about, what types of segments or classes are formed and how they evolve. In contrast, this work will investigate the specific nature of the segments or typical preferences and what specific features they are build up from. Another aspect of recommender literature is that little distinction is made between choice and rating. The emphasis is on predicting future rating behaviour. Persons are grouped when their rating behaviour is similar not when their choice of movies is. However, as was explained, decision making regarding a product and the evaluation of the experience afterwards is a two tier process. Rating is a post consumption act that induces new meanings on the preferences. Moreover, as stated by Jin et al. (2004), similarity in taste for items does not imply that users hold equal rating patterns. This has been partly addressed by normalization methods. A limited number of authors, Gantner et al. (2010), Gunawardana & Meek (2009), work with the probability that a user will act on an item and this is probably what recommender systems ought to envisage. Whether an individual will opt for a product or not might be the first concern for most stakeholders in the movie industry.

The analysis of this thesis is aimed at getting insight into the decision making factors behind movie choice behaviour. Given the multitude of person and item information, techniques for segmentation are needed to reduce dimensionality. Wishing to combine tag based methods with a latent class approach, hybrid systems appear suitable to build an inclusive model. The starting point will be the Latent Class Regression model by Kagie et al. (2009), integrating feature elements elegantly in the PLC-CF model of Hofmann & Puzicha (1999). As was argued in the second chapter, logistic regression is an expression of random utility theory and fits with a discrete choice approach. From that perspective, latent class logistic models bridge economic

theory and computer science, touching upon theories of Tversky and Kahneman and offering a tool to segment both items and consumers based on co-occurrence of their characteristics. The method offers other advantages, such as being more resistant to overfitting and being able to deal, in a model conform way, with missing observations, Porteous et al. (2010). Moreover, the Bayesian approach makes it possible to bring in domain knowledge by defining priors on them. What is most interesting about Bayesian Latent Class models compared to other cluster methods is that people are not deterministically appointed to one cluster but can have varying interest over segments and more importantly different characteristics can be attached in probabilistic terms to several latent classes. Therefore also, this statistical method corresponds with the framework presented in the introductory chapters, where users are sketched to be uncertain over their own preferences due to unknown ex ante quality, but can learn after ex post observation. Their uncertainty translates into probabilities over alternative latent classes. Using a dynamic variant, it opens the potential to look at the probability of an individual belonging to a segments at different points in time.

The proceedings of the thesis will consist of presenting a Bayesian Latent Class Logistic Regression Model (LCLR) where the explanatory variables are not mechanical features but tags. Unlike the model by Kagie et al. (2009), the dependent variable isn't rating but the act of having rated a movie or not. This implies a binary model replacing the Gaussian dependent of the reference paper. Tag diversity is high, therefore a preceding chapter will be devoted to tag relevance, where relevance is derived from a Latent Dirichlet Allocation (LDA) model. The reason why LDA is chosen for tag selection is that, if one uses numeric indices such as number of tags connected to items, one ignores keywords that are used by sub-segments of the population. At the same time, a selection imposes itself, since too many variables would cause issues of identification. Then, starting from a subset of tags, a LCLR model is tested to learn how the features connect into latent segments in order to gain insight in the determining factors of the established classes. The tag based results will be compared to genre results as this variable is at the core of many studies performed, both in economics and recommender theory for movies. In a third chapter, dynamics

will be introduced into the model. A Markov transition model will be applied to the data to verify how loyal individuals are to the distinguished classes. As a last exercise, the transition probability is made dependent on the rating. Doing so, the idea of experience goods will be fully exploited, the multi-characteristic items establish the segments, while the intertemporal preference switches become dependent on past evaluation.

The entire analysis will be performed on MovieLens data, which were assessed throughout this overview as a valuable micro-level open access dataset. The motivation for the choice of MovieLens will be discussed in the next chapter. Also the data manipulations are explained, clarifying the constraints that are imposed to achieve the samples used in the thesis. Tags are selected as independent variables. They are seen as representatives of the multitude of features that steer consumer decisions. Being a text variable, their handling asks for a text mining approach to deal with synonyms, errors or double entries. The GroupLens research group performed intensive research to improve on tag quality and their findings will also be credited for in next data chapter.

Chapter 4

The Movie Data Sets and the Data Management

4.1 The MovieLens data set

A number of movie data sets are prominently used in empirical scientific literature, be they the study of recommendation systems or to investigate social networks, namely Rotten Tomatoes, IMDB, Netflix, Flixster and MovieLens (p. 226). The last three base their recommendations on ratings entered by users, the first two systems start from attribute specification and similarities in search behaviour. A topic search in Web of Science provides 451 hits for MovieLens/GroupLens over the past 10 years, compared to 374 for Netflix, 177 for IMDB, 17 for Flixster and 8 hits for Rotten Tomatoes. Acknowledging that databases can be the subject of investigation without being noted in an abstract, keyword or title, it is fair to say the Netflix and MovieLens data are dominating the empirical recommendation literature. Moreover, IMDB, as was argued in previous chapter, is often used in conjunction with MovieLens or Netflix data to augment them with movie features. Netflix started initially as a US DVD rental company which evolved into a global video on demand supply system. In 2006, they launched a competition for the collaborative filtering based recommendation system that proved most accurate in predicting users' rating. It was based on a training data set of 100.480.507 ratings

that 480.189 users gave to 17.770 movies. Data were made available in a quartet <user, movie, rating, date of rate>. Judged on prediction precision, measured by the Root Mean Squared Error (RMSE), the performance of the competing algorithm is evaluated on a qualifying set of 2.817.131 ratings. Results were expressed in terms of gain over Netflix's own system Cinematch. The grand prize was awarded to a team succeeding in improving the RMSE by at least 10 percent. It was granted in 2009 to "BellKor's Pragmatic Chaos", a collaboration of researchers from Commendo Research & Consulting, AT&T Labs, Robert Bell, Yahoo! and the Pragmatic Theory contender group.

While the Netflix prize gained the involvement of academics to improve their algorithms, MovieLens was conceived as a scientific project from the start. The research lab behind the recommender system, namely GroupLens, is organized around topics of recommendations, online communities and digital libraries. They are part of the Department of Computer Science and Engineering at the University of Minnesota, comprising around 30 researchers and around 80 alumni. MovieLens is only one of their projects. It consisted of building a web interface which offers users a service to help them find movies they like to watch, improving on rating and tagging interfaces. Their algorithms result from fundamental research into automated collaborative filtering.

Harper & Konstan (2015) provide a recent comprehensive overview of the history, features and context of the MovieLens data. They are generated as a by-product of the online recommendation system, that was released in 1998 as a successor of EachMovie, replacing a propriety collaboration algorithm for a user-user CF technique. Since its release, the recommendation system was characterized by a steady growth of 20 to 30 new users a day, achieving a total of up to 250.000 users in 2015. Over time, the web interface was enriched. In 2000, a new interface was launched, including reviews, groups and importing external metadata such as box office data and releases. Five years later, in 2005, discussion fora were included as well as tagging facilities. Later, the tagging data were released in combination with the user rating data. After the closure of the Netflix competition in 2009, some of their features were incorporated into MovieLens, such as poster art and plot synopsis.

The main interface of Movielens is shown in figure 4.1. It consists of a number of movie lists, including top picks, recent releases, favourites from last year and new additions. The user can rate a selection of self chosen movies on a 1 to 5 scale as displayed in 4.2, using star values which have since 2003 incorporated half points. Conditional on a minimal number of ratings added to the system, the user receives a set of recommended movies. When a user initiates the rating activity, it induces a feedback loop as other movies are subsequently shown that might be of interest. Underlying those suggestions is a collaborative based predictive result. In a later version, this information was blended with overall movie popularity figures. Also the CF algorithms evolved over time from item-item to user-user. The cold start issue was dealt with in different ways, corresponding to different versions of the MovieLens interface, requiring the participants to rate initially 5 randomly chosen, later 15 movies selected according to popularity. In addition, the search facilities changed over the years and filter opportunities were offered on genre, director and other attributes. Also titles can be suggested by the registered members, that way completing the movie database. The design of the interface, the shape of the process and the choice of the CF mechanism all influence the nature of the rating behaviour. Hence, the adaptation of several features has affected the rating nature over the years.

In 2005, a tagging interface was added to the MovieLens database. They are displayed next to the movies and by clicking them, a list of movies is displayed. The sorting is based on a likelihood metric. Later, in 2007, a tag rating system was introduced, followed two years later by tag expressions. Quite some research went into the amelioration of the quality of the entered tags. Different options were compared by asking user opinions through surveys. The research performed by the GroupLens team on tag quality and tag rating is discussed in more detail in the next section.

In general, little is known of the identity of the users of MovieLens. Demographic information is only available for the early years. Since 2007, personal and group profile pages were introduced, added to discussion fora, but they were not used with high intensity. It is clear from analyzing the MovieLens data that there is a big difference in the intensity of use, with some persons posting a high number of ratings

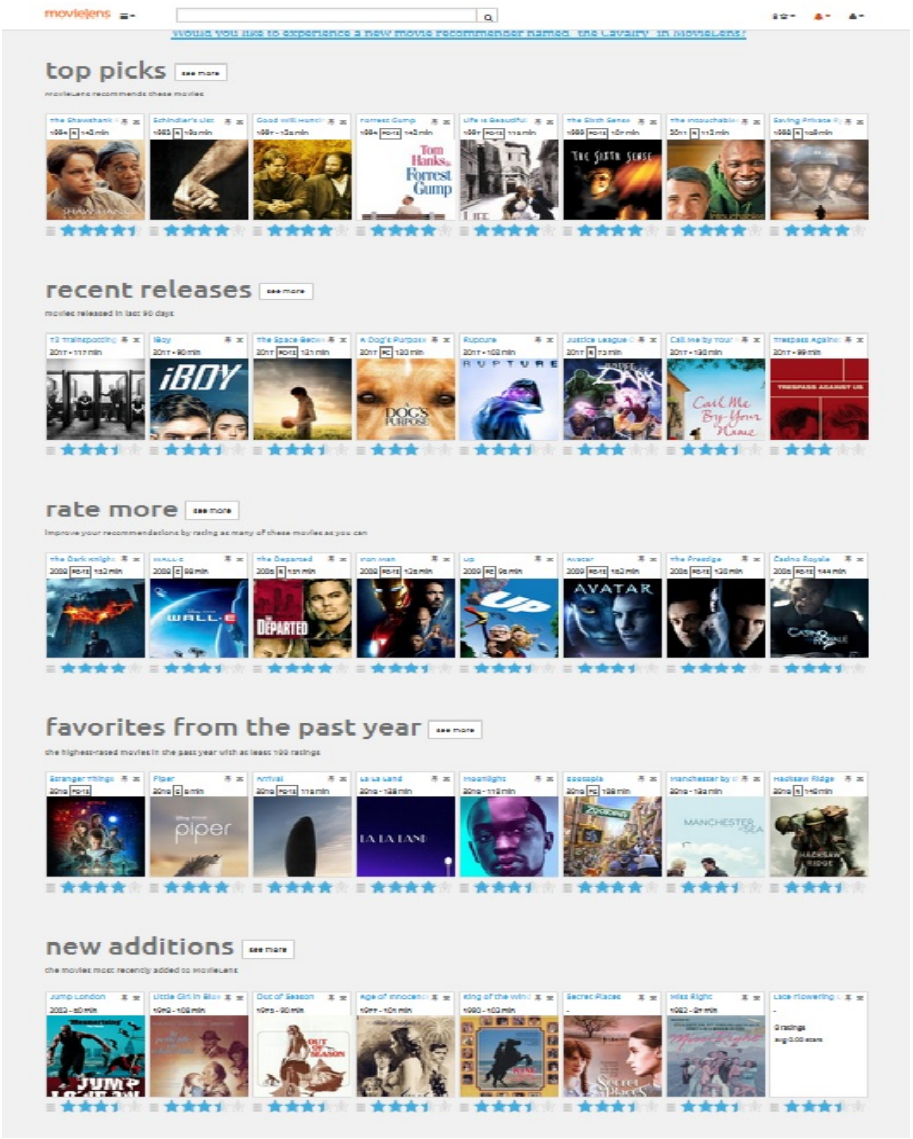

Figure 4.1: Screenshot of MovieLens Interface

while a majority adds only a few. The interface works with registered membership, but registration is easy. The data entry is connected to timestamps, which are as such not necessarily connected to the date of consumption. At a certain moment, a large number of items is rated, including movies that were watched a long time ago.

The MovieLens data were chosen for several reasons. First, they are open data made available for research purposes. The recommendation system is built by and for research. Both the user interface and the collaborative filtering algorithms are the result of scientific study. Secondly, the online gathering of movie ratings pro-

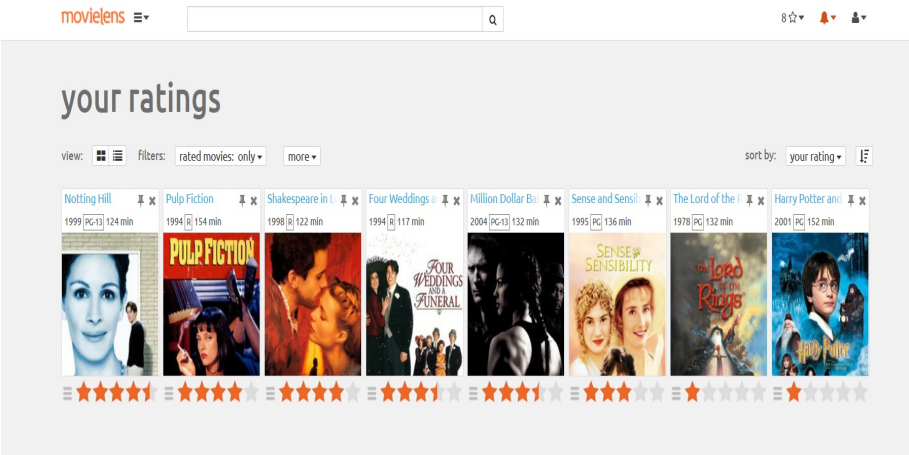

Figure 4.2: MovieLens Rated Movies Overview 5 scale with half values

vides the opportunity to organize a large online field experiment, making it easy to assemble information on user judgement at a large scale. This is strengthened by the observation that the user community showed commitment over the years, engaging in research surveys that were aimed at improving services. Finally, the MovieLens data were released as open access data for research purposes, packaged in three distinct sets, labelled as the 1m, 10m and 20m, referring to the included number of ratings. The fact that the data have been made public encouraged their use. By 2014, the dataset was downloaded 140.000 times. Searching for the term MovieLens in google books provides (accessed June 2016), 5.660 results and about 9.540 hits in google scholar. Moreover, the dataset was used intensively for research on recommendation systems and visualization of big data and for that reason form a valuable benchmark from which to compare results, Pentreath (2015).

4.2 MovieLens and the quest for tag quality

A number of authors connected to the GroupLens research team, like Shilad Sen and John Rieldl, performed intensive research on tag quality and how a tag interface is optimally constructed to generate the most relevant tags. In their article, Sen et al. (2007), "the quest for a qualitative tag", an experimental design is introduced embedded in the MovieLens interface. It consist of a thumbs up and thumbs down rating widget, not for the item but for the tag, appearing alongside the search pages.

In addition, they performed an online survey addressed to 2.531 active MovieLens users, asking them for specific feedback in relation to tag quality. Starting from personal opinions, they compare their revealed evaluations to a number of indicators that proxy tag value. Based on aggregate user behaviour, the authors test the indicative value of the number of times a tag was applied or searched. The established relationship between the induced five star rating scale, coming from the survey, and the number of searches or applications is mildly linear. Tags that are not often searched/applied were rated low. However, tags applied/searched very often obtain a somewhat lower rating than those in the middle range. This shape is attributed to the fact that some of the most appearing tags are related to personal classifications and are therefore considered less relevant to the overall user. The authors also compare their thumbs up/down system to the survey results, which shows that thumbs down clearly links with a rating value of 1 or 2, while thumbs up has an almost equal likelihood over the ratings 3-4-5, also slightly declining towards 5. When focusing on personal and aggregate behaviour, individuals show high persistence in their evaluation of tag quality, independent of what item the keyword is connected to. Taken in aggregate, users seem to agree well on a subgroup of tags, but show more consistency on bad tags than on good tags, indicated by lower average variance levels for low rated tags. At the same time, some tags are divisive or controversial, as measured by their entropy value. Moreover, the downside of applying total counts is that some singular power users may have a big influence, specifically on the negative evaluations.

In selecting valuable tags, there is an apparent trade-off between between coverage, circumstances in which the method can generate a prediction, and the precision of the top-n ranked. Systems of non-involvement, such as search values are in the advantage, as they do not perform bad on precision while being characterized by good coverage. Hybrid systems, like is the case for recommender systems, generally perform better. The same conclusion is made in a later study, "Learning to recognize valuable tags" by Sen et al. (2009*a*), where based on the same survey results, a larger number of implicit (without additional user effort or interface modification) and explicit systems were tested. The performance of all implicit-explicit combinations

outperforming singular measures in terms of precision.

In other writings by the same authors, Sen et al. (2009*b*), Vig et al. (2011), the issue of tag quality was formalized in terms of tag-preference, tag-relevance and the tag-genome. Tag preference refers to a person's sentiments towards a tag. It can be asked directly and expressed in 'like' or 'dislike' variables, or it can be inferred. Inference is established by the authors given different algorithms to calculate a user's preference from the interaction user-movie such as movie-clicks, log-odds clicks, movie ratings or a Bayesian movie rating generative model. Tag relevance signals how tight a tag applies to an item. This can be measured in terms of tag popularity or correlation between users' preferences for a tag and their preference for the movie. The so called "tag genome" for movies, refers to a vector of tag relevance values among all tags. The tag genome is further explored in a test setting where persons could ask for more or less of a certain tag in their selection for a movie: "I want a movie like pulp fiction but with less violence". This revealed a stronger preference for descriptive tags compared to discriminating tags, the latter being measured by the degree of entropy.

The analysis of the GroupLens team on tags is aimed at improving tag recommender algorithms, meaning they want to influence the tagging behaviour of the users of the system. At the same time, tags are used to improve the recommendation of movies. Indeed, Sen et al. (2009*b*) use the term tagommender, referring to tag based recommendation systems. Their approach results in them inducing a normalized inferred tag preference into a set of equations, composed of a cosine, linear and regression tag predictor. The first predicts the rating of a person for an item as a weighted average of the user's preference for the movie's tags, the linear tag as an average of the values of the prediction estimates of user's inferred tag preferences and their rating, while the regression predictor applies a linear equation for each movie where the input variables are all users' inferred tag preference. When compared to traditional CF algorithms, the tag based systems perform better in recommending a specific product, but not so much in terms of Mean Absolute Error.

4.3 The data structure

Handling the MovieLens dataset is a form of big data processing using server based techniques. The data can be presented in a relational database structure as follows:

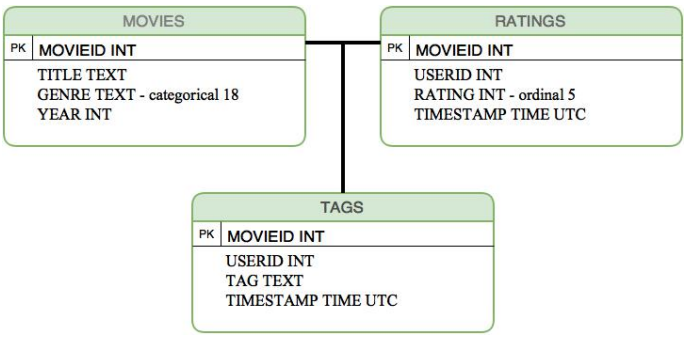

Figure 4.3: Database Structure MovieLens

The movie titles are consistent with those of the IMDB database. The data structure used for this study includes some minor modification in relation to the original. The movie year, which is part of the title e.g. "Casino Royale (2006)" in the original set, has been converted into a separate variable. The genre variable consists of 18 categories: Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. The ratings dataset provides for each user what rating is attached to the movies of choice. There is a timestamp, which reflects the moment of input of the rating, not the moment of watching the movie. The UTC time stamp was converted in a day, month and year variable. Users are anonymized by means of their userID, which also appears in the tag-set.

This study works with subsets of the MovieLens dataset, reducing it to a size manageable by PC software. Data are manipulated using SQL queries. A number of filters was imposed on the data. First, the set of movies was limited to films released after 1998. This was done, partly with the aim of data reduction, but also because remakes of older movies cause similarity of titles and generate doubles in case of merging feature sets. An additional requirement imposes that only user/movie combinations are selected where the rating year is within two years after the release of the movie. While also meant as a data reduction process, this is done mainly to

single out ratings entered in a close time span after the release of the movie. This potentially increases the likelihood that the movie was experienced in a movie theater and guarantees that the movie is well remembered at the moment the rating process took place.

Dataset A
MovieID
Movieyear>1998
Userid
Rating
Ratingyear <= Movieyear + 2

Table 4.1: Core Data Set

4.4 Data processing

The Latent Dirichlet Allocation analysis that will be performed in the next chapter is an exercise in text mining. Text mining, intelligent text analytics or natural language processing, aims at discovering patterns or processes in initially unstructured information provided in natural language. As words are indexed and placed in word document matrices, they can be analysed in a quantitative way. Text mining can be considered as the quantification of text. This can be done by calculating the degree of occurrence, represented in word frequencies or through visualization in word clouds. More advanced techniques come from data mining such as automated classification or clustering. They allow the categorization of a large corpus of texts as well and the detection of topics deducted from the co-occurrence of words.

Text mining techniques have been applied in many disciplines such as bibliometrics, to delineate disciplines, find emerging topics or, like is the case for Pubgene, to mine and connect various sources of medical information. It is also used in the automated processing of online messages, to improve search engines or to generate user profiles for various consumer targeting purposes. In the field of movie studies, text mining

has been of importance. One can refer to the influential work on opinion mining and sentiment analysis. Based on texts extracted from social media, user's views are identified and classified in degrees of evaluation. The aim is to identify feature words that are grammatically connected to sentiments, Pang & Lee (2008), Sharma et al. (2014). The aim of sentiment analysis is to extract adjectives, set of words or phrases that express positive or negative opinions. Technically, opinion mining starts from plain text and therefore demands more operations in terms of data pre-processing. Tags are a limited character expression of relational meaning. This doesn't imply that no further cleaning is required before a tag set can be used for analysis.

The user/movie variables of the filtered MovieLens set were merged with the tag dataset, meaning that tags are attached to the items where available. This resulted in a table of 10.893 lines of 1.299 movies tagged. A total of 4.060 singular tags were found and presented for cleaning. Since it is mainly words or group of words, rather than a full text, some standard text mining techniques such as stopword removal are less relevant. However, tags are very diverse and text analytics is more meaningful if the tags are cleaned to a certain degree. In this study, this was done manually following certain steps:

- Stemming: this means that a word is reduced to its stem. It is a classical procedure in text mining. A word like alcoholism is replaced by the word alcohol or Hitchockian by Hitchcock. Sometimes a tag was also converted into its noun, because converting it into the stem appeared to be meaningless. An example is activist that was replaced by activism and not by active

- Synonyms: words are replaced by one tag of equal meaning: tags such as 007, James Bond are mapped into Bond

- Typographical errors: errors appeared mainly in name spelling of actors or directors; e.g. Tobey Maguire was also found as Macguire and Maquire, Judi Dench vs Judy Dench

- Double entries: sometimes multiple tags are placed into one field entry such

as "sequel; history; treasure; president". They were separated as distinct tags

- Type classification: proper names are replaced by their type category; e.g. "MovieLens top pick" or "'Yahoo top pick"' were replaced by "list" to indicate that a list source was the inspiration to the choice

- Indirect meaning: "not true to the book" or "book adoption" is converted to "based on book" whereas "based on the life of" is replaced by biography

This table with cleaned words was then merged back on the original user-movie-tag set based on equality of the original tag in order to obtain a final set with cleaned and unified tags. Dataset B is represented in a database structure, meaning that one line identifies a triple <user, movie, cleaned tag>. To perform the LDA analysis of next chapter, this was further linearised to one line for each user with the names of all the movies he/she watched and the tags attached to them all lined up.

Dataset B
MovieID
Movieyear>1998
Userid
Rating
Ratingyear <= Movieyear + 2
Tag

Table 4.2: Core Data Set Tag Augmented

For the Latent Class Logistic Regression of Chapter 6 and the Latent Transition analysis of chapter 7, some further constraints were added. The rating years were limited to a period between 2001 and 2006. Moreover, the set of users was restricted in two ways: First, only those users were retained that had ratings over more than 5 years. This was done in the light of the Latent Class Markov Model (LCMM) where dynamic analysis is more meaningful when dealing with users that are active over several years. Moreover, users with less than 150 and more than 1.500 ratings

were omitted from the empirical exercise. This was done because some "power users" might appear too influential on the system. The resulting set consists of 470 users evaluating a set of 2.214 movies. Further downsizing of Dataset A was necessary, as for the latent regressions, the consumption of the 2.214 movies was considered for all the users, substantially amplifying the number of lines.

Dataset C
MovieID
Movieyear>1998
UserID having rated at least 5 years
having rated between 150 and 1500 movies
Rating
Ratingyear <= Movieyear + 2
Ratingyear between 2001 and 2006

Table 4.3: Data set for Latent Class Regression and Transition analysis

The resulting Dataset C is the core set for the analysis of the thesis. To be used for the LCLR and LCMM, the data were converted in terms of a choice decision exercise by 470 individuals over 2.214 movies. For each movie, one first determines if a person with a certain UserID has rated it. This is represented by a binary 1/0 variable. That operation generated 1.040.580 lines. Then, a second matrix was created. Starting from a selection of tags, a search was done if that tag was connected by at least one user to a particular movie. That way, a matrix was created with movieID on one axis, the selection of tags on the horizontal and a yes/no if the tag belonged to that movie or not. That way, one turns the tags into a kind of objective characteristic of that movie. Finally, the two matrices were merged. The movieID was replace by its tags, creating a model where the decision to rate a movie was related only to the presence of an array of tags.

Chapter 5

Latent Dirichlet Allocation for Movie-Tag Segmentation

5.1 Introduction

This chapter studies profiles of movie choice inferred from tag labelling. The tags are added by users of recommender systems to annotate the object and can be considered as proxies for the prime features users attach to an object. Looking at large scale data, profiles can be disentangled, based on co-occurence of tags. It is not uncommon to classify objects or users based on similarity in their tag profiles. Gemmell et al. (2008) explore clustering of tags to ameliorate personal search and navigation profiles. They see it as a means to reduce noise and redundant ambiguous tag assignment. Other examples include Dattolo et al. (2011), Deutsch et al. (2011), Cui et al. (2011), the latter introducing TagClus, and more recently, Li et al. (2016), focusing on co-occurrence group similarity to measure the relevance of tags.

Cluster analysis is a form of exploratory data mining applied to group items characterized by a degree of association. This association is mainly founded on co-occurrence and measures are used to signal similarity, often distance measures such as cosine or Jaccard. By applying cluster analysis, a partition is installed of objects with similar features or individuals that are alike in their behaviour. Cluster models often attach one user or item to a particular profile. In reality however, users' motives are driven by different decisive dimensions. These dimensions are not unique, nor

have they stable contours. For example, it is most likely that not the genre dimension in isolation or the presence of an actor is decisive in the consumer's choice for movies, but rather a mixture of different motives or criteria. Some however will be more influential to a particular segment of consumers. The features that typify the segments are thus not uniquely attached to one partition. To accommodate for that, a Bayesian model is opted for. The Bayesian approach is a well-considered choice, first because a researcher can make explicit prior beliefs highlighting the uncertain feature appraisal. Secondly, heterogeneous and hidden preference structures can be inferred by calculating the posterior probabilities for each individual. That way, estimation results not only explicit how characteristics are probabilistically related to latent segments but also how consumers are not necessarily deterministically attached to one particular preference profile. The uncertain nature of experience goods and taste formation can then be fully tested. Moreover, Bayesian statistics offers a coherent framework to structure massive data. Also factor models could be addressed to reduce dimensionality. Especially SVD was mentioned in chapter 3 as a valid method underpinning model based recommenders. However, factor analysis is less suited in the current context of discovering preference profiles for consumers. Factor models are more prone to overfitting and the interpretation of the factors is less straightforward. Opposite to that, the likelihood statistics that come with Bayesian estimation form reference points to compare models with different classes, times points and variables.

The aspect model by Hofmann & Puzicha (1999) discussed in chapter 3 formulates a representation of the probabilistic dependencies of variables to the latent classes. It is applied to segment movies based on ratings by consumers. However, the idea found more resonance in the realm of natural language processing. They borrowed insights from Latent Semantic Indexing (LSI) by Deerwester et al. (1990), a technique for analysing the hidden themes in a corpus of texts, taken from a spacial representation of the words they contain, the so called semantic space. It bears on the assumption that words appearing together in a text share a common meaning. A corpus of texts is transformed in a matrix of documents and terms. Then, singular value decomposition is used to distract the main latent dimensions. Both are "bag

of word" approaches, meaning that the order of words is ignored. Both methods perform dimensionality reduction, relating each document to a place in a lower dimensional topic space, Blei et al. (2003). Probabilistic Latent Semantic Indexing (PLSI) by Hofmann (1999) and Cohn & Hofmann (2000) relaxes the assumption that a document is generated from only one topic, but rather is a mixture weight of topics. Moreover, each topic is a probability distribution over words. The parallel with the section of model based recommendations is obvious: the SVD techniques were used in the models by Koren & Bell (2011), the probabilistic variant by Hofmann & Puzicha (1999) and Kagie et al. (2009). Users replace documents, movies replace words. In this section, whilst working with movies, it is a linguistic topic analysis. A corpus of texts is replaced by a set of users, the words are replaced by a combined set of movie titles and the tags that individuals attached to them.

The topic model used is Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003), which is a three level hierarchical Bayesian model. It takes the generative process a step further than the Probabilistic Latent Semantic Indexing (PLSI) by Hofmann (1999). In PLSI, documents are not induced by a generative process. In LDA, each document is a mixture of topics where the mixture proportions are distributed as a latent Dirichlet random variable. Contrary to PLSI, the topic node is sampled repeatedly within the document, not once for the corpus. In the light of the analysis to be performed, this is an interesting addition. It implies that the method not only estimates the distribution of features related to topics but also the degree of assignment of an individual to the topics. Those are not trained but inferred from the sampled topic distributions. Given that the array of titles and tags attached to a person represents his/her choices and opinions, topic models are expected to shed a light on the latent decision layers and the degree to which users are connected to them.

Looking for patterns in the semantic space is more than merely a technical statistical exercise. Griffiths et al. (2007) elegantly motivate the concordance between the world of semantic representation and the probabilistic latent methods used in language processing. They argue that the difference between LSI and LDA masks the difference between distance measures and the contrast model, opposed by Tver-

sky (1977). Topic models, quoting the authors: "can be thought of as providing a feature-based representation of the meaning of words, with the topics under which a word has high probability being its features". Indeed, the interpretation of the weight values does agree with that of the contrast model. The association between two words increases by each topic that assigns a high probability to both and decreases if latent classes assign a high probability to one feature but not to another. This agrees with Tversky's argumentation that common and distinctive features ought to affect the notion of similarity. It is an important insight because it provides a justification to use topic models as a means to search for latent classes based on tag similarity, where similarity agrees with the concepts put forward in chapter 2 of this thesis. Tags, used as proxies for user's opinions on movies, are represented in the semantic space. Based on tag similarity, dimensionality is reduced and latent classes are labelled in accordance to the features with the highest probability. Latent classes can be thought of as consideration sets.

The next sections will apply a topic model on the set of movie titles and tags provides by MovieLens. The objectives are twofold. Firstly, discovering the main latent classes revealed from the user's attachment of tags to movies. Secondly, selecting the most relevant tags. A tag will be considered relevant when being given the highest probability to define a topic. The features heading the topics will be used further in the decision making models presented in the last two chapters. This chapter will start by formalizing the generative process and estimation techniques behind LDA. This is followed by a descriptive statistical analysis of the tags, before the actual LDA exercise is presented. The chapter closes with explaining the value added of one of the distinctive features of LDA, namely the fact that it generates a probability distribution revealing the likelihood of individuals being attached to a particular latent class.

5.2 Topic Models

5.2.1 The generative process

In Latent Dirichlet Allocation, data are seen as observations that arise from a generative random process with hidden variables. Users base their decisions on mixtures of topics which are probability distributions over tags. A generative model is based on the assumption that there are simple probabilistic sample rules which describe how tags might be generated on the basis of latent random variables, Griffiths et al. (2007). The goal is to fit the best set of latent classes optimally explaining the set of observed data. Here, the observed data consist of users expressing an array of tag/title combinations. LDA is strongly founded on the conjugacy between the Dirichlet and the Multinomial distribution. The Dirichlet is a distribution on the simplex belonging to the exponential family. The dimensionality of k, the number of topics or classes, is considered known and fixed. Given α, a K-dimensional random variable θ has the following probability density on the simplex, Blei et al. (2003):

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\theta_1^{\alpha_1-1}\cdots\theta_K^{\alpha_K-1} \tag{5.1}$$

The parameter α is a K-vector with components $\alpha_k > 0$, and $\Gamma(.)$ is the Gamma function.

Consider an individual specific set of tags $\omega_i = (w_1, w_2, ..., w_N)$ of length N and a set of individuals $\iota = \{i_1, ..., i_I\}$. Now define the following parameters:

δ is the parameter of the Dirichlet prior on the tag distribution
α is the parameter of the Dirichlet prior on the topic distribution
θ_i is the topic distribution for individual i specified as $\theta \sim Dirichlet(\alpha)$
β_k is the tag distribution determined for topic k specified as $\beta \sim Dirichlet(\delta)$

An array of N tags expressed by a user is considered to be generated by the following process: 1. First θ is sampled from a Dirichlet distribution. This implies that θ lies in the $K-1$ dimensional simplex. It can be thought of as the degree the topics are

attached to the individuals. 2. Then for each of the N tags: 2a. Choose a topic c_n from a $Mult(\theta)$ 2b. Choose a tag w_n from the multinomial conditional distribution $p(w_n|c_n, \beta)$.

Figure 5.1 is the graphical representation of LDA. The outer plate represents users while the inner plate shows the choice of topics and tags. The parameters α is a hyperparameter whereas β is sampled K times. The variable θ is sampled once per user and the variables c and w are word level and sampled for each tag by each user. It shows the conditional structure and presents it as a hierarchical model structure.

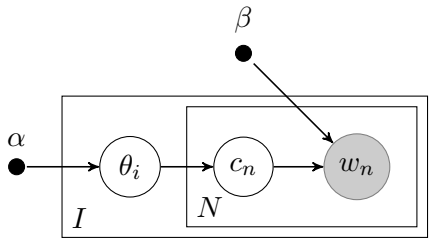

Figure 5.1: Plate Diagram of LDA.

5.2.2 The distribution

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of latent classes c and a set of tags w is specified by Blei et al. (2003) as:

$$p(\theta, c, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(c_n|\theta)p(w_n|c_n, \beta) \tag{5.2}$$

By integrating over θ and summing over c, the marginal distribution is obtained:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{c_n} p(c_n|\theta)p(w_n|c_n, \beta))d\theta \tag{5.3}$$

The LDA model, like PLSA, imposes the important assumption of exchangeability of words in topics, i.e. each token is independent from the previous one, meaning that the order can be ignored. It is semantically a strong assumption. The succession of several words can sometimes provide important information when classifying

texts, however in the context of this analysis, the order of tags has little meaning. Therefore, exchangeability seems a fair assumption to make. Mathematically, it has strong consequences. A set of random variables is exchangeable if the joint distribution is invariant to permutation. Applying De Finetti, any collection of exchangeable random variables has a representation as a mixture distribution, Aldous (1985). The probability of the sequence of words and topics then has the form:

$$p(w,c) = \int p(\theta)(\prod_{n=1}^{N} p(c_n|\theta)p(w_n|c_n))d\theta \tag{5.4}$$

5.2.3 Estimation

What needs to be estimated is the distribution of the hidden variables. There are different methods such as EM with variational inference, the EM with expected propagation and Gibbs sampling. For this exercise Gibbs sampling is employed. Gibbs sampling is a Markov Chain Monte Carlo method. It is specified for LDA by Griffiths & Steyvers (2004). Given the training data, a posterior is sampled, given this sample the model parameters are inferred. Draws from the posterior distribution $P(c|w)$ are obtained by sampling from:

$$\mathrm{P}(\mathrm{c}_n = K|w, c_{-n},) \propto \underbrace{\frac{q^{(j)}_{-n,K} + \delta}{q^{(\cdot)}_{-n,K} + V\delta}}_{topic-word} \underbrace{\frac{q^{(d_n)}_{-n,K} + \alpha}{q^{(d_n)}_{-n,\cdot} + k\alpha}}_{document-topic}$$

The factor $q^{j}_{-n,K}$ indicates how often the jth tag in the vocabulary is assigned to topic K without word n, Grun & Hornik (2011). The dot . refers to summation over the index where d_n indicates the user in the set to which tag w_n belongs. In the Bayesian models δ and α are parameters of the prior distributions over tag and topic. The predictive distributions of the parameters θ and β given w and c are defined by:

$$\widehat{\beta}_K^{(j)} = \frac{q_K^{(j)} + \delta}{q_K^{(\cdot)} + V\delta}$$

$$\widehat{\theta}_k^{(d)} = \frac{q_K^{(d)} + \alpha}{q^{(d)} + k\alpha}$$

5.3 Experimental setup

5.3.1 Data collection

GroupLens is a research lab of the department of Computer Science and Engineering at the University of Minnesota. It puts available for research three movie databases of different size, with data gathered through movielens.org, each containing a table of ratings and a movie genre table. The data set used for this study is the 10M data set, which does not contain personal features, but includes information on user tags. The data set contains 10.000.054 ratings and 95.580 tags applied to 10.681 movies by 71.567 users of the online movie recommender service.

A filter was imposed on the data set. The MovieLens data set table MOVIES was reduced to items of production year no earlier than 1998 and the <user, item, rating> triplets were filtered from the MovieLens table RATINGS where the rating date lays within two years after the production date. That set contains 10.987 user-movie combinations with tags. Herein, 1.299 movies are tagged with 4.060 distinct tags. There is an average of 8 tags per movie and 10 tags per user, however, the distribution is skewed, making that averages have to be read with caution. The maximum number of tags were attached to the movie "V for Vendetta", namely 124 and the most prolific user entered 1.260 tags.

5.3.2 Tags: first analysis

The text mining and cleaning activities related to the tag database where covered in the data management chapter. The cleaned tags were divided in two categories; those that can serve as movie characteristics and those that can be seen as mere

judgements. In some cases, it is clear one deals with a judgement such as "better than expected" or "best movie ever". However, some judgement adjectives can equally be interpreted as a potential feature; tags such as too long, kitsch, bittersweet or fanciful refer to evaluation as much as to more objective features of the movie. Some were therefore given a double classification. A part of the judgement tag descriptions reflect an entire opinion "entertaining for the wrong reasons" or "attempts comedy and drama and fails in both". For judgements, contrary to features, the sentences were left as entities. In some cases, features were extracted from it e.g. "I don't think arms dealers look anything like Nick Cage" the features arms dealer and Nick Cage were extracted and added as objective characteristics. Some tags will not be treated in the analysis. It involves tags like "seen 2005", "see also"; "what if"' or specific dates which most likely refer to the date the movie was watched. In total 4.060 expressions were cleaned. Table 5.1 shows a shot of the tag cleaning file. It consists of the original tag, which is first replaced by a cleaned concept, than by a single tag word.

Figure 5.2 shows that the occurrence of tags follows a long tail distribution with a few tags being omnipresent and a large number of tags appearing a few times, often only once.

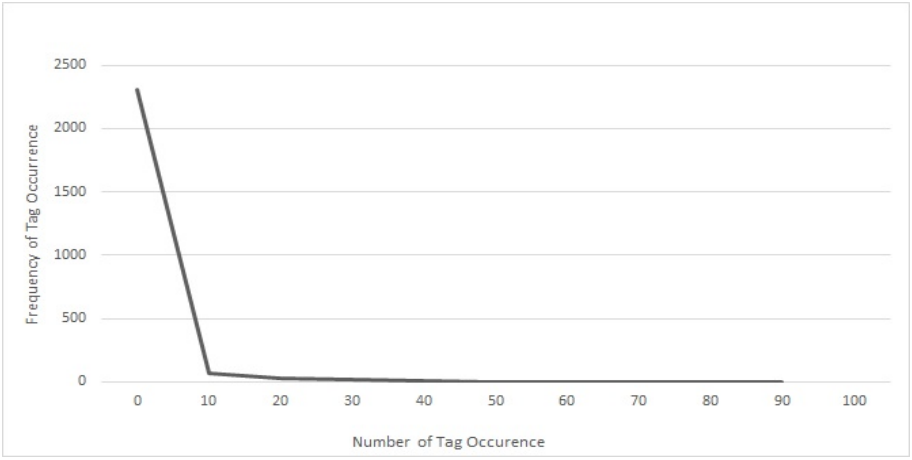

Figure 5.2: Distribution of tag occurrence

Table 5.2 and 5.3 line up the top 30 of most occurring tags under the type "characteristic". In a first count, the tag attached to a person is taken only once, even if the user employs the keyword on different movies. The results show that genre remains

Tag Description	Concept	Word	Type
Aardman	aardman	aardman	char
Aardman studios	aardman studios	aardman	char
abduction	kidnap	kidnap	char
accountants	accountants	accountant	char
Adaptation	adaptation	based on	char
Adapted from	adapted from	based on	char
Alcohol	alcohol	alcohol	char
Alcoholism	alcoholism	alcohol	char
Amazing	amazing	amazing	judge
amazing acting!	amazing acting	acting	char
amazing acting!	amazing acting	amazing	judge
amazing martial arts scenes	amazing martial arts	amazing	judge
amazing martial arts scenes	amazing martial arts	martial art	char
animated	animation	animation	char
animation remade as live action	animation remade as live action	animation	char
animation remade as live action	animation remade as live action	live action	char
based on a true story	based on true story	based on	char
based on a TV show	based on tv	based on	char
Based on a TV show?	based on tv	based on	char
based on a video game	based on a game	based on	char
based on book	based on a book	based on	char
Based on comic	based on a comic	based on	char
based on comic book	based on a comic	based on	char
based on novel	based on a book	based on	char

Table 5.1: Screen Shot of MovieLens Tag cleaning

an important qualifier, with terms such as commedy, action, drama, animation appearing in the top 15 of most frequent tags. However, the top delimiter seems to be the fact that the movie is based on a secondary source. The term "based on" is a container word aggregating other concepts. Over one third is based on a book or novel, others involve based on comic, play, true story, tv show or game. Also the fact the film is a sequel or a remake, which can be seen as based on a previous item, seems to matter. This confirms the hypothesis that often movies are chosen where known or familiar characteristics come into play.

Viewers appear to judge strongly on the movies' content or screenplay when tagging. The keyword "plot" is an important indicator and connected to that, how the story evolved or ended. Ending is the fourth most important word appearing. Apparently, this very much determines the overall feel connected to the movie. It is a mapping from great/bad ending, tearful ending, haunting ending, confusing ending. Also the term "twist", pointing to an evolution in the plot, is a distinguishing keyword to describe a movie. Looking at the type of movies, there seems to be a strong preference for comedy and funny pictures, followed by themes of war and politics, then fantasy, superhero and animation.

When allowing for double counting in the tags, where users attach the same tag to different movies, the conclusions are overall similar with some small differences. The based on factor remains important, however for some people, the fact that the movie received an Oscar or appears on an online suggestion list (which emphasizes the value of recommender systems) is added as a tag to a large segment of the movies they rated. These findings are in line with studies in movie economics relating box office revenues to awards. Singular actors or directors appear to have less influence, with the exception of Tarantino, no one appears in the list of most used tags. When looking at the full tags set, a lot of actors or directors are named more than once, so when reduced to a single variable "presence of top actor" the variable might turn out to be influential.

A last factor showing in both counts is the target audience. The word children is mostly used as in: for children/not suited for kids. Also terms as "family movie"

appeared in the top 25 whereas others such as "chick flick", are not represented in the top 25 but are listed in the top 50.

Tag	Count	Tag	Count
based on	97	magic	34
comedy	76	sciencefiction	33
funny	67	bond	33
ending	52	children	32
war	45	sex	31
politic	41	history	31
fantasy	40	twist	29
action	40	remake	29
drama	39	plot	28
superhero	39	family	27
animation	37	Pixar	27
love	36	zombie	26
violence	36	acting	25
murder	35	Tarantino	25
romance	35	sequel	25

Table 5.2: Top 30 of most occurring tags of type CHAR/ no double counting tag-person allowed

Tag	Count	Tag	Count
based on	341	div	59
murder	147	history	56
dvd	128	sequel	56
comedy	122	fantasy	55
list	117	magic	53
Oscar	98	pg	52
funny	92	drama	51
superhero	82	romance	50
war	78	violence	50
ending	75	newyork	50
politic	73	love	48
animation	70	bond	48
nudity	65	family	47
action	62	remake	47
sex	61	children	46

Table 5.3: Top 30 of most occurring tags of type CHAR/ double counting tag-person allowed

Tag	Count	Tag	Count
great	41	mustsee	14
good	34	horrible	13
overrated	34	disturbing	13
funny	30	dark	13
boring	27	excellent	13
bad	20	predictable	12
disappointing	17	complex	12
best	15	quirky	12
better than expected	14	crap	12

Table 5.4: Top 15 of most appearing tags of type JUDGE

Table 5.4 shows the top 15 most common words under the type judgement. They include straightforward evaluation adjectives such as great, good, bad. Part of them also relate to the storyline, such as dark and complex. Noteworthy is the reference of various judgement tag descriptions to prior expectations. The term "better than expected" appears in the top 10, but also terms like overrated, disappointing or predictable reflect that prior expectations were made before watching the movie.

Sen et al. (2006) compare, in one of their first studies on MovieLens tagging behaviour, about 3.000 tags entered in the 2006. The authors divide into factual tags, identifying facts about movies, subjective tags expressing user opinions and personal tags meant for self use such as library. The later includes things like bibliothek, DVD. The subjective tags are not attached in the same way as was done in current study. Rather, they are based on word-type-noun versus adjective. The study classifies genre descriptions as the dominant appearing factuals, with the first five being action, drama, Disney, comedy and teens. The top listed subjective keywords are classic, chick flick, funny, overrated and girl movie.

Finally, an overall count of all MovieLens tags (restricted to movies past 1998), excluding massive taggers and without pre-processing (table 5.5) confirms the results of the sampled analysis. The tag based on a book is listed at the top, only surpassed by classic. The genres comedy, action, drama appear to be influential. This is also the case for external opinions like Oscar winning or being listed IMDB top 250.

Tag	Count	Tag	Count
classic	681	owned	406
Tumey's DVDs	641	erlend's DVDs	405
less than 300 ratings	504	sci-fi	391
based on a book	502	Oscar (Best Picture)	382
comedy	483	seen more than once	377
R	478	drama	348
action	475	movie to see	345
70mm	460	fantasy	309
Nudity (Topless)	442	Disney	307
dvd	433	imdb top 250	305

Table 5.5: Overall tag count of MovieLens data

5.3.3 Preprocessing and best model selection

The language used to estimate the LDA model is R, with the R-code inspired by Grun & Hornik (2011). The R package "topic models" is founded on the textmining package "tm" by Feinerer et al. (2008), which offers the code elements to construct a corpus and transform it into a document term matrix. Those packages need predecessors like Snowball, Hornik (2007), that, along with the textmining R software, provides tools to process terms with natural language processing operations like stopword, number, punctuation removal, tokenizing and stemming. For this exercise, most of those operations were performed manually. For the calculation of the posterior distributions, the R software code movMF is used, Hornik & Grun (2014). The basis for a training model of LDA is a segmented vector W. It is a vector of users and their tags augmented by the movie titels to provide extra information for topic labelling.

For each user, one starts with an array of movie titles and the tags connected to them. The LDA analysis is thus based on the hybrid mapping of two information sets, the tags on the one hand, the movies on the other. Movie titles were restricted to those having tags attached to them. Then, a corpus is constructed with a "bag of words" per individual. It is entered in R as a dataframesource.

For $\iota = \{i_1, ..., i_I\}$, define the corpus:

$$W = \begin{bmatrix} \{tagsbyuser1, ..., movietitlesbyuser1\} \\ \{tagsbyuser2, ..., movietitlesbyuser2\} \\ ::: \\ \{tagsbyuserI, ..., movietitlesbyuserI\} \end{bmatrix}$$

The corpus is converted into a document term matrix which is the input for the estimation of LDA topic models. In a document-term matrix, the rows correspond to the users and the columns to tags/movietitels. The entry of $DTM_{j,i}$ indicates how often the j^{th} term occurs with the i^{th} user.

$$DTM = \begin{bmatrix} n^o_{1,1} & n^o_{2,1} & ... & n^o_{W,1} \\ n^o_{1,2} & n^o_{2,2} & ... & n^o_{W,2} \\ : & : & : & : \\ : & : & : & : \\ n^o_{1,I} & n^o_{2,I}, & ... & n^o_{W,I} \end{bmatrix}$$

The number of rows equals the number of users, the number of columns reflects the size of the total word corpus. This tag corpus was not selected a priori, but was determined out of the vector W, which involved that the set of vectors had to be tokenized. No tags were removed, not even the terms with low frequency. The matrix used for this exercise consists of 971 users and a total word corpus of 3.612 words. The DTM matrix is the input for the LDA analysis which consists of estimating the probability that users are lead by one of the K pre-specified latent dimensions and the probabilities that one of the 3.612 words are attached to a specified class. The LDA exercise is based on the hybrid mapping of two information sets, the tags on the one hand and the movies on the other. It is a tag extended topic model.

The number of classes k needs to be fixed. The selection of k is a somewhat ad hoc exercise of trial and error. However, there are some methods available to select the best model. One method, suggested in the paper of Griffiths & Steyvers (2004), is

SEED	NUMBER OF CLASSES	LOGLIK
100	15	-96550
500	14	-96491
1000	14	-96398
1500	17	-96518
2000	14	-96464

Table 5.6: Best Model Selection LDA

Number of Classes: The value that maximizes the log likelihood for different posterior draws of k

to evaluate the log likelihood for different posterior draws of the parameter. This was done for a sequence of 1 to 50 classes. To initiate the Gibbs sampling based estimation, iteration values, burnin values and an initial seed have to be set. The value of the initial seed generates a particular sequence of pseudo-random numbers and therefore alternative seeds can induce different simulation paths. The log likelihood maximization exercise to find the optimal number of classes was performed for different initial seed values. Table 5.6 shows a combination of initial seed with the optimal number of classes. Generally, the optimal number of classes equals 14, once 15 and once 17. The number of classes that is retained is 14 because the log likelihood is larger for all $k = 14$ compared to $k = 15$ or $k = 17$. Moreover, the distinct classes can be labelled in a meaningful way.

5.4 LDA analysis

5.4.1 Estimation results

Tags are diverse and as discussed above, their distribution is long tailed. A lot of tags are unique. Contrary to some text mining clustering exercises, the decision was taken to keep those variables in the analysis, if only for the reason that working with a threshold frequency would results in a very small set of tags remaining. However, the high variety of terms in combinations with their low occurrence are ingredients

for unstable segments and it was expected that, in the best case, the tag analysis would be decisive on only a limited number of topics. The correlation between words is not always very strong and therefore no specific meaning can be deducted from it. However, in a trial of both $k = 14$, $k = 15$ and $k = 17$, some classes did fluctuate, but others remained stable and a meaningful interpretation could be attached to them.

Table 5.7 shows the results of the Latent Dirichlet Allocation based topic model with 14 specified classes. The first class is headed by the word *based on*. It is clear, also from the tag frequency analysis, that whether or not a movie is inspired by a third source, majorly but not exclusively a book or novel, is a determining factor for movie choice. It stands with the tag "murder", which is not surprising since a lot of murder plots are based on detective novels. The interpretation of this class relates with that of topic 11, which is also a stable class, with words like ending, twist and plot. Both refer to the importance users seem to attach to the storyline of the movies. Together, both latent classes represent a clear content dimensions, pointing to the importance of the quality of the scripts, often denied in empirical analysis because it is difficult to measure their presence and impact. However, looking at tagging behaviour teaches us that consumers indicate often and clearly that the story matters. Where the script factor has been largely ignored when explaining box office revenues, recent studies brought some interesting insights into its role and importance in the context of movie sales. A study published by Goetzmann et al. (2013) examines soft information and hard information and their relation to screenplay prices and box office revenues. Soft information is proxied by screenplay complexity, by using measures of the number of words in the logline, the number of other movies mentioned in the logline and the number of genres assigned to the screenplay. Hard information is measured by the screenwriter's experience and past success. The hard information is shown to stand in a positive relation with revenues and screenplay prices, the later being higher with writer's experience but lower with fuzziness of the script. Also Eliashberg et al. (2014) used natural language processing, consisting of word frequency counts and genre-content analysis to investigate the effect of script quality characteristics on a movie's return on investment.

Topic1	**Topic2**	**Topic3**
basedon	superhero	oscar
murder	animation	pg
setting	anime	newyork
bechdeltest	pixar	rrated
ghost	incrediblesthe	family
comedy	starwarsepisodeiii-revengeofthesith	britain
dystopia	batmanbegins	london
animal	eternalsunshineofthespotlessmind	suicide
sport	japan	losangeles
timetravel	martialart	england
Topic4	**Topic5**	**Topic6**
littlemisssunshine	nudity	sequel
johnnydepp	sincity	remake
musical	topless	teen
piratesofthecaribbeandeadmans	acting	zombie
tumeys	nocountryforoldmen	bond
harrypotterandthegobletoffire	charliekaufman	chickflick
kingkong	munich	funny
prestigethe	satire	pirate
casinoroyale	easternpromises	vampire
judelaw	stevecarell	maffia
Topic7	**Topic8**	**Topic9**
vforvendetta	war	list
brokebackmountain	politic	redbox
dani2006	history	bibliothek
capote	gay	ratatouille
syriana	death	jacksonville
thankyouforsmoking	revenge	ironman
dani2007	drug	classic
apocalypse	religion	scorsese
childrenofmen	surreal	simpsonsmoviethe
davincicodethe	violence	walktheline
Topic10	**Topic11**	**Topic12**
funny	ending	owned
tarantino	fantasy	pan'slabyrinth
killbillvol.2	children	casinoroyale
brucewillis	twist	brokebackmountain
tomhanks	love	fairytale
nicolascage	christian	prideprejudice
adamsandler	plot	woodyallen
bournesupremacythe	magic	goodnightandgoodluck
jimcarrey	chroniclesofnarniatwatw	historyofviolencea
kingkong	violence	luckynumberslevin
Topic13	**Topic14**	
comedy	sex	
action	quirky	
drama	atmospheric	
romance	departedthe	
sciencefiction	humor	
magic	style	
yearoldvirginthe	adventure	
harrypotterandtheprisonerofazkaban	crime	
documentary	future	
serenity	juno	

Table 5.7: Latent Dirichlet Allocation with 14 topics based on tags and movie titles

The most important tags belonging to topic 6 are *sequel* and *remake*. It potentially refers to a characteristic of creative good consumption which is inherently dynamic in the sense that like or dislike of a previous good might lead to the consumption of more of the same in the future. Consumers not only want to reduce consumption risk that results from the experience good but they also seek for familiar characteristics, Bohnenkamp et al. (2015). Ravid (1999) showed how sequels perform above average, though often not better than the original, and how it represents a high return low risk investment for the film industry. Interestingly, they are reviewed less and worse. Hennig-Thurau et al. (2009) define sequels as a form of "brand extension", an idea put forward earlier by Luehrman & Teichner (1992) in their suggestion to use option theory to measure brand extension values for motion picture sequels. Also the importance of remakes was recently analysed and discussed by Bohnenkamp et al. (2015). They state that "remakes are a special type of brand extension in that they retell an existing narrative in the same modality in which it has been told before. Like other brand extensions, remakes are of limited risk for consumers and offer audiences familiar branded attractions as a result of their established characters and story". The LDA results of this analysis bring both terms together into one class. Consumers seem to be joined under "*brand familiarity*".

Topic 2 is a clear expression of a specific subgenre, namely animation in combination with movies such as Starwars and Batman. This segment points to a particular niche of the consumer market. Rather than labelling it as a subgenre, one could speak of a style or even a way of life to some. Therefore, topic 2 is classified as "*Geek*" or "*Nerd*" movies. Topic 12 could, in the same line of interpretation be seen as assembling those liking "dreamy movies", with words like "fairy tail" and a title like Pride and Prejudice, however, this class appears altogether difficult to pin down. The additional presence of movies like Casino Royal and Lucky Number Slevin makes labelling more diffuse.

Segments 4 and 7 gather box office successes with the emphasis on adventure, the first group includes *family type movies* while the latter signals more the +18 category with overall a somewhat darker line of drama. Class 7 includes hits *surpassing mere entertainment*, where theme, storyline and staging matters. This is also the case

for class 8, which contains tags like history, war, religion, and can be labelled as *"engaged"* or *"value added seeker"*. Class 5 refers to adult movies connected with tags like nudity and topless, a choice dimension rarely explicitly considered in analysis. The tag sex also heads layer 14, but the mixture of words in this topic points to an amalgam of terms, such as quirky, atmospheric and humour. In contrast, class 10 seems well outlined as a group, but presented itself as rather unstable under different try runs of the analysis. It can be read as "*actors/directors matter*", not entirely distinctive from class 4. Overall, the presence of actors or directors does not seem to dominate the tag analysis, which suggests it is less of a decision factor than commonly assumed. Also in the movie economics literature, the significance of actors was ambiguous. Occasionally positive effects are established for directors. This is also the case for class 10 containing the combination Tarantino/KillBill. The dominance of director over actor was also noted in some of the recommendation studies.

Classes 3 and 9 cover a number of tags that are omnipresent in the movie economics literature on box office prediction, namely Oscars and ratings for topic 3 and recommendation in listings for topic 9. Movies, being experience goods are seen to benefit from prior information provided by external sources. However, research shows that the effect of information sources is often not conclusive. The two classes are distinct, blogs and recommendation lists on the one hand, which are of growing importance since the early years 2000, and awards on the other. The first attaching more importance to the opinion of the crowd, while the second group values certification by experts. The search for information seems broader in segment 3, also including Motion Picture Association of America film ratings and provenance or scene location plays a role.

There is a specific layer, namely topic 13, that assembles mainly genre tags. Indeed, some individuals seem to label movies systematically by the genres attached to them by the producers or movie websites. In chapter 3, it was argued that genre is a way of recognizing resemblance, a factor of common knowledge through which meaning is passed. The omnipresence of genre tags proves that genre is an important benchmark to label movies. This study shows however that in general, the motivational

classes steering the consumers' decision are more extended. While some topics directly point to elements intensively researched in the movie economics literature, such as the impact of Oscars or ratings, others highlight the importance of dimensions often ignored or only recently brought to attention. One such dimension is contents or storyline. The fact that a movie has a *based on* factor, which can be based on a novel, life story, a comic or even a game seems to matter. Given that tags are free word associations by users which are very divers and low in occurrence, tag based probabilistic latent class analysis is not expected to bring strong contoured results. Employed as indirect signals or proxies of user's underlying motivations, it does unveil the importance of the familiarity factor and supports theories of brand extensions of movies. Not only does the consumer seek familiarity in making connections with a book, they also choose what they know by opting for sequels or remakes. Those particular features of a movie seem to surpass others such as genre or even the presence of known actors or directors. The latter seem to strengthen certain layers rather than shape them. Genre is a manifest label for some, but the movie-tag bundle points to the fact that genre is an always changing concept, better to be extended to broader definitions such as style, class, segment or consideration set.

5.4.2 The value added of LDA

As explained in the introductory sessions of this chapter, the LDA model is an expression of the uncertainty of users over feature sets in combination with uncertainty of tags over classes. This contrasts with cluster methods where one tag is allocated to one topic or one latent dimension to one user. The tag probabilities were used to order the tags in the latent classes and label the topics. The probability values over the entire word set for topics 1 and 2 are displayed by figure 5.3 and 5.4. The profile of class 1 has two peeks at the words *based on* and *murder*. The other terms in the top 10 have rather equal and lower values. This is different from topic 2, the tags standing out are less manifest, but at the same time more numerous. That makes the labelling more solid.

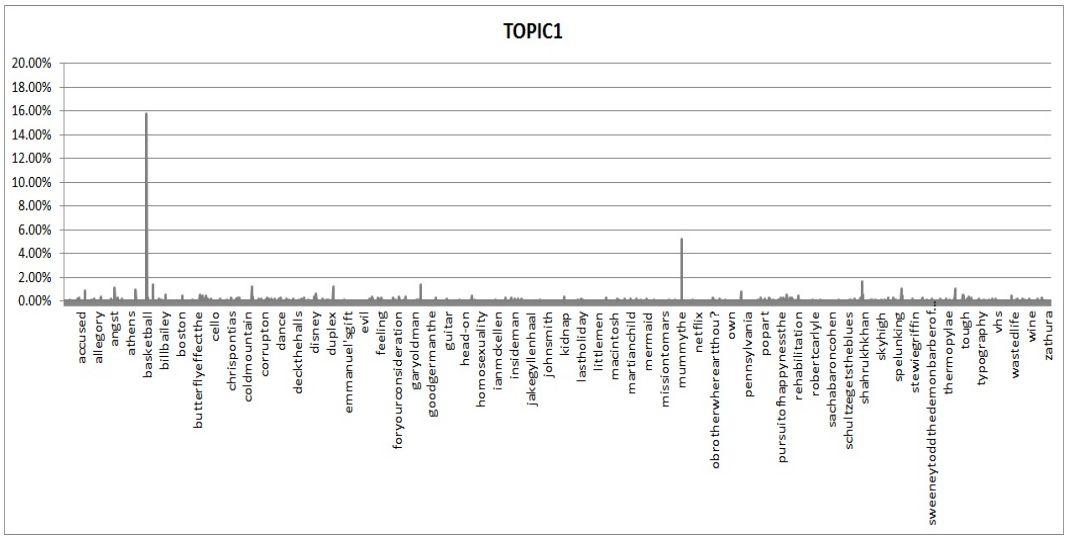

Figure 5.3: Probability distribution over tags for topic 1

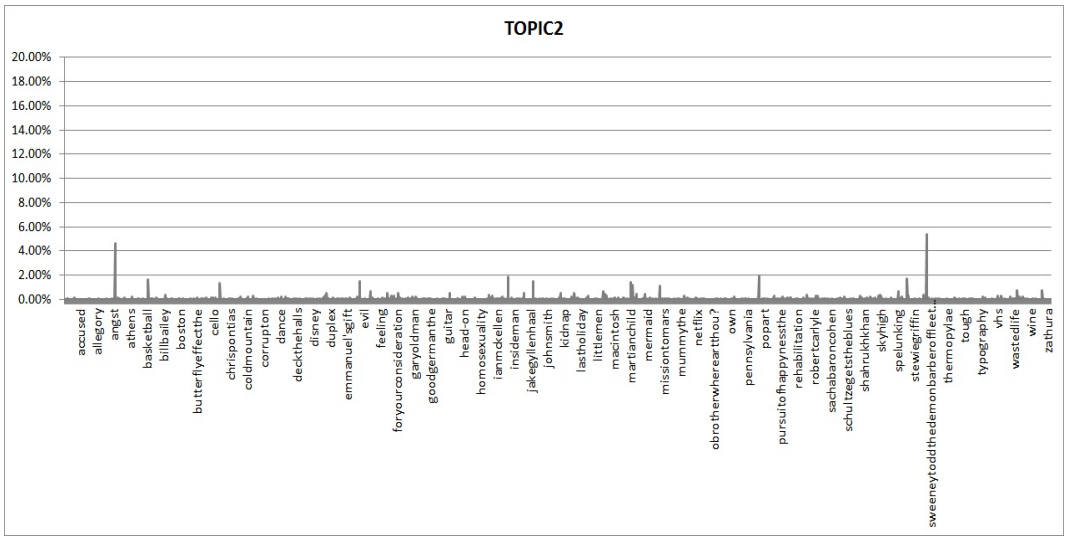

Figure 5.4: Probability distribution over tags for topic 2

U U	P1 P8	P2 P9	P3 P10	P4 P11	P5 P12	P6 P13	P7 P14
U1	0,07	0,07	0,07	0,09	0,07	0,07	0,07
U1	0,07	0,07	0,07	0,11	0,07	0,07	0,07
U2	0,07	0,17	0,06	0,06	0,06	0,06	0,07
U2	0,06	0,06	0,06	0,11	0,06	0,06	0,06

Table 5.8: Probability Distribution for user 1 and user 2 over different topics

Interesting information can also be distracted from table 5.8. It displays the probabilities that a user is attached to a specific class. The distributions are very uniform. This is partly the result of the Gibbs sampling method starting with a uniform prior and the final estimation being attracted to that. Therefore, it is not possible to say that most people are equally inspired by all dimensions in their decision for movie attendance. Results do lead to the observation that there are noticeable differences between users that are practically egalitarian and a small group who are up to 20 percent concentrated in one layer. Table 5.8 displays the distributions of the first two users of our sample. The first has an almost equal probability among classes, the second clearly has a more pronounced link to dimension 2, which is the so-called "Geek layer". That person has the following array of tags and movies *"'anime anime anime fantasy fantasy fantasy SpiritedAway Howl'sMovingCastle Steamboy LordoftheRingsTheFellowshipoftheRingThe LordoftheRingsTheTwoTowersThe LordoftheRingsTheReturnoftheKingThe*". The first user can be typified as more egalitarian, which is largely due to the lack of information that characterizes him/her. The user only added a few tags. This implies that for prediction, one acts as if that person had an equal interest over different dimensions. When there is sufficient information, it potentially signals that users are difficult to pin down to one class and are thus omnivorous. It is a factor worth knowing in case a company wants to target consumer segments. Moreover, the estimated value α of the Dirichlet distribution equals 3,5, which is high. It is more common in GIBBS sampling that the α value is lower than the initiation value. The higher this value, the lower the percentage of users that are assigned to one specific topic. Therefore, this exercise provides some evidence that decision making for movies, for most individuals, involves different consideration sets. It also shows that tags can provide insights in the consumer's motives despite their diversity. Results certainly have to be interpreted with caution, the probabilities attached to single tags are small. However, tag-extended methods allowed to perform a meaningful latent class segmentation.

5.5 Summary

This chapter presented a first exploration in the potential of movie tagging with the aim of discovering the incentives of movie consumers. Tags are linguistic expressions that users associate with movies, and are used as proxies for the features underlying choice behaviour. Discovering meaningful latent classes based on the entire tag set strengthens the belief that the motives of consumers can be pinned down to a number of segments or considerations sets. Moreover, the probabilistic expressions reveal that features are not uniquely attached to one dimension, but can be part of multiple classes.

Latent Dirichlet Allocation is a three level hierarchical Bayesian method which is popular for textmining purposes. It adds to probabilistic semantic indexing that users are randomly attached to topics. The method is statistically well suited to identify patterns in large sets of words. Equally important however, topic models agree with the concept of similarity put forward by Tversky (1977). That way the statistical methods of big data are connected to the conceptual views described in chapter 2.

Looking at some descriptive statistics on the tags expressed by the sampled users, indicates that the linguistic terms express a combination of the variables put forward in movie economics and the technical characteristics integrated in content based recommender systems. They include elements that are important to consumers searching for prior certification, such as ratings and Oscars as well as objective characteristics such as actor/director. Listing tag usage reveals that some important aspects are overlooked by both disciplines. First, consumers express a great interest in content, finding it important that the movie is a derivative of a book, a comic or a game. Also tags like plot and ending appear in the top of user statistics. Secondly, some categories are ignored in standard genre classifications and therefore also in research analysis. It involves "adult movies", where keywords like nudity are manifest and engaged movies, including terms as politics, history, religion. Looking at the subjective tags, expressing judgements, supports the vision that movies are

experience goods endowed with a prior belief dimension. It is reflected in terms like "better than expected" or "overrated".

An additional goal of this exercise is to select the most relevant tags which can be used for further analysis. Listing up the tags provides first insights. However, methods like LDA join tags based on contiguity and also place terms on top of classes that do not stand out in count tables. It is therefore important to look at tag relevance in terms of their probabilistic co-occurrence rather than on their occurrence. Else, one might miss out on factors that are important for particular subgroups. The results also make clear that it is meaningful to segment users based on latent classes. Some dimensions are clearly steered by the more traditional genre based elements, others point to the importance of content or information. Some tags manifestly shape the classes, but all are attached to the segments with some probability. Moreover, individuals are clearly not deterministically related to one partition, as is often forced upon in cluster based methods, but show a diversified interest over different consideration sets. The LDA analysis is however explorative. Give that many tags occur with low probability, results have to be looked at with caution. Moreover, topic models are not based on a choice model. The next chapter will keep the latent class approach, but the set-up will be converted in a decision model over the selected movies. Only the most relevant tags will serve as explanatory variables.

Chapter 6

Bayesian Latent Class Consumer Model for Movie Choice

6.1 Introduction

In their article "a rational analysis of categorization" Anderson & Matessa (1990) question the validity of a disjointed classification of featured artefacts. While some objects can be assigned to distinct segments, often features seem to be cognitively connected with multiple segments. It makes that a probabilistic theory of the feature-object relationship describes the decision making process more adequately. The LDA analysis presented in previous chapter suggests that contiguity of tags is useful to shape distinct segments. Results were based on the full tag set and a large number of keywords were attached to classes with modest probability. The LDA results in previous analysis served as an exploratory data analysis to discover patterns in the multitude of tags. It is not the estimation of a choice model for movies, but the applied technique is a form of Latent Class Analysis (LCA), a group of methods designed to empirically discover sets of latent types steering the observed indicators, McCutcheon (1987). In this chapter, another variant, namely a latent class logistic regression model will be presented, where the dependent variable is the probability a consumer picks a movie. A selection of tags from previous chapter will serve as explanatory variables.

Latent class analysis refers to a statistical method to infer latent structures from observed variables. The term latent means that the variable is not directly observable nor directly measurable. The premise of latent class analysis is that the co-variation observed among the manifest or observed variables is linked to latent variables steering, influencing or explaining the relationship between the observed. Clearly, the causal flow goes from latent to observed. The latent variable is therefore considered the true or driving source or cause. Collins & Lanza (2010) define the purpose of latent class analysis as "*to arrive at an array of latent classes that represents the response patterns in the data, and to provide a sense of the prevalence of each latent class and the amount of error associated with each variable in measuring these latent classes*".

The earliest core citation in the theory of LCA goes back to Lazarsfeld & Henry (1968). They describe the technique as the use of mathematical models for characterizing latent variables in the analysis of attitudinal measures for survey research. Their work on Latent Structure Analysis provided one of the first coherent and comprehensive theoretical treatments of the topic, Uebersax (2010). Goodman (1974a), (1974b) improved the approach by developing an implementable method for the maximum likelihood estimations of the LCA algorithm, a method closely related to the EM algorithms later developed by Dempster et al. (1977), Collins & Lanza (2010). Haberman (1974) added methods on maximum likelihood algorithms, this time in the context of a log-linear model. Up to date, both the EM methodology and the log-linear framework form the heart of the methods currently used.

Latent Class Analysis bears resemblance with factor analysis, however, the latent classes reflect more qualitative differences between categories of individuals or objects, Ruscio & Ruscio (2008). They are therefore represented by a categorical variable. LCA progressed over the years to deal with both nominal, ordinal and count variables as well as with the handling of sparse data. It can be considered a qualitative data analog, be it that it allows the multidimensional typological classifications from a set of observed discrete measures, McCutcheon (1987). In contrast, factor analysis is a method to detect more quantitative differences that can be represented on a continuous scale. In relation to that, factor analysis can be seen

as a more variable orientated approach, establishing patterns among variables that apply to all individuals while the focus of LCA is often more person-orientated, aimed at detecting patterns of similarity in individual's characteristics, Collins & Lanza (2010), Bergman & Magnusson (1997). LCA comes in different shapes such as Latent Class Clustering or Latent Class Regression. The former is related to the LDA method used in previous chapter, seeking for latent patterns in an array of variables. The latter additionally contains a single dependent variable, which can however be observed more than once. Current analysis works with a Latent Class Logistic Regression model (LCLR), a general linear model variant where each class can be explained by the difference in importance of the explanatory parameters.

A major distinctive feature of the presented model, when compared to recommendation models in existing literature, is that the key variable of consideration is the choice made by individuals. Current model starts from the question of whether an individual did or did not decide to watch the movie. In contrast, the main strand of recommendation literature aims at predicting rating levels, which in this work will be introduced in the dynamic model covered in the last chapter. For the empirical exercise undertaken in this chapter, the choice for a movie is approximated by the item being evaluated by the individual who takes part in the MovieLens rating system. It is a proxy and by definition imperfect, as it is possible that movies were watched that are not entered in the system, however, this is also the case when using ratings as a dependent. Four arguments support the approach. First, it is in accordance with the economic theoretical models presented in the introductory chapters where ex ante choice and ex post evaluation are separated. Second, the data show that agents do enter movies without an evaluation score and therefore, the variable can be considered as properly reflecting the choice set rather than the ratings set. Thirdly, it is a big data analysis. This implies that the scale of the analysis is of that nature that existing trends between variables are manifested where the amplitude of the data partially counts away the errors in the variables. Finally, from an economics point of view, choice is the variable that is of interest to the movie business. With regards to the supply side of the industry, the question whether an individual or group watch a movie is at least as important as the ratings that are

granted afterwards.

Latent Class Logistic Regression bridges economics with recommender literature. It was argued in the second chapter that, given a discrete choice setting and under the assumption of IIA, the random utility model translates into a logit model. This framework was singled out as equipped to deal with the multi-featured nature of creative goods, where each choice relates to a new product. After studying the various alternatives in recommender theory, also the hybrid probabilistic model based approach was put forward as dealing in the most comprehensive way with the multitude of object features. Users can be segmented, while features are attached to items in a probabilistic way. The model of Kagie et al. (2009) is inspirational, however the authors work with ratings as a dependent, which they translate into a Gaussian distribution. The explanatory variables include IMDB genre and keywords. Here the dependent will be binary choice, rating is moved to an intertemporal setting, and the explanatory variables are tags. The results of the tag-based LCLR will be compared to using merely the genre characteristic, as a much used variable in both the economics and recommender literature.

6.2 The consumer decision model as a logistic regression

6.2.1 The formal model

Expressed formally, the latent class model is represented by a response variable or indicator, one for each case i, denoted by y_{im}. There are M response variables $1 \leq m \leq M$, while the latent classes are represented by latent nominal variables c with K categories, $1 \leq k \leq K$, called classes, Vermunt & Magidson (2005). The exogenous or prediction value connected to a case is denote by z_i. The relation between response, latent class and explanatory variable can be represented by using the following probability structure:

$$f(y_i|z_i) = \sum_{c=1}^{K} P(c|z_i) f(y_i|c, z_i) = \sum_{c=1}^{K} P(c|z_i) \prod_{m=1}^{M} f(y_{im}|c, z_{im}) \tag{6.1}$$

The model is distinct in the fact that the probability density of a particular set of indicator values, given a constellation of explanatory z_i values, is connected through the unobserved latent classes. The connector $P(c|z_i)$ reflects the probability that an individual belongs to a certain class, given that individual's realized covariate values of z (the mixing weights). The mixture values $f(y_i|c, z_i)$ reflect the density of y_i conditional on c and z_i. Distinctive for the Latent Class Regression Model, compared with cluster or factor models, is that the single dependent variable may occur more than once per case. Assuming local independence:

$$f(y_i|c, z_i) = \prod_{m=1}^{M} f(y_{im}|c, z_{im}) \tag{6.2}$$

6.2.2 Logistic regression

In the Latent Class Logistic Regression model, the dependent variable y_{im} is conditioned on latent class and predictors. The probability of an individual being represented by a certain class potentially depends on covariates that are invariant across observations per case z_i^{cov}. Those covariates are not used in the analysis of this chapter, but will return in the dynamic analysis.

$$f(y_i|z_i) = \sum_{c=1}^{K} P(c|z_i^{cov}) f(y_i|c, z_i^{pred}) = \sum_{c=1}^{K} P(c|z_i^{cov)} \prod_{m=1}^{M} f(y_{im}|c, z_{im}^{pred}) \tag{6.3}$$

The response variable y_i consists of two possible disjoint outcomes 1 or 0, corresponding to the act of choosing or not choosing the good. The link function, linking the y to the independent variables is the log odds or logistic regression:

$$P(y_{im}|c, z_i) = \pi_{m,c,z_i} = \frac{exp(\eta_{c,z_{im}})}{1 + exp(\eta_{c,z_{im}})} \tag{6.4}$$

The linear predictor equation being:

$$\eta_{c,z_{im}} = \beta_{c0} + \sum_{q=1}^{Q} \beta_{cq}.z_{imq}^{pred} \tag{6.5}$$

Here β_{c0} is a class specific intercept and β_{cq} represents the class specific regression coefficients. This is a logistic regression model. The value of π is in the set (0,1); as the value of $\eta \to \infty$, the value $\pi \downarrow 0$ when $\beta < 0$ and $\pi \uparrow 0$ when $\beta > 0$. The link function is the log odds transformation. The odds, or the quotient that compares the probability that an event occurs to the probability of failure is expressed as:

$$\frac{\pi}{1-\pi} = \exp(\eta) \tag{6.6}$$

The log odds transformation converts the right hand side into a linear relation which eliminates the skewness inherent in estimates of the odds ratio. It ranges from $-\infty$ to $+\infty$, O'Connell (2006):

$$\log \frac{\pi}{1-\pi} = \beta_{x0} + \sum_{q=1}^{Q} \beta_{cq}.z_{imq}^{pred} \tag{6.7}$$

The latent classes are assumed to come from a multinominal distribution which is parameterized as:

$$P(c_i|z_i) = \pi_{c|z_i} = \frac{exp(\eta_{c|z_i})}{\sum_{c'=1}^{K} exp(\eta_{c'|z_i})} \tag{6.8}$$

6.2.3 Estimation

The model is estimated using the package Latent Gold by Vermunt & Magidson (2005). To estimate the latent regression model, a successive run is performed of the EM algorithm followed by the Newton-Raphson algorithm. Let ϑ be the vector of unknown model parameters which are estimated by finding the values that maximize the likelihood function:

$$log\mathcal{L} = \sum_{i=1}^{I} logf(y_i|x, z_i, \vartheta) \tag{6.9}$$

A prior $p(\vartheta)$ is used on ϑ to prevent boundary solutions. The implementation of a prior on ϑ results in maximizing the log posterior:

$$\begin{aligned} log\mathcal{L}_{post} &= log\mathcal{L} + p(\vartheta) = \sum_{i=1}^{I} logf(y_i|z_i, \vartheta) + p(\vartheta) \\ &= \sum_{i=1}^{I}\sum_{c=1}^{K} s_{ci}\, logP(c|z_i, \vartheta) f(y_i|c, z_i, \vartheta) + p(\vartheta) \end{aligned} \tag{6.10}$$

where $s_{ci} = P(c|z_i, y_i, \vartheta)$ and $\vartheta \sim Dirichlet(\alpha)$. Latent Gold sets the α values default to 1, which makes the influence of the prior small. Giving it a value of zero would result in a maximum likelihood estimation. The Expectation step of the EM algorithm consists of estimating $P(c|z_i, y_i, \vartheta)$, filling in $\hat{\vartheta}^{\nu-1}$, the E step. In the maximization step M, the complete data loglikelihood, that is the likelihood value when known to which latent class each case belongs, is maximized with respect to ϑ generating new values ϑ^{ν}. Latent Gold uses multiple sets of random starting values, that way avoiding local maxima, and a pre-specified number of iterations. Within the best 10 percent in terms of log-posterior, an extra 2 times iterations is performed till either a maximum number of iterations is reached or convergence falls within an EM tolerance limit. Then, the program switches to the Newton-Raphson algorithm (NR) until the maximum number of iteration or an overall tolerance level with respect to convergence is reached. The Newton-Raphson method updates parameters according to the following equation:

$$\hat{\vartheta}^{\nu} = \hat{\vartheta}^{\nu-1} - \epsilon H^{-1} g \tag{6.11}$$

The vector g is the gradient vector of first-order derivatives of the log-posterior evaluated at $\hat{\vartheta}^{\nu-1}$, H the Hessian matrix of second order derivatives, and ϵ is a scalar representing the step size, introduced to prevent decreases of the log-posterior

to occur. The derivatives are calculated analytically by Latent Gold. Iteration is stopped when the change in the log-posterior is smaller than 10^{12}, Vermunt & Magidson (2005).

Non-identification can occur when different parameter estimates yield the same log-posterior or log-likelihood value. Then, the observed information matrix, $-H$ does not reach full rank. Also weak identification may occur when the data are not informative enough to obtain stable parameter estimates. The choice of the priors influences the identification issue, as does the number of selected variables open for estimation. For the presented model, in some cases, the number of parameters or classes will need to be restricted in the light of establishing full identification of the model.

6.3 Assessing model fit

6.3.1 Likelihood statistics and information criteria

Recommender theory focusses mainly on predictive accuracy. Because of the complexity of the feature structures attached to movies and the diversity of users, the error values corresponding to most movie recommendation models are high. This work wants to find ways to deal with the heterogeneity of consumers and the multi-characteristic nature of the creative product. The aim is to explore if tags can serve as connectors forming meaningful segments. To achieve this, methods, criteria and/or measures will be needed to compare different models, taking account of the fact that data are sparse and that the large number of estimation variables challenges identification. In this study, model comparison relies on information criteria founded on log-likelihood statistics. They follow naturally from the Bayesian structure of the model and its estimation. Moreover, information criteria are widely used in the context of Bayesian model comparison as they offer a common scale discarding differences in parameterisation.

When different models have the same number of parameters estimated in the same

way, one might simply compare their best-fit log predictive densities directly:

$$log\mathcal{L} = \sum_{i=1}^{I} logf(y_i|x, z_i, \hat{\vartheta})$$

However, when comparing models of different size, and working with in-sample data, it is recommendable to adjust the deviance for fitted parameters. Information criteria typically work with the log predictive density of the data given the posterior distribution or a point estimate of the fitted model adjusted for the number of parameters, Gelman et al. (2014). Here, a log predictive density of the provided data given the maximum likelihood point estimate is used. The lower the estimated log likelihood, the worse is the fit of the model. Information criteria additionally penalize complexity of the model.

A first measure is the Akaike Information Criterion. This metric, introduced by Akaike (1973), (1974), is grounded into information theory. Asymptotically it estimates the difference in information loss of two candidates. It agrees with the relative Kullback-Leibler (KL) distance of the likelihood function specified by the fitted candidate from the unknown true likelihood function. The -2 is chosen for so called "historical reasons", as -2 times the logarithm of the ratio of two maximized likelihood values is asymptotically chi-squared under certain assumptions, Burnham & Anderson (2002). The joined components provide the following expression:

$$AIC_{\log \mathcal{L}} = -2log\mathcal{L} + 2npar$$

The first term indicates the reward for the goodness of fit of the model where the second term adds a penalty for increasing the number of estimation parameters, in this way correcting for overfitting.

Bozdogan (1987) overviews a number of information criteria in terms of what he calls "dimension consistency". Provided that a so called true model exists and is among the set of candidate models, then such criteria imply the selection of the true model with probability 1 as the sample size increases asymptotically. The above AIC

criterion does not obey this type of consistency requirement, but it can be adapted in a way that it does:

$$CAIC_{\log \mathcal{L}} = -2log\mathcal{L} + [(logN + 1]npar$$

A measure closely related is the Bayesian Information Criterion (BIC). It is initiated by Schwarz (1978) as a competitor to the Akaike (1973), (1974) information criterion. The measure is dimension consistent and serves as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Contrary to AIC, BIC does not provide an estimate of Kullback-Leibler information. The Bayesian Information Criterion is defined as:

$$BIC_{\log \mathcal{L}} = -2log\mathcal{L} + (logN)npar$$

If the generating model is of finite dimension and the true model is represented among the candidates, BIC will asymptotically select the candidate model having the correct structure with probability one. Moreover, Kass & Raftery (1995) demonstrated that the Schwartz criterion gives a rough approximation to -2 times the logarithm of the Bayes factor.

All presented criteria feature the same goodness of fit component, but differ in their penalty term. The latter value is bigger under BIC, therefore, one could state that the Bayesian Information Criterion favours smaller models. Cavanaugh (2012) states that BIC could be advocated when the primary goal of the modelling application is descriptive; to build a model that will feature the most meaningful factors based on an assessment of relative importance rather than mere prediction. The BIC criterion is popular in use, partly because of the mentioned consistency and partly because it performs better under larger datasets. Given the big size of the dataset used for this study, given also the interpretation of the BIC criterion in terms of posterior odds, it will be the preferential criterion to compare the different models suggested in this work. However, the other statistics will additionally be provided.

6.3.2 Classification statistics

The purpose of LCA is to discover classes or segments. Therefore, it is important to judge models also by their capacity to predict to which latent classes an individual case belongs, given its observed values of y and z, Vermunt & Magidson (2005). Classification of cases is based on the posterior class membership probabilities:

$$\hat{P}(c|z_i, y_i) = \frac{\hat{P}(c|z_i)\hat{f}(y_i|c, z_i)}{\hat{f}(y_i|c, z_i)} \tag{6.12}$$

This definition holds for the maximum likelihood estimates of the terms involved and for response patterns i. Response patterns are groupings of identical cases having the same covariate and predictor values and do generate the same response.

The classification of cases can be visualized by a classification table that cross-tabulates modal and probabilistic class assignments, Vermunt & Magidson (2005). Each entry (c, c') contains the sum of class c posterior membership probability for the cases allocated to modal class c'. The diagonals indicate correct classifications per latent class whereas the off-diagonal elements signal the wrong allocation of cases. A good model therefore has a high weight on the diagonal.

Apart from using a table, a metric similar to BIC can be defined, weighting not only the fit of the model, but also classification performance. The so-called Approximate Weight of Evidence or AWE is a measure due to Banfield & Raftery (1993), and is defined as

$$AWE = -2log\mathcal{L}^c + 2(\frac{3}{2}\log N)npar$$

Like with the BIC criterion, the lower the value the better. When classes are well separated, the value of AWE and BIC are expected to be close.

6.3.3 Prediction statistics

The central performance statistics in recommendation literature are the Mean Absolute Error (MAE) and the Mean Squared Error (MSE). They are based on the difference between the observed and predicted responses.

The predicted values are obtained as weighted averages of the class-specific estimates where the posterior membership probabilities $\hat{P}(c|z_i, y_i)$ serve as weights, Vermunt & Magidson (2005):

$$\hat{y}_{im} = \sum_{c=1}^{K} \hat{P}(c|z_i, y_i)\hat{E}(y_{im}|c, z_{im}^{pred}) \tag{6.13}$$

Then both measures are expressed as:

$$MSE = \frac{\sum\limits_i v_i[y_i - \hat{y}_i]^2}{\sum\limits_i v_i}$$

$$MAE = \frac{\sum\limits_i v_i|y_i - \hat{y}_i|}{\sum\limits_i v_i}$$

Here v_i denotes the replication weights connected to response pattern i. Both are common indicators for the forecast error or for the lack of fit of the estimated model, therefore one looks for a model with the smallest possible values.

When dealing with a regular regression model, generally the goodness of fit is indicated by using the R^2. However, in the context of a logistic regression, where the response variable is a binary 0/1, this statistic loses its meaning. A number of alternative or so-called "pseudo R^2" statistics have been defined to provide an indication of the goodness of fit for this type of models, McFadden (1973), Cox & Snell (1989), Menard (2000). The pseudo R^2 reflects the reduction in error of the basic model or null model compared to the suggested model. The null model can be seen as a prediction without information, based merely on constant term estimates.

$$R_y^2 = \frac{Error(nullmodel) - Error(fullmodel)}{Error(nullmodel)}$$

When the 0/1 choices of individuals are well predicted, the error of the estimation models goes to zero, making the pseudo R^2 larger in value. This value however rarely approximates 1, but rather varies in the range of [0,1 0,4] where the latter value indicates a "good fit". It is a rough index, important to interpret the results in context, but better accompanied by other measures.

6.4 LCLR Analysis

6.4.1 Data description

The LCLR analysis is performed on the 10M MovieLens dataset. A number of filters were placed on the data. Like for the LDA exercise, the set was restricted to user-movie combinations for which the user evaluated the movie within two years after the production date. This was done because the selection might include a bigger percentage of movies attended live at theatres. Moreover, the ratings attached to the films are potentially more up to date. In addition to the analysis in previous chapter, some extra filter mechanisms were imposed. One is that the rating years were limited to a time frame [2001-2006]. Additionally, in order to generate a meaningful user decision set, users that voted too less or too many movies were removed, to make sure that profiles are representative. To that aim, users with less than 150 and more than 1.500 ratings were deleted from the, already restricted data set. The application of several filters results in a set of 470 users evaluating a set of 2.214 movies.

Two types of models will be estimated, one based on genre characteristics and a tag driven choice model. An important task was to convert the data so that they were suited to be used to estimate a decision model. To achieve that, for the 470 distinct individuals, each of the 2.214 movies were converted in a binary decision variable, a 1 if the item was consumed, a 0 otherwise. This implied a total of 1.040.580

decision lines. Then, each movie was replaced by its characteristics. For the set of of tags, this implied a role back operation to the non-cleaned version of the word. The procedure went as follows:

- A tag is taken and is rolled back to its original equivalents. For example, the term "based on" represents a variety of inputted tags such as based on book, movie, comic and alternatives such as "adapted from" or even full text evaluations such as "book was better". The set of tags originally connected to the selected word compose the rolled back set. - The set of 2.214 movies is then inspected for the presence of the expressions pinned down in the rolled back set. This again is translated into a binary set with one expressing the presence of the feature item and zero its absence. A value 1 is given when the tag is attached to a movie by any individual in the dataset. - Finally, the choice set is merged with the set of movie features. Only tags appearing over 40 times in different movies were selected, combined with those tags that were heading the LDA segments. This provided a series of 43 explanatory variables. The movie decision feature set thus consists of a matrix of 1.040.580 lines and 44 columns.

The aim is to estimate to what extent the decision to choose a movie is driven by the presence of a particular feature and if latent classes can be discovered that can be labelled by the importance of a group of variables. The results of the tag based decision model are compared to a model were the movies are only typified by the genre characteristics attached to them. The genre characteristic is also part of the MovieLens dataset and is provided for each movie. A total of 18 genres is identified and often multiple genre features are attached to one movie. Like for the tag based model, the variables were translated into a binary feature/decision set.

The genre item is occasionally expressed by individuals as a tag and therefore the two models are partly overlapping. To accommodate for that, the genre item and the tag-genre item sets were mixed. This means that if a particular genre tag appears, than the genre data are added. As an example, if "action" gets a value of 1, it is because it is mentioned by at least one individual or it is indicated by MovieLens as an action movie. Therefore the tag set used here can be considered a hybrid or a

genre-augmented tag set. This does not imply that all genre types are represented.

In summary, the LCLR model which is investigated relates a binary yes/no choice set for all 470 individuals on the features of 2.214 movies. The explanatory variables are a set of genre variables, also represented as one/zero dummies for the first exercise. The second set consists of a genre-augmented array of tags, also binary variables. The question to be answered is if latent classes can be discovered which improve a one class model, merely relating the explanatory variables directly to the dependent. A second issue under investigation is whether the genre augmented tag model outperforms a model based simply on genres, a set which is always easily accessible and therefore often used in prediction models.

6.4.2 The Latent Class Genre Based Decision Model

The results presented show the estimation of a log-linear model, relating movie decision as a dependent to genre features. Like for the LDA model in previous chapter, the first stage of the experiment consists of determining the optimal number of classes. This is done using the BIC criterion. The BIC values for a 1 to 10 class simulation are shown in table 6.1.

	LL	BIC(LL)	Npar	L	R-value
1-Class Regression	-293609,6298	587336,1615	19	581435,6908	0,0191
2-Class Regression	-285036,6613	570313,2792	39	564289,7539	0,0382
3-Class Regression	-282669,1999	565701,4111	59	559554,8312	0,0432
4-Class Regression	-280855,1144	562196,2947	79	555926,6601	0,0486
5-Class Regression	-279860,215	560329,5505	99	553936,8612	0,0511
6-Class Regression	-279127,1164	558986,4079	119	552470,664	0,0535
7-Class Regression	-278690,5069	558236,2436	139	551597,445	0,0551
8-Class Regression	-278283,3943	557545,0732	159	550783,2199	0,0559
9-Class Regression	-277955,01	557011,3591	179	550126,4512	0,0567
10-Class Regression	-277656,1578	556536,7095	199	549528,747	0,0576

Table 6.1: LCLR GENRE BASED class iteration

LL = Loglikelihood, BIC = Bayesian Information Criterion, Npar = Number of Parameters, L =Likelihood, R-value = Reduction of error

The models can be compared based on the likelihood values and the Bayesian Information Criterion, where the number of co-variates was kept stable. The BIC values

continue to improve as the number of classes increases. The percentage decrease of the BIC value stabilises to 0.1 around the 6th and 7th class regression. Also in terms of segments appearing, the interpretation of the latent classes is stable, however, the 7 class solution is maintained because an additional niche class is revealed. It is clear from the model selection statistics that segmenting agents into feature classes improves the model fit, meaning that the inclusion of latent classes positively affects likelihood compared to a model where future consumption is solely based on user and genre. The largest gains are realized when going from zero to two and three classes.

The overall estimation statistics are shown in the table 6.3. Provided that each case is considered over 2.214 movies results in 1.040.580 observation points. The p-value has to be interpreted with caution given the sparsity of the data. Also the size of the observation set has a positive impact on the significance levels. However, the low rejection rates do suggest that the null hypothesis can be rejected for the proposed array of parameters, or that the dependencies in the model are strong enough not to be generated by chance. Also the Wald (=) values, testing for equality of the parameters between classes are high. The overall log likelihood value is low, which was to be expected under the binary and sparse structure of the model. The classification table shows a major emphasis on the diagonal axes and the classification errors are low. The classification table presents the probability of belonging to class c, given that a person was assigned to class $\hat{c}$ under modal assignment. Modal means that cases are assigned to the class with the highest posterior probability. A heavy diagonal points to correct classification. It indicates that the presented genre based model does well in terms of predicting class assignments of cases and that no systematic misallocation seems to have taken place. The goodness of fit is also demonstrated by the low absolute error (0.1468) and squared error (0.073).

The estimation results are presented in table 6.4. The exponential translation of the parameters is represented in 6.5. They indicate that the genre variables are significant in terms of explaining movie choice decisions, with the variables "children" and "western" being the weakest factors in terms of estimation precision, but still significant. Given that movie rating data in the MovieLens computer systems are

mainly entered by adults, the first seems like a natural result while westerns are a less common genre.

Collins & Lanza (2010) suggest two dimensions for judging latent class estimation results, namely homogeneity and class separation. A model is considered homogeneous when members of that class are likely to provide the same observed response pattern or that one response pattern stands out as very likely for that specific latent class. Separation refers to the fact that a response pattern, having a large probability of occurrence conditional on one latent class, shows smaller probabilities conditional on any of the other latent classes. The classification tables suggest that the genre based model is well separated. Cases in the modal are precisely allocated to particular classes ex post. The model scores less good in terms of homogeneity, but the response patterns are sufficiently pronounced to allow to label the latent classes in a meaningful way.

The first class strongly features the parameter "adventure". This variable has as such not a great separation value as it is a manifest and significant parameter for the seven classes, but it is more pronounced for classes 1 and 6. The global profiles of the two latter classes are however substantially different. Class 1 is characterized by a genre profile that combines terms like adventure, comedy, romantic. Additionally emphasizing words as children, fantasy and musical, it profiles a consumer segment targeted at family movies. There is an overall like for all genres, except for western, and therefore this group can be labelled as "genre omnivores". In contrast, class 6, scoring higher on homogeneity, has 4 factors lightening out, namely action, adventure, sci-fi and war, the last to a lesser but still significant degree. This latent class therefore points to a segment of consumers holding a preference for adventure-action type movies. They indicate through their profile to dislike documentaries and drama. Profile 6 also shows similarities with 2, however class 2 combines an interest in crime with positive affection towards mystery and a distinguishing lower interest in action and sci-fi than the agents of group 6.

When comparing and describing the nature of the classes this way, the strength of the LCLR method is revealed: where in a one dimensional log linear regression, the

adventure variable might have been pinned down as important, being significant for all groups, here it is shown that the effect on consumer decisions is different when this factor is taken in combination with alternating constellations of genres. It teaches that not one factor in isolation drives the decision making process for movies but rather shifting sets of features such as represented in the various classes.

Class 3 is characterized by a high odds ratio on the genre factor "film noir". The value of the parameter equals 4,2 which is the highest value in the analysis. Apart from film noir, this group of consumers shows a more explicit preference for documentaries and drama than do members who belong to the other segments. One could label this latent class as "value added seekers". Members behave opposite to group 6, clearly showing less interest in action, war and sci-fi compared to others. At first glance, they might seem like a niche group. However, table 6.2, showing class sizes in terms of $P(c)$, indicates that class 3 stands for a significant 15 percent of cases. Another apparent niche class is 7, which is typified by low homogeneity and separability with only a limited number of parameters standing out. When looking more closely, this appears to be the only segment with a positive estimation value connected to the genre "western". While also positive towards war and thriller, this subgroup holds negative emotions towards factors such as animation, children, documentary and drama. This particular class is small in size, representing less than 5 percent of the cases, but is well targeted within the class. The profile mean indicates that for the class, the number of hits equals 20 percent. This percentage represents the class specific success probabilities for the binomial counts. Table 6.2 gives us the important insight that, while a class might be bigger in size, it is not necessarily stronger in terms of the proportional number of successes observed within this class or vice versa. Niche classes can be identified which are small but well targeted. The profile of the 5th class separates only slightly from segment 3. Individuals in the latter show a high interest in film noir, a positive though lower interest in documentary, combined with a more pronounced preference for the genres thriller and war. Both classes hold a negative attitude towards action movies. Class 4 shows a more distinct profile, endowed with a higher preference for comedy and romance, while disliking documentary, drama and animation. They represent a consumer

segment focused on feel-good/boredom avoidance. Like group 3, they represent about 15 percent of consumers and the estimated success rate equals 12,8 percent.

	Class1	Class2	Class3	Class4	Class5	Class6	Class7
Class Size	0,2359	0,2058	0,15	0,1441	0,122	0,0994	0,0428
Mean	0,0351	0,0725	0,0686	0,1281	0,1253	0,0885	0,2084

Table 6.2: LCLR 7 CLASSES GENRE BASED estimated class sizes

An important observation is that some genres unite while others seem to separate agents. Some variables, such as adventure, comedy, crime, fantasy, film-noir, musical, romance, thriller and war do have a positive effect on the log odds for movie choice for all segments. Others seem to be far more divisive. This is particularly the case for action, generating an estimated coefficient of 0.8 for one group going to -0.12 for another segment. Genres like animation and documentaries are a positive factor in the decisions of only a limited number of segments while horror seems to have an overall negative coefficient for all of the 7 subgroups.

To conclude, when relating the decision making process on one dimension only, namely genre, the analysis demonstrates the advantages of the LCA methodology in terms of identifying consumer segments. A singular log-linear model of movie choice decisions for individuals would result in overall highly positive estimates for some of the genre parameters and lower or negative evaluations for others. However, introducing latent classes not only improves the likelihood values, it makes the interpretation more meaningful. A group of individuals can share with others an interest in a particular genre while simultaneously being substantially different in his/her liking for others. It is the connection of variables that typifies consumer behaviour, not the variables in separate. Some individuals do clearly show niche preferences while others show omnivore consumer behaviour. The last group has slightly higher preference for some genres but is positive on most of the variables. The first group turns visible if the number of classes increases. Results additionally reveal opposite profiles, with manifest like and clear dislike. Some groups show a major dislike for

most genres except for a limited number of variables that are distinguishing positive.

Overall, augmenting the number of classes increases model fit in terms of the Bayesian Information Criterion. The most manifest improvement occurs at the beginning when adding 2 to 4 classes. Here, the number of classes was cut off there where the marginal improvement started to stabilize, which was not the minimum of BIC. When taking the number of segments too high in this type of exercise, parameters start to be unidentified. After all, when increasing the number of latent classes for this model from 1 to 10, the number of variables to be estimated raises from 19 to 199. Therefore, estimating this type of models is always a trade-off between gaining information, keeping the number of variables under control and providing results which are open for interpretation. The latter however is less important when the model is entered into a recommendation model where what matters is that prediction errors decrease, whether it can be rationally explained or not. However, when producers want to target certain segments, for example through advertising, then the ability of a model to be expressed in meaningful segments clearly becomes relevant.

As explained, genre is limited when thinking of potential features that could influence consumer decision behaviour for movies. One potential point of criticism on using only that variable, is that some correlations influencing the shape of the segments were initiated by supply restrictions rather than by the choice factors of the consumers. Romantic comedies is a specific movie subcategory, hence the combination of both variables is likely to occur in a regression. However, the set of movies is sufficiently broad to offer a diversity of connector combinations that might occur. This does not take away that the analysis could improve by adding features having less a priori ties. The next analysis will apply a similar methodology, this time based on a set of movie tags that have been attached by consumers. Some of those tags also relate to genre, but other express alternative motives influencing the individual's decision making process.

7-Class Binomial Regression Model								
Number of cases	470							
Number of replications	1040580							
Number of parameters (Npar)	139							
Random Seed	181991							
Best Start Seed	782999							
Chi-squared Statistics								
Degrees of freedom (df)	331	p-value						
L-squared (L)	551597,445	3,1e-119178						
X-squared	1,68E+111	0,00E+00						
Cressie-Read	1,95E+75	0,00E+00						
BIC (based on L)	549560,8905							
AIC (based on L)	550935,445							
AIC3 (based on L)	550604,445							
CAIC (based on L)	549229,8905							
SABIC (based on L)	550611,4208							
Dissimilarity Index	1							
Log-likelihood Statistics								
Log-likelihood (LL)	-278690,5069							
Log-prior	-2,3979							
Log-posterior	-278692,9047							
BIC (based on LL)	558236,2436							
AIC (based on LL)	557659,0137							
AIC3 (based on LL)	557798,0137							
CAIC (based on LL)	558375,2436							
SABIC (based on LL)	557795,0843							
Classification Statistics	Classes							
Classification errors	0,0163							
Reduction of errors (Lambda)	0,9787							
Entropy R-squared	0,9789							
Standard R-squared	0,972							
Classification log-likelihood	-278708,842							
Entropy	18,3351							
AWE	559545,1437							
ICL-BIC	558272,9138							
Classification Table	Modal							
Latent	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Total
Class1	109,3081	0,9237	0,7321	0	0	0,0187	0	110,9825
Class2	0,1813	93,8394	1,8962	0,1929	0,0005	0,656	0	96,7663
Class3	0,5107	0,9905	68,2633	0,001	0,7337	0	0	70,4991
Class4	0	0,1058	0,0057	67,4842	0,133	0,0003	0	67,729
Class5	0	0,0001	0,1029	0,0832	57,1328	0	0	57,319
Class6	0	0,1404	0	0,2302	0	46,325	0	46,6957
Class7	0	0	0	0,0085	0	0	20	20,0085
Total	110	96	71	68	58	47	20	470
Prediction Statistics								
DECISION								
Error Type	Baseline	Model	R					
Squared Error	0,0777	0,0734	0,0551					
Minus Log-likelihood	0,2906	0,2668	0,0818					
Absolute Error	0,1554	0,1468	0,0556					

Table 6.3: LCLR 7 CLASSES GENRE BASED estimation statistics

Number of cases=number of individuals, number of replications=number of individuals times the number of movies, Log-Likelihood statistics where IC=Information Criterion, Classification statistics with assignment of cases to class with highest probability (modal), AWE=Approximate Weight of Evidence, Prediction statistics: error depends on difference between observed and predicted response, Baseline=average predicted response, R=Reduction of error

Model													
	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Overall					
R	0.013	0.026	0.017	0.030	0.022	0.083	0.059	0.055					
DECISION	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Wald	p-value	Wald(=)	p-value	Mean	Std.Dev.
Intercept	-3.940	-3.039	-3.351	-2.382	-2.504	-2.810	-1.809	55496.108	0.000	2573.636	0.000	-3.063	0.610
Predictors	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Wald	p-value	Wald(=)	p-value	Mean	Std.Dev.
Action	0.253	0.386	-0.088	0.315	-0.120	0.826	0.533	1404.683	0.000	644.564	0.000	0.261	0.273
Adventure	0.866	0.721	0.746	0.677	0.577	0.826	0.606	3709.130	0.000	59.704	0.000	0.741	0.096
Animation	0.034	-0.114	-0.035	-0.309	-0.199	-0.173	-0.089	73.718	0.000	28.534	0.000	-0.110	0.113
Children	0.152	0.012	-0.018	0.104	-0.188	0.021	-0.012	27.938	0.000	26.616	0.000	0.029	0.103
Comedy	0.272	0.080	0.257	0.358	0.061	0.093	0.240	598.616	0.000	144.178	0.000	0.198	0.109
Crime	0.248	0.301	0.324	0.078	0.270	0.192	0.251	467.238	0.000	52.988	0.000	0.243	0.077
Documentary	0.102	-0.413	0.443	-0.648	0.153	-1.061	-0.843	503.185	0.000	407.974	0.000	-0.211	0.478
Drama	0.132	0.033	0.609	-0.114	0.535	-0.438	-0.128	1283.361	0.000	1200.971	0.000	0.129	0.316
Fantasy	0.512	0.376	0.396	0.319	0.303	0.494	0.368	832.924	0.000	29.164	0.000	0.405	0.078
FilmNoir	1.141	0.967	1.435	0.675	1.204	0.797	0.814	591.317	0.000	31.862	0.000	1.042	0.240
Horror	-0.771	-0.696	-1.314	-0.524	-0.942	-0.067	0.007	1147.084	0.000	393.929	0.000	-0.719	0.361
Musical	0.498	0.312	0.384	0.206	0.181	0.350	0.012	220.386	0.000	37.411	0.000	0.326	0.128
Mystery	0.463	0.492	0.410	0.339	0.453	0.266	0.259	692.892	0.000	27.889	0.000	0.414	0.078
Romance	0.335	0.176	0.361	0.411	0.274	0.016	0.198	961.721	0.000	131.709	0.000	0.272	0.117
SciFi	0.374	0.407	0.196	0.299	0.170	0.722	0.440	822.648	0.000	135.934	0.000	0.356	0.151
Thriller	0.211	0.459	0.137	0.470	0.301	0.546	0.570	1548.097	0.000	166.071	0.000	0.348	0.149
War	0.342	0.630	0.188	0.472	0.409	0.633	0.557	628.661	0.000	38.986	0.000	0.443	0.154
Western	-0.818	-0.286	-0.727	-0.270	-0.108	-0.393	0.029	75.632	0.000	26.561	0.000	-0.451	0.279

Table 6.4: LCLR 7 CLASS GENRE BASED parameter estimates

	Class1	Class2	Class3	Class4	Class5	Class6	Class7
Predictors							
Action	1.288	1.470	0.915	1.370	0.887	2.285	1.704
Adventure	2.378	2.057	2.108	1.968	1.781	2.285	1.834
Animation	1.034	0.892	0.966	0.734	0.820	0.841	0.915
Children	1.164	1.012	0.982	1.109	0.829	1.021	0.988
Comedy	1.312	1.084	1.293	1.430	1.063	1.097	1.272
Crime	1.281	1.351	1.383	1.082	1.310	1.211	1.286
Documentary	1.108	0.662	1.557	0.523	1.165	0.346	0.431
Drama	1.141	1.034	1.839	0.892	1.708	0.645	0.880
Fantasy	1.669	1.456	1.486	1.376	1.354	1.638	1.445
FilmNoir	3.130	2.630	4.201	1.965	3.333	2.219	2.257
Horror	0.462	0.498	0.269	0.592	0.390	0.935	1.007
Musical	1.646	1.366	1.467	1.228	1.199	1.420	1.012
Mystery	1.589	1.636	1.506	1.403	1.574	1.304	1.295
Romance	1.398	1.192	1.435	1.508	1.315	1.016	1.219
SciFi	1.453	1.502	1.216	1.348	1.185	2.058	1.553
Thriller	1.235	1.583	1.146	1.600	1.352	1.727	1.768
War	1.407	1.878	1.206	1.603	1.506	1.884	1.745
Western	0.441	0.751	0.483	0.764	0.897	0.675	1.029

Table 6.5: LCLR 7 CLASS GENRE BASED exponential parameter estimates

6.4.3 The Latent Class Tag Based Decision Model

This paragraph, as one of the central parts of this thesis, covers the stages and results of performing a Bayesian Latent Class Regression on the tag sets connected to the movies. The dependent variable consists of the decision to rate a movie, as a proxy of consumer choice. The explanatory factors are tags that have been attached to the items, partially expressing what individuals consider determinants of their

choices. Once a tag is connected to an item, it is further considered as a feature that the movie is "endowed with", irrespective of the individual who induced the tag-item relationship. The same holds for the genre variable. A number of genre features are part of the set of tags. Once that genre was tagged by at least one individual, the movies were enriched with the genre information of the MovieLens data, meaning that the explanatory variables represent hybrid information of "once tagged and given a tagged genre". To avoid that the number of independent variables gets too high, only tags that appear in at least 40 movies were kept for analysis. This resulted in a selection of 43 explanatory variables.

The pattern of BIC values, as before, is put in relation to the number of classes. The design is quite similar to that of the genre categories. Here also, the 7 segments solution is chosen, for the same reason as before, namely it balances information gain, a manageable number of degrees of freedom with the potential to provide a meaningful interpretation. Moreover, it is important to maximize comparability with the genre-based model, as it is one of the purposes of this exercise to investigate if tag based information outperforms genre to predict consumer choice. From 11 segments on, the model suffers from under-identification as the number of parameters to be estimated becomes too high. Therefore, table 6.6 only goes up to 10 classes.

	LL	BIC(LL)	Npar	L	R
1-Class Regression	-273151,7024	546574,125	44	540519,836	0,0707
2-Class Regression	-263752,6993	528052,9919	89	521721,83	0,0933
3-Class Regression	-261092,8896	523010,2455	134	516402,2106	0,0996
4-Class Regression	-258994,42	519090,1792	179	512205,2713	0,1074
5-Class Regression	-257927,7201	517233,6524	224	510071,8715	0,1107
6-Class Regression	-257084,5794	515824,2439	269	508385,5901	0,1135
7-Class Regression	-256507,963	514947,884	314	507232,3572	0,1155
8-Class Regression	-255997,3759	514203,5828	359	506211,1831	0,1171
9-Class Regression	-255602,7819	513691,2678	404	505421,995	0,1186
10-Class Regression	-255313,864	513390,305	449	504844,1593	0,1196

Table 6.6: LCLR TAG BASED class iteration

LL=Loglikelihood, BIC=Bayesian Information Criterion, Npar=Number of Parameters, L=Likelihood, R-value=Reduction of error

Overall, the results show a profile which is in many ways similar to previous results in terms of fit and accuracy statistics. Like for the genre based model, it is clear that introducing segmentation leads to information gain compared to a one class model. The classification matrix of the overall results (table 6.8) shows that items are mainly assigned to the diagonal cells rather than to the off diagonal cells, signalling an acceptable class segregation. Comparing current classification table with the results of the previous paragraph shows that the tag based model has more off diagonal null cells and therefore is expected to be better separated. This is also reflected in the error statistics, although very similar in magnitude, the absolute error in the tag model equals 0,1365 versus 0,1468 for the genre model.

Comparing the tag and genre based model shows that the BIC, AIC and CAIC are close in value, yet there is a clear difference. For the tag-based model, the BIC values range between 513.000 to 547.000 where for the genre based model they vary in the interval [556.000 587.000]. This implies that, despite the fact that the number of estimation variables for the tag model is substantially higher, namely 314 versus 139 in previous model, and given that all information criteria introduce a penalty factor for an increase in the number of variables, the tag-based model still scores better. The tag matrix is also far more sparse than the genre matrix. Yet, the general log likelihood raises from -278.690 to -256.507. Combined with the lower error values of prediction, it is legitimate to state that the tag based consumer model performs better as prediction model for movie choice decisions.

Looking at the estimation results of the tag based model, shown in table 6.9, indicates that most of the parameter estimates are significant, with p-values suggesting significance at 0.05 as well as 0.01 level. The variable PG does worst in terms of significance, but still passes under both significance levels. PG stands for Parental Guidance, provided by the British, American or Australian board of Film classification. When looking at the variables against the hypothesis that all variables are equal between classes, some fail acceptance levels at 0.05 significance rate, superhero, topless, history, teen and atmospheric. The highest Wald levels are for the variables list, funny, sequel, Oscar, based on, completed with a number of genre variables that were also important in previous exercise such as adventure, action,

drama, and thriller. This does, not surprisingly, largely agree with the dimensions detected in the Latent Dirichlet Allocation, a three level hierarchical Bayesian model applied on an elaborated set of tags. Expressed in terms of log-odds, the values are close to 1, however, some ratio values exceed two or stand out as more pronounced.

Like is the case for the genre model, separation values are good and the homogeneity is sufficient to adequately distinguish tendencies in the observed segments. The parallel with the LDA translates into the identity of the observed latent classes. The first two classes can be considered to represent dimensions related to information seeking behaviour. The highlighted variables here are sequel, list and Oscar, having odds ratios above 2. Also the style factor is more pronounced for this segment. As explained in previous chapters, creative goods are often described as "experience goods". Only after consumption, the individual can make a full judgement of his/her valuation of the good. Agents can diminish their uncertainty factor by seeking prior information, which can be provided by experts through reviews, by the movie appearing on recommendation lists, or by word of mouth. The uncertainty reduction can also be a result of own experience resulting in a movie of the same "type" being chosen. This is particularly the case for sequels where the repetition factor makes that a number of uncertain elements can be filled in as consumers partly fall back on their knowledge of the previous episode. Also the fact that peers or a panel of experts assert a quality label under the form of a prize, reward or Oscar can be considered an element of uncertainty reduction which influences individuals' choice, mainly persons either lacking experience or those characterized by high uncertainty avoiding behaviour. It is noticeable that for this group of information seekers, the movie being endowed with a PG or R-Rating does have a rather negative impact on decision making. Overall however, the information variables have a significant positive impact. Class 2 also distinguishes from class 1 in several ways. Elements of engagement and knowledge search are somewhat more pronounced for this segment with tags such as political, documentary and plot receiving higher values. They can be seen as a group of individuals seeking added value. The two groups are the biggest in terms of class size, representing 23,69 and 21,09 percent in share of the total.

	Class1	Class2	Class3	Class4	Class5	Class6	Class7
Class Size	0,2369	0,2109	0,1689	0,159	0,0857	0,0852	0,0534
Mean	0,0352	0,068	0,0735	0,1099	0,1262	0,1267	0,2016

Table 6.7: LCLR 7 CLASSES TAG BASED estimated class sizes

One of the most influential factors for group 3 is the fact that the movie casts an important actor. It also shows in the positive evaluation of the term acting. The variable "have actor" has a value 1 when the actor mentioned is either: Johnny Depp, Jude Law, Charlie Kaufman, Steve Carell, Martin Scorsese, Bruce Willis, Tom Hanks, Nicolas Cage, Jim Carrey or Woody Allen. They represent a selection of actors frequently named in the tagging information. Here, only the tagging information is used, or the variable haveactor gets a value 1 if at least one user made the connection. However, in an enhanced recommendation system, the true actor information could be added in the same way as was done for the genre information in this analysis, turning it into an augmented hybrid tag based system. It is yet another way to redefine that variable, not necessarily facing the shortcomings of other approaches listed in the literature review. The actor variable here included is overall significant and important to the members of most classes. Apart from attaching great value to the presence of a famous actor, members of class 3 value the fun factor, while simultaneously expressing a preference for action movies, a segment also labelled in the previous genre-based analysis. Their decision to watch a movie can be connected to tags as action, war, sci-fi, violence and thriller. Being more into action, negative values are expressed to engagement movies, with documentary, drama, gay and drugs getting negative values as do tags such as romance and animation. They attach less value to the fact that the movie is "based on" compared to the other groups.

Segments 4 and 7 are difficult to separate. Like class 3, their members value acting, but are both distinct in their preference for movies directed by Quentin Tarantino. The first group values content and information elements more, with tags such as list, political, quirky and based on being more pronounced, as are the sex and nudity variables. Group 7 on the other hand seems to be less negative towards drugs and violence, is more into action movies and shows more dislike towards documentaries.

This is a small group of slightly over 5 percent. Latent class 6 displays an odds ratio higher than 2 for the fun factor, combining it with the tags comedy and romance. For this segment, the "feel good" factor seems crucial. Members of this group combine the fun elements with negative appreciation for variables such as drama, crime, drugs and documentary as well as gay, nudity and sex. Apart from romantic comedies, this group is also into the thriller genre. Group 5 also appreciates more romantic movies, but adds content features to it. The "based on" factor lights out, as do tags such as politic, drama, murder, plot and ending. This group shows negative towards drugs and violence and rather dislikes action and sci-fi movies. Also the odd ratio on "family" is very high.

Like for the genre based model, it is important to point to the unifying as well as the divisive nature of the variables presented for the analysis. Some variables are clearly significantly positive over all groups. This is the case for tags such as sequel, based on, fun, being Oscar nominated or having a top actor casted in the movie. However, contrary to a uni-dimensional regression, the latent class analysis shows how those factors, whilst positive, effect decisions in a more or less pronounced way. Other variables separate proponents and opponents. This is the case for themes such as sex, gay, violence or drugs. As stated before, also certain genres make the segments have a divisive impact. Documentaries, thrillers, drama or fantasy movies are motivating factors in the choice profile of a part of the consumers while having an outspoken negative impact on the choice behaviour of other groups.

The tag based model outperforms the genre based model, not only in terms of statistical performance, but even more so in terms of interpretation. The descriptive statistics of tagging already indicated that, while genre is often named by users, a lot of other keywords are attached to films. Those keywords are largely related to the experience good characteristics of movies and they translate into other types of dimensions that can be outlined. Combined with genre, tags allow clearer profiling. Information seekers represent a preference profile as do those emphasizing the importance of certain genres such as action or war movies. However, when making decisions, all elements come into play, to a lesser or larger extent, with positive or negative impact. It is therefore important to look at the choice for creative products

as a decision over multidimensional features sets.

6.5 Interim Conclusions

This chapter presented some of the core results of this PhD work. Using a latent class logistic regression model, the decision to watch a movie, proxied by the item being evaluated in MovieLens, was investigated in relation to a set of prominent tags. In both the choice of dependent variable and the explanatory variables, the model differs from the probabilistic hybrid models used in recommender literature. Imposing latent classes clearly shows valuable to discover heterogeneity in consumer profiles. Increasing the number of classes improves accuracy as expressed in terms of the BIC criterion. The segmented model therefore outperforms a uni-dimensional model. The higher the number of segments, the more also that "niche groups" turn visible. For this type of models however, there is always a trade-off between increasing information and the identification of the model.

The segments generated appear to be meaningful. They reveal classes that can be labelled by feature combinations being more or less pronounced. That way, profiles or consideration sets can be outlined. The conclusions are not unlike those of the prospective LDA analysis. Some profiles are affirmed, such as "information seekers", "value added seekers" and "action or fun seekers". Importantly, the decisive factors put forward in economics and the features attached to movies in the recommender literature are not opposed. They give rise to different profiles. However, this type of analysis permits to be more fine grained. It teaches that not a single variable is important in explaining consumer choice, but rather the interplay between features. A segment might have in common with others that it includes fun or comedy while at the same time being very different in its appreciation towards action or content. Some characteristics are overall positively appreciated, be it more pronounced in certain segments, others however are more divisive, very much liked by some groups and clearly disliked by others. It partly explains that in a econometric model, working with a representative consumer, results can be inconclusive or turn out to

be very dependable on the group that is questioned. For the movie industry, the results point to the importance of targeting consumers. For economic theory, the manifest presence of segments challenges the representative consumer hypothesis.

The presented model supports the view of creative products sketched at the beginning. A creative product can be seen as a bundling of various decision making features of different nature, objective and subjective. Consumers categorize and each time a decision is made, an individual compares if the new item possesses to a larger or lesser extent the features that she/he values. In a way, the segments or profiles can therefore be interpreted as a kind of consideration sets. Technically, it is possible to tract an individual back to the segment she/he is most likely related to. Classification statistics provided by the latent regression estimation contain the predictions of what latent classes users belong to, given their observed values for y and z, which are the posterior class membership probabilities. In case of good separation of classes, individuals will ex post be related to one, sometimes to a few classes.

While certainly showing clear advantages, the latent class methods have to be interpreted with care. The sample used for this thesis is big in comparison to many survey based investigations, it is not of a size to calculate all errors away. Moreover, the information provided for empirical estimation is minimal, a few choices made by individuals on a large portfolio of movies, placed in relation to features attached to objects. The features are tentative, imperfect proxies. They are what some individuals themselves indicated as keywords related to items. Unlike econometric analysis, the parameter values are therefore not to be interpreted in an absolute way. When estimation techniques are sample based, alternating simulation values can result in slightly different parameter values. However, the results generated with LCLR and with LDA show similar patterns and results are powerful when it comes to detecting dimensions and segmenting the group of users.

7-Class Binomial Regression Model								
Number of cases	470							
Number of replications	1040580							
Number of parameters (Npar)	314							
Random Seed	1040580							
Best Start Seed	738740							
Chi-squared Statistics								
Degrees of freedom (df)	156	p-value						
L-squared (L)	507232,3572	7,1e-109842						
X-squared	1,64E+111	0,00E+00						
Cressie-Read	1,90E+75	0,00E+00						
BIC (based on L)	506272,5309							
AIC (based on L)	506920,3572							
AIC3 (based on L)	506764,3572							
CAIC (based on L)	506116,5309							
SABIC (based on L)	506767,6449							
Dissimilarity Index	1							
Log-likelihood Statistics								
Log-likelihood (LL)	-256507,963							
Log-prior	-2,4251							
Log-posterior	-256510,3881							
BIC (based on LL)	514947,884							
AIC (based on LL)	513643,9259							
AIC3 (based on LL)	513957,9259							
CAIC (based on LL)	515261,884							
SABIC (based on LL)	513951,3084							
Classification Statistics	Classes							
Classification errors	0,0062							
Reduction of errors (Lambda)	0,9918							
Entropy R-squared	0,99							
Standard R-squared	0,9883							
Classification log-likelihood	-256516,627							
Entropy	8,664							
AWE	517839,1701							
ICL-BIC	514965,212							
Classification Table	Modal							
Latent	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Total
Class1	111,1608	0,256	0,0101	0	0	0	0	111,4269
Class2	0,7187	97,9153	0,2958	0,2755	0,0036	0	0	99,2089
Class3	0,1206	0,5259	78,6111	0,1431	0	0,0002	0	79,4009
Class4	0	0,1251	0,0778	74,467	0,0523	0,024	0	74,7462
Class5	0	0,1777	0	0,1003	39,944	0	0	40,222
Class6	0	0	0,0052	0,0017	0	39,9758	0	39,9828
Class7	0	0	0	0,0124	0	0	24,9999	25,0124
Total	112	99	79	75	40	40	25	470
Prediction Statistics								
DECISION								
Error Type	Baseline	Model	R					
Squared Error	0,0777	0,0687	0,1155					
Minus Log-likelihood	0,2906	0,2456	0,155					
Absolute Error	0,1554	0,1364	0,1221					

Table 6.8: LCLR 7 CLASSES TAG BASED estimation statistics

Number of cases=number of individuals, number of replications=number of individuals times the number of movies, Log-Likelihood statistics where IC=Information Criterion, Classification statistics with assignment of cases to class with highest probability (modal), AWE=Approximate Weight of Evidence, Prediction statistics: error depends on difference between observed and predicted response, Baseline=average predicted response, R=Reduction of error

Model													
	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Overall					
R^2	0.061	0.089	0.100	0.115	0.087	0.088	0.106	0.116					
DECISION	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Wald	p-value	Wald(=)	p-value	Mean	Std.Dev.
Intercept	-4.406	-3.806	-3.332	-3.108	-2.896	-2.665	-2.146	79892.488	0.000	3072.424	0.000	-3.493	0.657
Predictors	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Wald	p-value	Wald(=)	p-value	Mean	Std.Dev.
basedon	0.450	0.480	0.250	0.465	0.471	0.388	0.400	1592.885	0.000	51.870	0.000	0.419	0.081
superhero	0.573	0.427	0.366	0.487	0.416	0.479	0.423	336.024	0.000	6.896	0.330	0.464	0.072
oscar	0.732	0.769	0.380	0.570	0.575	0.230	0.345	1809.731	0.000	165.333	0.000	0.578	0.182
nudity	-0.130	0.154	0.184	0.245	0.251	-0.051	0.093	91.304	0.000	40.300	0.000	0.094	0.147
sequel	1.026	0.845	0.887	0.930	0.749	0.714	0.797	2224.056	0.000	26.488	0.000	0.887	0.099
war	0.282	0.128	0.477	0.290	0.102	0.233	0.198	243.436	0.000	44.562	0.000	0.260	0.120
list	0.835	0.819	0.596	0.737	0.631	0.493	0.595	3432.419	0.000	88.896	0.000	0.716	0.117
funny	0.647	0.585	0.804	0.715	0.466	0.705	0.794	3214.514	0.000	70.763	0.000	0.669	0.097
ending	-0.104	0.027	-0.014	0.170	0.280	0.213	0.203	77.345	0.000	43.162	0.000	0.059	0.129
comedy	0.220	0.134	-0.047	0.053	0.036	0.353	0.162	283.934	0.000	139.288	0.000	0.123	0.115
seks	0.540	0.445	0.239	0.417	0.397	-0.016	0.163	414.914	0.000	72.039	0.000	0.370	0.162
murder	0.498	0.509	0.390	0.581	0.493	0.242	0.431	780.240	0.000	33.699	0.000	0.470	0.090
animation	0.135	0.075	0.005	0.043	-0.177	-0.242	-0.170	52.229	0.000	50.728	0.000	0.011	0.122
PG	-0.139	-0.067	-0.005	0.062	-0.001	0.042	0.124	18.869	0.009	18.867	0.004	-0.028	0.079
HaveActor	0.477	0.456	0.772	0.623	0.408	0.628	0.556	1543.825	0.000	63.866	0.000	0.557	0.120
topless	0.432	0.290	0.258	0.222	0.219	0.343	0.379	188.434	0.000	10.453	0.110	0.311	0.080
politic	0.575	0.755	0.355	0.671	0.600	0.339	0.558	656.683	0.000	34.336	0.000	0.572	0.146
tarantino	-0.509	0.304	0.138	0.699	-0.063	-0.031	0.636	39.740	0.000	33.180	0.000	0.104	0.416
fantasy	0.072	-0.043	0.083	0.037	-0.126	0.106	0.048	28.429	0.000	23.650	0.001	0.029	0.068
action	0.029	-0.141	0.674	0.126	-0.326	0.455	0.395	1014.069	0.000	726.221	0.000	0.143	0.314
quirky	0.500	0.707	0.294	0.686	0.524	0.021	0.335	829.869	0.000	128.703	0.000	0.491	0.207
rrated	-0.167	-0.059	-0.148	-0.063	0.110	-0.265	-0.079	97.128	0.000	57.729	0.000	-0.104	0.090
drama	0.103	0.446	-0.298	0.105	0.536	-0.303	-0.078	1005.908	0.000	943.503	0.000	0.101	0.290
sci-fi	0.336	0.248	0.449	0.403	-0.062	0.371	0.187	512.902	0.000	82.670	0.000	0.308	0.136
crime	0.135	0.208	0.127	0.140	0.083	-0.122	0.090	128.305	0.000	54.218	0.000	0.121	0.083
children	0.003	-0.056	-0.183	0.024	-0.158	0.214	-0.097	49.571	0.000	47.176	0.000	-0.039	0.107
romance	0.249	0.296	-0.036	0.221	0.417	0.387	0.286	848.186	0.000	167.128	0.000	0.235	0.135
adventure	0.512	0.493	0.521	0.528	0.277	0.469	0.397	1507.176	0.000	35.353	0.000	0.482	0.069
documentary	0.257	0.456	-0.743	0.054	0.306	-1.010	-0.581	506.350	0.000	506.150	0.000	-0.051	0.513
thriller	0.087	0.078	0.474	0.300	0.107	0.482	0.440	977.570	0.000	218.302	0.000	0.239	0.172
family	0.569	0.596	0.244	0.434	0.748	0.243	0.323	653.383	0.000	73.548	0.000	0.473	0.163
style	0.512	0.416	0.291	0.174	0.259	0.118	-0.001	224.299	0.000	49.465	0.000	0.318	0.153
musical	0.417	0.251	0.241	0.264	0.090	0.068	0.052	127.806	0.000	28.353	0.000	0.251	0.117
history	0.278	0.216	0.290	0.263	0.235	0.150	0.270	144.704	0.000	3.842	0.700	0.250	0.040
newyork	0.361	0.508	0.264	0.491	0.625	0.288	0.462	468.936	0.000	29.171	0.000	0.418	0.112
violence	-0.094	-0.303	0.100	-0.073	-0.434	0.173	0.026	77.132	0.000	64.117	0.000	-0.102	0.180
gay	0.023	0.015	-0.220	-0.095	0.088	-0.360	-0.098	30.563	0.000	24.235	0.000	-0.072	0.130
plot	0.253	0.347	0.158	0.244	0.355	-0.015	0.365	146.923	0.000	26.461	0.000	0.247	0.105
teen	0.261	0.375	0.302	0.351	0.207	0.224	0.302	163.207	0.000	6.019	0.420	0.301	0.057
acting	0.115	0.257	0.409	0.444	0.177	0.557	0.424	285.370	0.000	40.403	0.000	0.306	0.147
author	-0.073	-0.041	0.235	0.147	0.056	0.236	0.186	45.932	0.000	26.866	0.000	0.072	0.127
drug	-0.347	-0.161	-0.244	-0.085	-0.083	-0.463	0.123	66.562	0.000	32.211	0.000	-0.211	0.143
athmospheric	0.197	0.193	0.256	0.281	0.158	0.120	0.178	82.874	0.000	4.762	0.570	0.209	0.047

Table 6.9: LCLR 7 CLASSES TAG BASED parameter estimates

Predictors	Class1	Class2	Class3	Class4	Class5	Class6	Class7
basedon	1.568	1.616	1.284	1.591	1.601	1.474	1.492
superhero	1.774	1.533	1.441	1.628	1.515	1.614	1.527
oscar	2.079	2.157	1.463	1.768	1.778	1.259	1.412
nudity	0.878	1.167	1.202	1.277	1.285	0.950	1.098
sequel	2.790	2.328	2.429	2.535	2.114	2.043	2.219
war	1.326	1.136	1.611	1.336	1.108	1.262	1.219
list	2.304	2.268	1.816	2.089	1.880	1.637	1.812
funny	1.909	1.795	2.235	2.045	1.594	2.024	2.212
ending	0.901	1.027	0.986	1.186	1.324	1.238	1.225
comedy	1.246	1.143	0.954	1.055	1.037	1.424	1.176
seks	1.716	1.560	1.270	1.518	1.487	0.984	1.177
murder	1.646	1.664	1.476	1.787	1.638	1.273	1.539
animation	1.145	1.078	1.005	1.044	0.838	0.785	0.843
PG	0.870	0.935	0.995	1.064	0.999	1.042	1.132
HaveActor	1.611	1.578	2.163	1.865	1.504	1.874	1.744
topless	1.540	1.336	1.294	1.249	1.245	1.409	1.461
politic	1.777	2.128	1.426	1.956	1.822	1.404	1.748
tarantino	0.601	1.356	1.148	2.011	0.939	0.969	1.889
fantasy	1.074	0.958	1.086	1.037	0.882	1.112	1.049
action	1.029	0.868	1.962	1.134	0.722	1.577	1.485
quirky	1.649	2.029	1.342	1.987	1.689	1.021	1.398
rrated	0.846	0.943	0.863	0.939	1.116	0.767	0.924
drama	1.108	1.563	0.742	1.111	1.709	0.738	0.925
sci-fi	1.399	1.282	1.566	1.496	0.940	1.448	1.206
crime	1.144	1.231	1.135	1.151	1.087	0.885	1.094
children	1.003	0.946	0.833	1.024	0.854	1.238	0.908
romance	1.283	1.345	0.965	1.247	1.517	1.473	1.331
adventure	1.668	1.638	1.684	1.696	1.320	1.598	1.487
documentary	1.294	1.577	0.476	1.055	1.359	0.364	0.559
thriller	1.091	1.081	1.606	1.350	1.113	1.619	1.552
family	1.767	1.815	1.276	1.543	2.112	1.275	1.381
style	1.668	1.516	1.338	1.190	1.296	1.126	0.999
musical	1.518	1.285	1.272	1.302	1.094	1.070	1.054
history	1.320	1.241	1.337	1.301	1.265	1.162	1.310
newyork	1.435	1.662	1.303	1.633	1.869	1.333	1.587
violence	0.910	0.738	1.105	0.930	0.648	1.188	1.026
gay	1.024	1.015	0.803	0.910	1.092	0.698	0.907
plot	1.288	1.415	1.171	1.276	1.425	0.986	1.441
teen	1.298	1.455	1.353	1.421	1.230	1.250	1.352
acting	1.122	1.293	1.505	1.560	1.194	1.745	1.528
author	0.930	0.960	1.265	1.158	1.058	1.267	1.205
drug	0.707	0.851	0.783	0.919	0.920	0.629	1.131
athmospheric	1.218	1.213	1.292	1.325	1.171	1.128	1.195

Table 6.10: LCLR 7 CLASSES TAG BASED exponential parameter estimates

Chapter 7

Movie Choice in Dynamic Perspective: A Latent Class Markov Model

7.1 Introduction

In previous chapters an argumentation was developed. The reasoning started from the definition of creative products as bundles of characteristics where each element is consulted to a smaller or larger extent by individuals in their decision to consume a movie or not. In both economic theory and computer sciences, a discrete probabilistic choice model is put forward as a potential way to deal with this type of commodity in empirical studies. A Bayesian Latent Class Logistic Regression model was applied to MovieLens data, identifying latent consideration sets. Using tags attached by users as proxies for what are connectors between objects and decisions, a number of dimensions were identified. Some relied more on technical features driving the recommender literature, others showed the importance of information factors, thereby underlining the experience good nature of movies emphasized in economics literature. While recognizing its strengths, certainly in terms of dealing with heterogeneity of taste, the approach is static in nature, assuming taste patterns are constant over time. Given that the analysis was done over a period of six years, this is a reasonable assumption to make. However, even over a short time span, consumption patterns are likely to change; people might suddenly come to prefer

one genre more than another, stop or start valuing certain actors or pay more or less attention to the opinion of others.

The second chapter focused on the description of experience good as a two angled concept. The emotional perspective, referring to the notion of "erlebnis", encompasses the immediate hedonistic effect of consumption, incorporating feelings of arousal or regret. Picturing the pallet of emotions following the experience of artistic or other goods has been very much in the domain of marketing literature. A second interpretation, supporting the vision of experience as "erfahrung", agrees more with a cognitive framework. When a person is seen as experienced, it means she/he accumulated knowledge in the acts of sequentially going through performances. The time aspect is innate to both definitions, one has to experience the commodity to know or to feel the content or emotion, however the second view stresses the learning aspect, the mental process of remembering the past and evoking it for future decision making. Being faced with prior uncertainty, the individual looks back on the consumed item afterwards, makes a judgement and processes the acquired information into a new set of beliefs. These beliefs are the source an individual relies on to make future decisions. An element of cognition and recognition is underlying the static latent class view of the previous chapters. It is hidden in the assumption that individuals are comparing new objects with mental categories they recognize from the past. However, the interpretation of accumulated knowledge goes deeper, as it may encompass change of taste over time, potentially caused by a series of positive or negative evaluations once the uncertainty is lifted. In terms of sociological or economical theories, this process touches what is labelled as cultural or human capital formation. They embody disposition of the mind capturing symbolic elements that steer taste or the capital stock of skills, knowledge, training a person can appeal to when acting. The assumption of human capital formation was used by Stigler & Becker (1977) to model shifts in the appreciation for music. When defining an experience good as a commodity endowed with prior uncertainty, than the stock formation can be seen as a process of belief formation and updating, gradually but not totally removing uncertainty, given that each movie remains a new product, Amez (2003).

The recommender systems literature largely passes by the phenomenon of preference shifts and base future predictions mainly on the observation of the past. When expressing future advances in collaborative filtering, Koren & Bell (2011) acknowledge the importance to take into account the shifts in user preferences over time and incorporate their views into a time changing factor model. Accounting for intertemporality is done by including time drifting parameters into the predictive equations of their SVD++ model. They recognize that the temporal effects are the hardest to capture as they are split over the multiple factors. Real life recommender systems do update prediction models frequently and are therefore likely to capture preference shifts. Results showed that individuals can be attached to classes in a meaningful way. The question is now raised if, taken over a longer period, individuals are faithful to the consideration set associated with them, or whether they are likely to move to other segments.

Latent Class Analysis can be extended by adding an intertemporal layer, by letting the model take the form of a Latent Markov Model. In this type of model, stability or shift is captured by estimating transition probabilities. They express the probability that individuals stay within a category or move from one segment to another. Moreover, the variables or features determining the different classes, such as presented in previous chapter can be separated from the covariates that influence the transition. In the light of the expressed views of experience goods, it would be interesting to see if past period evaluation has a significant influence on the agent's decision to stay loyal to a class or to move. The analysis here differs fundamentally from that found in the recommender literature, in which rating is treated as a predictor, while here it plays a role in evaluation. Thus, choice is based on preferences and beliefs that are updated after the experience. Choice and rating are considered as sequential acts that do not simultaneously occur. The next session will present a Latent Markov Model applied on the same set of users and movies that underlie the results of last chapter. The time span will be split in two blocks to investigate if there is a transition switch between segments and if past rating plays a role in preference shift.

7.2 Latent Class Markov Models

Latent Class Markov Models (LCMM), also known as Hidden Markov or Hidden Transition models, Baum et al. (1970), Collins & Wugalter (1992), Vermunt et al. (1999), like LCA are based on the assumption that a process is characterized by latent states that are not directly visible while a number of linked observed indicators are. Those observed variables are considered at different time points and class participation is not seen as static. Instead, the question is asked with what probability users stay or move to other response categories. LCMM is used to study discrete time longitudinal data, initially for speech pattern recognition, recently they find a broader application in behavioural sciences, Paas et al. (2007).

Say that $t : 0, ..., T$ indexes a sequence of $T + 1$ measurement occasions. Like before, one considers K latent classes. The LCMM specifies the probability of generating a particular array of choices at the $T + 1$ measurements points, given the individual's covariates, Paas et al. (2007). The probability structure has three components: the initial latent state probability and the latent transition probability, designing the evolution of latent segments across time points, together forming the structural part, and a measurement component that ties the particular time points to the observed responses. The latter is similar to the static LCA model, the first is endowed with a first order Markov structure, meaning that class membership at time t is only affected by membership at a previous time point. Formally, one can distinguish: (a) The initial state probability $P(c_0|z_{io})$, denoting the probability of belonging to a particular class at the initial measurement occasion, conditional on that person's feature profile. (b) The latent transition probability $P(c_t|c_{t-1}, cov_{it})$ providing the probability of being at latent state c_{t-1}, thereafter switching to latent class c_t at occasion t, given an individual-specific covariate. (c) The response probability $P(y_{it}|c_t, z_{it})$, referring to the probability of an observation at time point t, given the state and the explanatory variables, Vermunt (2008).

When taken together, the probability of the dependent, given an external variable is:

$$P(y_i|z_i) = \sum_{c_0=1}^{K} \sum_{c_1=1}^{K} ... \sum_{c_T=1}^{K} P(c_0|z_{i0}) \prod_{t=1}^{T} P(c_t|c_{t-1}, cov_{it}) \prod_{t=0}^{T} P(y_{it}|c_t, z_{it}) \quad (7.1)$$

As is the case for previous model, the response model is a logistic regression where choice is dependent only on class membership and the explanatory variables. The latent class membership at time t is conditional only on that of the previous state and potentially also on covariate variables, not on the observed past choices. Here, the importance of opposing choice and rating becomes crucial. Rating turns into an individual specific covariate that can influence the transition probability. In the presented analysis, both a model with and without covariate will be estimated, the variable being the average rating of an individual over the previous period. The choice over different products at t are mutually independent, same way as before.

The data choice and setup are in many ways similar to that in previous exercise. The variable y_i represents choices made over an array of movies. The independent variables, z_{it}, are binary variables, indicating whether that particular tag is associated with the movie. The observations are split in two time points, one covering movies over the period [1999 2002], the second referring to movies between [2003-2006]. The covariate, cov_{it}, the rating variable is the average rating given to movies over the period [1999 2002]. They were rounded to the unit, resulting in a 5 scale nominal variable. If significant, the parameters connected to the covariate will confirm there might be a relationship between a user shifting from one cluster to another and the average rating level. Without covariate, the estimation of the transition probabilities provides a test of the stability of individuals towards certain classes, thus investigating taste dynamics. The investigation is that of bounded rationality, without forward looking behaviour. Compared to the analysis of the previous chapter, the number of classes will be limited to 4 and the number of explanatory variables to 14, taking into account proper identification of the model. For the explanatory variables, a selection was made of those variables that were most pronounced in previous chapter. The estimations are performed using Latent Gold 5.1 software. The aim of next section is to investigate ways to integrate dynamics when dealing with feature based commodities such as creative goods. As stated before, data are

sparse and results have to be interpreted with some care. This last part must be read as an opener for further research.

7.3 LCMM Analysis

Three different models are compared. A LCLR with 4 classes and 14 selected variables, a LCMM without covariates and finally a LCMM with the average rating of the past period as a transition covariate. The first model was added to evaluate the effect of introducing dynamics. Given that the number of classes and variables are not the same as those in previous analysis, one cannot transfer the likelihood statistics. The non-dynamic case thus serves as a benchmark.

The overall estimation results of the LCMM without covariates is shown in table 7.5. As in previous chapter, the number of individual cases is 470 and the number of observations is 1.405.580, however since the database was split over two time periods, it generates 940 time points. Like for the tag based model of chapter 6, the model performs well in terms of separation, with the modal state classification table exhibiting few off diagonal cases. Taking account of the fact that sparsity and size potentially have an impact on the significance levels, the p-value of the model is very low. Also the 14 individual estimated parameters, table 7.6, are significant at 0.01 level, except for the variable war, being at the limit. The measurement component estimates the relationship between features and the dependent for the various states. Overall, in magnitude, the estimated parameters related to information seeking behaviour, such as sequel, list and Oscar are larger compared to the genre related variables. On the other hand, the genre indicators point to a greater diversity in taste, with some classes expressing very low to negative preference towards certain genres. The four classes comply, in large contours, with the groups pointed out in previous chapters. State 1 points to a group with more omnivore taste. The emphasis is on fun, they like thrillers and romantic movies and value the presence of a top actor. Members of that class show an aversion for "serious" content, being rather negative on drama and documentary and whilst important,

they are clearly less information seeking than the other groups. One could label the group as "feel good and not influenced". Opposite to that is state 4, with members expressing more affinity towards content features. For this group, higher values are observed for parameters such as based on, and they like drama and documentaries. Romance and sex is also more pronounced in comparison to the other classes, while at the same time they avoid action movies. Moreover, they are influenced by movies gaining an Oscar or being mentioned on pick lists. They will be referred to as the content group. Segment 3 can be labelled as action seeking. The action genre is highly valued as is the fun factor, but not as expressed in the comedy genre. Their preference for war and thriller is also slightly more pronounced, their appraisal for drama and documentary is outspoken negative and romance is valued low. They opt for sequels, but value external information somewhat less than the others classes. In that respect, as in their aversion for content, they act not unlike the feel good group. Segment 2 assembles mainly information seeking features such as sequel, list, Oscar and based on. In terms of genre, the information group is less outspoken, although, in reference to others, romantic comedy appears to dominate.

When comparing the estimated parameters to the analysis without dynamics (table 7.4), results are very similar in terms of estimation results and the same groups can be singled out (having a different class order). When comparing overall performance of the model (7.3 versus 7.5), the BIC value for the dynamic analysis is lower, dropping from 528.959 to 521.618. That isn't a big shift, but given that the parameters are clearly significant and that the dynamic model has more estimation parameters, it is fair to conclude that introducing dynamics does improve the fit of the model.

Focusing on the structural component of the LCMM, one can distinguish the initial state probabilities and the transition probabilities. The initial state is fairly equally distributed among groups, with the third state, the action state, being the largest with close to 30 percent. However, table 7.1 additionally shows that for most classes, the probability of switching is significant. The most stable state appears to be state 2 having a value of over 90 percent on the diagonal. This means that users in the information seeking segment in the period [1999 2002] will almost certainty be in the same segment the next period. Also the content seeking group is rather loyal to

Transition Probabilities				
Wald(0) = 169,8				
p-value = 5.3e-30				
	State[=0]			
	1	2	3	4
	0,2029	0,2547	0,2990	0,2434
	State			
State[-1]	1	2	3	4
1	0,2681	0,1426	0,2925	0,2968
2	0,0005	0,9073	0,0399	0,0522
3	0,0475	0,5296	0,3514	0,0715
4	0,0005	0,4413	0,0131	0,5450

Table 7.1: LCMM 4 CLASSES WITHOUT COVARIATE transition probabilities

their class, with about a 50-50 chance of staying or moving. When they move, they mainly go to the information seeking segment. The most unstable classes, looking at it from an intertemporal perspective, appear to be the feel good and action groups. The third group mainly moves to the second, the first has about an equal chance of reappearing in any class in the second time period. Where caution is due in making strong conclusions - supply of movies may be of different type over the different periods - it seems fair to say that information seekers and content groups express the highest levels of attachment towards certain preference classes.

In a last step, the rating of the previous period was added as a covariate in the transition probability, making that the state at time t becomes dependent both on the state at $t-1$ as well as on the rating. The estimated response model is almost identical to that without covariates (table 7.8 versus table 7.6). Also the general estimation statistics (table 7.7) show equal values in terms of overall significance, log-likelihood, classification and error values. The two dynamic models perform better than the static model in terms of the BIC value. Adding a covariate in the transition function does not contribute much to model fit, but provides a deeper insight in the potential dependency of switching behaviour on past assessment.

Transition Probabilities					
Wald(0)=152					
p-value=9.6e-13					
		State[=0]			
		1	2	3	4
		0,2053	0,2523	0,2984	0,2440
		State			
AVGRATE	State[-1]	1	2	3	4
2	1	0,9556	0,0150	0,0147	0,0147
2	2	0,2500	0,2500	0,2500	0,2500
2	3	0,0147	0,9559	0,0147	0,0147
2	4	0,0147	0,9559	0,0147	0,0147
3	1	0,2560	0,1605	0,2298	0,3537
3	2	0,0003	0,8600	0,0778	0,0619
3	3	0,0806	0,4343	0,3934	0,0917
3	4	0,0003	0,4544	0,0279	0,5175
4	1	0,2839	0,1145	0,4058	0,1958
4	2	0,0002	0,9389	0,0144	0,0465
4	3	0,0003	0,6308	0,3175	0,0514
4	4	0,0003	0,4221	0,0003	0,5773
5	1	0,2500	0,2500	0,2500	0,2500
5	2	0,0076	0,9773	0,0076	0,0076
5	3	0,2500	0,2500	0,2500	0,2500
5	4	0,0147	0,0147	0,0147	0,9559

Table 7.2: LCMM 4 CLASSES WITH COVARIATE transition probabilities

AVGRATE = the average rating over all movies in period t-1; the table shows the relationship between average rating values and the Markov transition probabilities between state t-1 and t

The most interesting finding relates to the composition of the transition probabilities at different rating levels displayed at table 7.2. The diagonal of each matrix at each level shows the group of stayers. This is the probability of being in a state and staying there. The basic hypothesis would be that the higher the average rating in the past period, the higher is the diagonal value. That appears to be the case for states 2 and 4, the information segment and content group, the first going from a 25

percent chance of staying at rating level 2, to 86 and 93 percent at rating 3 and 4, ending up at nearly 98 percent when the previous period average rating level was 5. It seems that information seekers are unlikely to stay when disappointed. Segment 4 goes from less than 2 percent over a 50-50 at ratings 3/4 to 95 percent when the past evaluation level was 5. The action group shifts states when the evaluations are bad and then settles at 25 to 30 percent chance of staying, diminishing slightly with the highest rating. The first group has the strangest pattern and does not alter when their average judgement of previous movies was low. After that, they stabilize at about 25 percent chance of staying compared to 75 percent of moving to another segment. Taken over the entire period and all rating levels, the information seeking segment appears to be most loyal to their segment, only moving away if the rating is really low. The feel good group is more difficult to pin down, strangely anchoring even when rating is low and exposing an even chance of moving to any class in a next period.

The considered period might not be sufficient to draw big conclusions regarding stability in taste. However, the estimated transition probabilities would suggest that individuals do switch segments, when looking at it from a longitudinal perspective. Basic economic theory designs demand as a function of quantities, largely assuming taste to be stable, at least over the short run. In cultural economics, stability of taste is mainly connected to genre stability. Views that taste are innate and stable, Peltoniemi (2015), Kivetz, Netzer & Schrift (2008) are confronted with those picturing taste as acquired through time, Blaug (2001), Throsby (2001). They are revealed by consumers getting an interest or becoming specialist in a particular style, Holbrook (1993), or adversely showing ample genre stability, Moon, Bergey & Iacobucci (2010). Current findings do reveal a degree of stability in taste; clear patterns can be discovered and for some segments, consumers are highly loyal. Other groups are characterized by higher switching behaviour. Moreover, switching behaviour seems to depend on previous evaluation, backing views on modelling experience goods as process of belief and belief update.

7.4 Interim Conclusions

This chapter added dynamics to the choice model of movies. The intertemporal aspect was introduced by letting the probability of being in a certain state at a particular time point vary and depend on the previous state. The values are reflected in the estimated transition probabilities. In terms of model fit, the Latent Markov Model performs better compared to its non-dynamic variant, but the improvement is marginal. However, the intertemporal parameters appear to be significant, so treating consumer choice as a dynamic decision process appears a better rendering of the underlying behaviour.

Introducing latent classes taught us that choice for movies can be segregated in a meaningful way. However, splitting the initial eight period time frame over different blocks contributed that the attachment to segments is time dependent. More importantly, individuals belonging to different classes show to be highly differentiated in their stay or switch behaviour. Those segments attaching high value to information features, remain doing so over the entire period. Also individuals looking for content appear consistent in their preference. The categories of action and fun seekers are more difficult to pin down, their behaviour is characterized by a high probability of transition to other segments.

The notion of experience goods installs dynamics. It is about living the experience. In a cognitive representation, the agents form expectations prior to consumption, evaluate their past, a judgement that is taken into account to form a new set of beliefs used in future decision making. If this theory is a valid reflection of the decision process of creative goods, one would expect the transition probabilities to be dependent on the level of rating; the likelihood of switching being higher if the rating value of the movies watched in previous time period is low and staying loyal to a consideration set if the appreciation in previous period appeared to be high. This was indeed confirmed for the information seeking and content valuing groups. Here also, the action and fun groups appear to be divergent, showing a more constant probability of switching, less dependent or sometimes even adversely related to the

rating level.

The findings of the last chapter are first experiments using Latent Markov Models on the relationship between movie tags and the user's decision making. In the light of identification, the number of classes and tags had to be reduced. Indeed, introducing a time variable augments the number of estimation variables substantially. Despite the more limited setting however, the big contours of the segments outlined in previous section were reaffirmed. Moreover, the results support the vision of creative products as multi-featured experience goods, not only in its meaning as "erlebnis", but even more so in its cognitive notion as "erfahrung", where the individual is seen as building up cultural capital. That, on its term challenges the assumption of stable and unchanging preferences underneath the majority of economic analysis and most recommendation systems. Certainly content based systems assume that like or dislike for certain features in the past are stable predictors for future behaviour. Information systems are often updated, and therefore deal with the problem in a technical way. Economic theory traditionally faced difficulties when integrating preference shifts in their modelling. More research is needed to confirm the presence of actual preference reversals. Data have to be taken over an even longer period of time. Online data of rating systems offer that potential. Looking at the problem in terms of consumers being loyal or moving away from a featured based segment offers a valuable framework to study preference evolution and the presence of taste shifts.

4-Class Latent Class Model					
Syntax Model					
Number of cases	470				
Number of replications	1040580				
Number of parameters (Npar)	63				
Random Seed	96291				
Best Start Seed	429302				
Chi-squared Statistics					
Degrees of freedom (df)	407	p-value			
L-squared (L)	522831,5203	8,5e-112816			
X-squared	1,664221132e+111	0,0e-2147483647			
Cressie-Read	1,929813078e+075	0,0e-2147483647			
BIC (based on L)	520327,3581				
AIC (based on L)	522017,5203				
AIC3 (based on L)	521610,5203				
CAIC (based on L)	519920,3581				
SABIC (based on L)	521619,0977				
Dissimilarity Index	1,0000				
Total BVR	0,0000				
Log-likelihood Statistics					
Log-likelihood (LL)	-264286,0959				
Log-prior	-1,8233				
Log-posterior	-264287,9192				
BIC (based on LL)	528959,8140				
AIC (based on LL)	528698,1918				
AIC3 (based on LL)	528761,1918				
CAIC (based on LL)	529022,8140				
SABIC (based on LL)	528759,8641				
Classification Statistics	Class				
Classification errors	0,0069				
Reduction of errors (Lambda)	0,9895				
Entropy R-squared	0,9855				
Standard R-squared	0,9850				
Classification log-likelihood	-264295,2450				
Entropy	9,1491				
CLC	528590,4899				
AWE	529554,7342				
ICL-BIC	528978,1121				
Class Classification Table	Modal				
Latent	1	2	3	4	Total
1	71,8346	0,2993	0,0000	0,1881	72,3220
2	0,1132	99,5896	0,5634	0,7650	101,0312
3	0,0000	0,0545	158,9052	0,6213	159,5810
4	0,0522	0,0566	0,5314	136,4256	137,0658
Total	72,0000	100,0000	160,0000	138,0000	470,0000
Prediction Statistics					
Choice					
Error Type	Baseline	Model	R		
Squared Error	0,0777	0,0705	0,0923		
Minus Log-likelihood	0,2906	0,2532	0,1287		
Absolute Error	0,1554	0,1403	0,0974		

Table 7.3: LCLR 4 CLASSES STATIC estimation statistics

Number of cases=number of individuals, number of replications=number of individuals times the number of movies, Log-Likelihood statistics where IC=Information Criterion, Classification statistics with assignment of cases to class with highest probability (modal), AWE=Approximate Weight of Evidence, Prediction statistics: error depends on difference between observed and predicted response, Baseline=average predicted response, R=Reduction of error

Model							
	term		coef	Wald(0)	p-value	Wald(=)	p-value
Choice	1	Class(1)	-2,3042	82922,1921	2,0e-18002	3258,5391	1,2e-706
Choice	1	Class(2)	-3,1184				
Choice	1	Class(3)	-4,1069				
Choice	1	Class(4)	-3,1418				
Choice	basedon	Class(1)	0,5876	4738,6989	2,4e-1026	39,2370	1,5e-8
Choice	basedon	Class(2)	0,6734				
Choice	basedon	Class(3)	0,7280				
Choice	basedon	Class(4)	0,5841				
Choice	oscar	Class(1)	0,4898	3576,3310	4,6e-774	151,4449	1,3e-32
Choice	oscar	Class(2)	0,7942				
Choice	oscar	Class(3)	0,9254				
Choice	oscar	Class(4)	0,6629				
Choice	war	Class(1)	0,3064	435,2980	6,5e-93	11,9427	0,0076
Choice	war	Class(2)	0,2664				
Choice	war	Class(3)	0,2913				
Choice	war	Class(4)	0,4036				
Choice	sequel	Class(1)	0,8789	3147,5615	5,2e-681	30,0189	1,4e-6
Choice	sequel	Class(2)	0,8566				
Choice	sequel	Class(3)	1,0889				
Choice	sequel	Class(4)	1,0365				
Choice	list	Class(1)	0,6093	5353,7786	7,4e-1160	113,1929	2,3e-24
Choice	list	Class(2)	0,8523				
Choice	list	Class(3)	0,9432				
Choice	list	Class(4)	0,7552				
Choice	funny	Class(1)	0,8385	4515,4787	6,8e-978	63,5798	1,0e-13
Choice	funny	Class(2)	0,6341				
Choice	funny	Class(3)	0,7123				
Choice	funny	Class(4)	0,8549				
Choice	comedy	Class(1)	0,1617	242,5795	2,6e-51	75,8346	2,4e-16
Choice	comedy	Class(2)	0,1448				
Choice	comedy	Class(3)	0,2201				
Choice	comedy	Class(4)	-0,0123				
Choice	seks	Class(1)	0,2293	739,9798	7,7e-159	80,3184	2,6e-17
Choice	seks	Class(2)	0,6104				
Choice	seks	Class(3)	0,5505				
Choice	seks	Class(4)	0,3318				
Choice	HaveActor	Class(1)	0,6922	2437,0444	7,7e-527	50,1128	7,6e-11
Choice	HaveActor	Class(2)	0,6068				
Choice	HaveActor	Class(3)	0,5584				
Choice	HaveActor	Class(4)	0,8120				
Choice	action	Class(1)	0,5421	2937,2786	2,2e-635	801,9510	1,6e-173
Choice	action	Class(2)	0,0466				
Choice	action	Class(3)	0,2816				
Choice	action	Class(4)	0,8151				
Choice	Thriller	Class(1)	0,4611	1383,9115	2,1e-298	188,4195	1,3e-40
Choice	Thriller	Class(2)	0,2405				
Choice	Thriller	Class(3)	0,0911				
Choice	Thriller	Class(4)	0,4409				
Choice	drama	Class(1)	-0,1501	1227,9436	1,4e-264	1184,6089	1,6e-256
Choice	drama	Class(2)	0,5333				
Choice	drama	Class(3)	0,1936				
Choice	drama	Class(4)	-0,2466				
Choice	romance	Class(1)	0,2724	704,4384	3,8e-151	94,9366	1,9e-20
Choice	romance	Class(2)	0,3319				
Choice	romance	Class(3)	0,2461				
Choice	romance	Class(4)	0,0790				
Choice	documentary	Class(1)	-0,5652	437,5636	2,1e-93	434,6385	6,9e-94
Choice	documentary	Class(2)	0,4110				
Choice	documentary	Class(3)	0,3239				
Choice	documentary	Class(4)	-0,4717				

Table 7.4: LCLR 4 CLASSES STATIC parameter estimates

4-Classes LCMM model	No covariate				
Syntax Model					
Number of cases	470				
Number of time points	940				
Number of replications	1040580				
Number of parameters (Npar)	75				
Random Seed	519903				
Best Start Seed	1543779				
Chi-squared Statistics					
Degrees of freedom (df)	395	p-value			
L-squared (L)	516126,3007	1,4e-111379			
X-squared	1,657145280e+111	0,0e-2147483647			
Cressie-Read	1,922235814e+075	0,0e-2147483647			
BIC (based on L)	513695,9713				
AIC (based on L)	515336,3007				
AIC3 (based on L)	514941,3007				
CAIC (based on L)	513300,9713				
SABIC (based on L)	514949,6253				
Dissimilarity Index	1,0000				
Total BVR	87,9845				
Log-likelihood Statistics					
Log-likelihood (LL)	-260578,3759				
Log-prior	-4,3220				
Log-posterior	-260582,6979				
BIC (based on LL)	521618,2068				
AIC (based on LL)	521306,7518				
AIC3 (based on LL)	521381,7518				
CAIC (based on LL)	521693,2068				
SABIC (based on LL)	521380,1712				
Classification Statistics	State				
Classification errors	0,0177				
Reduction of errors (Lambda)	0,9710				
Entropy R-squared	0,9667				
Standard R-squared	0,9637				
Classification log-likelihood	-260619,7040				
Entropy	41,3281				
AWE	522387,3179				
ICL-BIC	521700,8630				
State Classification Table	Modal				
Latent	1	2	3	4	Total
1	125,8065	0,0000	0,5886	1,0986	127,4937
2	0,0000	363,6900	2,1724	1,2610	367,1235
3	1,4181	2,0138	217,4276	3,1986	224,0581
4	0,7755	1,2962	2,8113	216,4418	221,3248
Total	128,0000	367,0000	223,0000	222,0000	940,0000
Prediction Statistics					
choice					
Error Type	Baseline	Model	R		
Squared Error	0,0777	0,0697	0,1036		
Minus Log-likelihood	0,2906	0,2490	0,1431		
Absolute Error	0,1554	0,1386	0,1084		

Table 7.5: LCMM 4 CLASSES WITHOUT COVARIATE estimation statistics

Number of cases=number of individuals, number of replications=number of individuals times the number of movies, Log-Likelihood statistics where IC=Information Criterion, Classification statistics with assignment of cases to class with highest probability (modal), AWE=Approximate Weight of Evidence, Prediction statistics: error depends on difference between observed and predicted response, Baseline=average predicted response, R=Reduction of error

Model							
	term		coef	Wald(0)	p-value	Wald(=)	p-value
Choice	1	State(1)	-2,2175	59948,7597	5,9e-13014	4102,9802	5,7e-890
Choice	1	State(2)	-4,2529				
Choice	1	State(3)	-3,0156				
Choice	1	State(4)	-3,3330				
Choice	basedon	State(1)	0,5991	5289,5807	6,4e-1146	53,2058	1,7e-11
Choice	basedon	State(2)	0,7827				
Choice	basedon	State(3)	0,6385				
Choice	basedon	State(4)	0,7300				
Choice	oscar	State(1)	0,5831	3576,9882	3,3e-774	68,9253	7,3e-15
Choice	oscar	State(2)	0,8154				
Choice	oscar	State(3)	0,6601				
Choice	oscar	State(4)	0,8549				
Choice	war	State(1)	0,2612	430,5079	7,1e-92	10,6502	0,014
Choice	war	State(2)	0,3695				
Choice	war	State(3)	0,3787				
Choice	war	State(4)	0,2752				
Choice	sequel	State(1)	0,8838	3175,1855	5,2e-687	31,3699	7,1e-7
Choice	sequel	State(2)	1,1444				
Choice	sequel	State(3)	1,0190				
Choice	sequel	State(4)	0,9068				
Choice	list	State(1)	0,7148	5891,3167	1,5e-1276	84,0767	4,1e-18
Choice	list	State(2)	0,9613				
Choice	list	State(3)	0,7570				
Choice	list	State(4)	0,9352				
Choice	funny	State(1)	0,8411	4710,6483	2,9e-1020	50,8059	5,4e-11
Choice	funny	State(2)	0,7297				
Choice	funny	State(3)	0,9075				
Choice	funny	State(4)	0,6971				
Choice	comedy	State(1)	0,1441	233,8058	2,0e-49	29,6355	1,6e-6
Choice	comedy	State(2)	0,1770				
Choice	comedy	State(3)	0,0449				
Choice	comedy	State(4)	0,1786				
Choice	seks	State(1)	0,3296	834,3141	2,8e-179	79,0278	5,0e-17
Choice	seks	State(2)	0,5407				
Choice	seks	State(3)	0,2839				
Choice	seks	State(4)	0,6678				
Choice	HaveActor	State(1)	0,7026	2571,7027	4,7e-556	43,0895	2,4e-9
Choice	HaveActor	State(2)	0,6380				
Choice	HaveActor	State(3)	0,8550				
Choice	HaveActor	State(4)	0,6233				
Choice	action	State(1)	0,5167	3074,1880	4,3e-665	739,1478	6,8e-160
Choice	action	State(2)	0,4528				
Choice	action	State(3)	0,8874				
Choice	action	State(4)	0,0933				
Choice	Thriller	State(1)	0,4848	1508,5143	2,0e-325	207,5882	9,6e-45
Choice	Thriller	State(2)	0,1259				
Choice	Thriller	State(3)	0,4771				
Choice	Thriller	State(4)	0,2488				
Choice	drama	State(1)	-0,0216	1257,1851	6,4e-271	1176,0745	1,1e-254
Choice	drama	State(2)	0,1363				
Choice	drama	State(3)	-0,2865				
Choice	drama	State(4)	0,5686				
Choice	romance	State(1)	0,2778	795,5795	7,0e-171	80,0360	3,0e-17
Choice	romance	State(2)	0,2737				
Choice	romance	State(3)	0,0959				
Choice	romance	State(4)	0,3544				
Choice	documentary	State(1)	-0,5036	620,5561	5,5e-133	613,2314	1,4e-132
Choice	documentary	State(2)	0,3906				
Choice	documentary	State(3)	-0,6210				
Choice	documentary	State(4)	0,5741				

Table 7.6: LCMM 4 CLASSES WITHOUT COVARIATE parameter estimates

4-Classes Latent Class Markov Model	rating is covariate				
Syntax Model					
Number of cases	470				
Number of time points	940				
Number of replications	1040580				
Number of parameters (Npar)	111				
Random Seed	143081				
Best Start Seed	2566797				
Chi-squared Statistics					
Degrees of freedom (df)	359	p-value			
L-squared (L)	516100,7962	1,6e-111430			
X-squared	1,657214180e+111	0,0e-2147483647			
Cressie-Read	1,922304062e+075	0,0e-2147483647			
BIC (based on L)	513891,9652				
AIC (based on L)	515382,7962				
AIC3 (based on L)	515023,7962				
CAIC (based on L)	513532,9652				
SABIC (based on L)	515031,3621				
Dissimilarity Index	1,0000				
Total BVR	88,7384				
Log-likelihood Statistics					
Log-likelihood (LL)	-260565,6237				
Log-prior	-4,5178				
Log-posterior	-260570,1415				
BIC (based on LL)	521814,2006				
AIC (based on LL)	521353,2473				
AIC3 (based on LL)	521464,2473				
CAIC (based on LL)	521925,2006				
SABIC (based on LL)	521461,9080				
Classification Statistics	State				
Classification errors	0,0166				
Reduction of errors (Lambda)	0,9730				
Entropy R-squared	0,9673				
Standard R-squared	0,9647				
Classification log-likelihood	-260606,2644				
Entropy	40,6407				
AWE	522911,4354				
ICL-BIC	521895,4821				
State Classification Table	Modal				
Latent	1	2	3	4	Total
1	126,3221	0,0000	0,9778	1,5110	128,8108
2	0,0000	361,9398	1,7615	1,3244	365,0257
3	0,2540	2,7893	218,8038	2,8218	224,6689
4	0,4239	1,2709	2,4570	217,3427	221,4946
Total	127,0000	366,0000	224,0000	223,0000	940,0000
Prediction Statistics					
choice					
Error Type	Baseline	Model	R		
Squared Error	0,0777	0,0697	0,1036		
Minus Log-likelihood	0,2906	0,2490	0,1431		
Absolute Error	0,1554	0,1386	0,1084		

Table 7.7: LCMM 4 CLASSES WITH COVARIATE estimation statistics

Number of cases=number of individuals, number of replications=number of individuals times the number of movies, Log-Likelihood statistics where IC=Information Criterion, Classification statistics with assignment of cases to class with highest probability (modal), AWE=Approximate Weight of Evidence, Prediction statistics: error depends on difference between observed and predicted response, Baseline=average predicted response, R=Reduction of error

Model							
	term		coef	Wald(0)	p-value	Wald(=)	p-value
Choice	1	State(1)	-2,2197	64368,0082	1,5e-13973	3999,5818	1,6e-867
Choice	1	State(2)	-4,2595				
Choice	1	State(3)	-3,0284				
Choice	1	State(4)	-3,3277				
Choice	basedon	State(1)	0,6002	5316,6020	8,7e-1152	53,9549	1,1e-11
Choice	basedon	State(2)	0,7853				
Choice	basedon	State(3)	0,6375				
Choice	basedon	State(4)	0,7292				
Choice	oscar	State(1)	0,5807	3578,8686	1,3e-774	70,9533	2,7e-15
Choice	oscar	State(2)	0,8151				
Choice	oscar	State(3)	0,6621				
Choice	oscar	State(4)	0,8560				
Choice	war	State(1)	0,2592	432,5097	2,6e-92	11,5851	0,0089
Choice	war	State(2)	0,3671				
Choice	war	State(3)	0,3854				
Choice	war	State(4)	0,2744				
Choice	sequel	State(1)	0,8893	3178,0257	1,3e-687	30,4147	1,1e-6
Choice	sequel	State(2)	1,1465				
Choice	sequel	State(3)	1,0124				
Choice	sequel	State(4)	0,9117				
Choice	list	State(1)	0,7144	5916,2533	5,9e-1282	85,1601	2,4e-18
Choice	list	State(2)	0,9640				
Choice	list	State(3)	0,7558				
Choice	list	State(4)	0,9357				
Choice	funny	State(1)	0,8408	4706,9235	1,9e-1019	48,5893	1,6e-10
Choice	funny	State(2)	0,7299				
Choice	funny	State(3)	0,9049				
Choice	funny	State(4)	0,7002				
Choice	comedy	State(1)	0,1440	233,2389	2,7e-49	28,6214	2,7e-6
Choice	comedy	State(2)	0,1794				
Choice	comedy	State(3)	0,0467				
Choice	comedy	State(4)	0,1751				
Choice	seks	State(1)	0,3253	839,0096	2,7e-180	82,0798	1,1e-17
Choice	seks	State(2)	0,5437				
Choice	seks	State(3)	0,2821				
Choice	seks	State(4)	0,6696				
Choice	HaveActor	State(1)	0,7054	2571,9747	4,1e-556	43,1885	2,2e-9
Choice	HaveActor	State(2)	0,6369				
Choice	HaveActor	State(3)	0,8547				
Choice	HaveActor	State(4)	0,6224				
Choice	action	State(1)	0,5189	3192,8705	7,6e-691	781,8908	3,7e-169
Choice	action	State(2)	0,4479				
Choice	action	State(3)	0,8918				
Choice	action	State(4)	0,0926				
Choice	Thriller	State(1)	0,4848	1499,5519	1,8e-323	207,2694	1,1e-44
Choice	Thriller	State(2)	0,1240				
Choice	Thriller	State(3)	0,4757				
Choice	Thriller	State(4)	0,2503				
Choice	drama	State(1)	-0,0260	1263,1925	3,2e-272	1179,3942	2,2e-255
Choice	drama	State(2)	0,1395				
Choice	drama	State(3)	-0,2863				
Choice	drama	State(4)	0,5685				
Choice	romance	State(1)	0,2782	796,6884	4,0e-171	84,0025	4,2e-18
Choice	romance	State(2)	0,2754				
Choice	romance	State(3)	0,0953				
Choice	romance	State(4)	0,3523				
Choice	documentary	State(1)	-0,5128	621,3032	3,8e-133	613,0518	1,5e-132
Choice	documentary	State(2)	0,3994				
Choice	documentary	State(3)	-0,6119				
Choice	documentary	State(4)	0,5705				

Table 7.8: LCMM 4 CLASSES WITH COVARIATE parameter estimates

Overall Conclusions

This thesis comprises a multidisciplinary quest to improve economic choice models for creative goods, with movies used as a particular case. The motivation arose from the observation that empirical cultural economics lacks a fundamental theoretical underpinning. Indeed, the concept of creative goods challenges economic theory relying as it does on basic premisses of a representative consumer and stable preferences. Models of rational addiction were addressed to deal with the capital formation nature of creative products, paradigms of Bayesian learning to shape the process of belief formation and revision. While valuable, the theories are hard to translate into a verifiable hypothesis. The reason is that the theoretical frameworks by-pass what makes a creative good, namely the novelty aspect. Each product is a new creation, an alternative composition of features that a consumer weights against alternatives. In the light of this, it is an open question as to how to test for persistency in taste when individuals are confronted with new choice circumstances each time a decision has to be made. Stability in taste has been partly tested in creative goods studies in terms of genre loyalty, but genre is but one of the characteristics entering a consumer's decision function. Economic theories by Nobel Prize winners such as Daniel McFadden and research pursued by Amos Tversky point to the restrictions of economic theory in dealing with large scale empirical micro economics. Random utility theory opens up perspectives to the inclusion of taste heterogeneity. Tversky's path breaking work presents an overall challenge to basic economic thinking, bridging insight from psychology with economics and in so doing broadens views on choice modelling in terms of comparison of features. Related to that are ideas of categorization and prototyping. It is a way out when it comes to reflection on novelty goods to think of heterogeneity of taste in terms of latent segments

representing typical preference profiles shaped by a variety of features.

While economic theory struggles to frame artistic goods, a branch of computer sciences, specialized in product recommendation has expanded quickly, and developed various types of algorithms to predict a consumer's potential interest in items. Their preferential topics were products offered through online sale, hence mainly creative products such as literature, music and movies. This has been particularly true in the case of movies, where research joint ventures such as GroupLens have developed specialist applications based upon recommender algorithms. Recommender theory research grew to become an independent subject field in the academic curriculum and a scientific community formed around it. When reviewing its contributions, there are two strands that can be distinguished. Content based theories incorporate taste persistency arising from the relationship between object and feature and recommend items similar to those opted for in the past: similarity expressed in terms of feature equality which can be indexed by a distance measure. It is important to notice how, for this methodology, past observation translates into future prediction, not through the item, but through the characteristics. That way, heterogeneity is absolute: it comes from singular profiling, person to object to feature, hence person to feature. A second approach, namely collaborative filtering, is intrinsically social, and attaches a group of peers to an individual based on similar interests. Predictions follow from observing consumption patterns of the like minded. Statistical techniques are different from those used in cultural economics, relying on data mining techniques such as clustering or Bayesian classification. Here taste patterns are reduced to representative groups of individuals expressing similar behaviour. Hybrid models incorporate both by augmenting the cluster base with object features. The probabilistic collaborative filtering model suggested by Thomas Hofmann can be seen as an outsider. It is an offspring of latent semantic indexing approaches, algorithms created to classify texts and can be seen as the probabilistic counterpart of the SVD-methods of Koren and Bell, which made them win the Netflix prize for the best predictive system. Persons are typified by their "typical preference pattern". Unlike the classical collaborative filtering, individuals hold an uncertain relation to the latent classes. When taken in its hybrid form, formally modelled as a latent

class logistic regression, the referential patterns are shaped by their representative features. These are not subsets of features each related to a class, but rather all features anchored to all segments with various probabilities.

What are the features driving decisions to opt for particular creative products? Recommender systems mechanically detach and incorporate objective, technical information such as actor, director or genre, but it can be assumed that prior judgement over products involves a broader range of elements. Through various fora, people formulate opinions on creative goods, describing the elements they value, judging past experience by attaching ratings to them. It gave rise to a subset of recommender systems, using algorithms similar to collaborative filtering or content based methods, but incorporating social information shared by persons through their online communication. Tags in particular are interesting to detect the user's connotation attached to an object. They are freely added keywords, when taken in aggregate the joint contributions grow into shared vocabularies or so called folksonomies. Tags are linguistic expressions and therefore diverse, but are direct communications of relational concepts. While subjective, partial and hard to homogenize, they appear to be among the few ways to detect some of the main variables in the consumer's decision function. Collected online tagging information is one element borrowed from computer science, along with the bottom-up approach to discover patterns by looking at the data, those patterns being latent segment or typical preference patterns. The big data underneath this work is provided by MovieLens, information extracted by the GroupLens research team from their eponymous recommender system and made available for research purposes. The 10M set also includes tag data, added to the traditional user-movie-rating information. A number of restrictions was imposed on the dataset making it a big, yet PC manageable sample set.

A frequency count of the most common tags provides some first insights into their use. Traditional genre annotation, such as action, drama, romance are popular by users, which was to be expected as this type of labelling is widely provided by production houses as well as movie websites. Individual actors or directors are less prominent in the top ranked tag list, with the exception of Quentin Tarantino. One of the more notable observations is that content or storyline seems to matter

substantially to movie watchers. A tag such as "based on" is listed as top word as are related terms like plot, ending or twist. "Based on" can refer to a book, comic, game or life story. The content factor is depthened by users widely referring to terms as politic or historic. It is a dimension often discarded in scientific analysis and, not unlike sequels, it indicates users avoiding uncertainty by opting for new items that have a comfortable degree of similarity to knowledge they posses. Other factors of quality certifications, often the basic estimation variables in cultural economics studies, are also manifestly specified. It includes terms such as Oscar or appearing on a recommendation list published by film sites. When looking at the judgement terms, tagged adjectives, a main part contains trivial qualifiers such as good, bad, excellent, but part of the expressions clearly indicate that the experience exceeded or felt short compared to prior expectations, pointing to the presence of belief formation. While providing some elementary insights, frequencies are as such not a good guide for tag relevance. Indeed, some tags may be important to a smaller segment of users or adversely some power users potentially influence the global vocabulary. Therefore, patterns of co-occurence are searched for in the vocabularies of users, where the tags labelling segments are withheld as the most relevant terms. The method used is Latent Dirichlet Allocation (LDA), a Bayesian hierarchical model to search for latent linguistic layers or topics in a corpus of texts.

LDA is a topic model approach, introduced by Blei et al. (2003). It presents a methodology to look for hidden abstract patterns in the semantic space. While initially used for document classification, here, it is translated into individuals each connected to a bag of words consisting of tags or movie titles. Distinguishing is that terms shape classes in a probabilistic way with individuals not deterministically attached to one segment. It provides estimates of the distribution of tags over classes and of persons' assignment to classes. The choice for topic models is not merely a methodological one. Griffiths et al. (2007), establish the concordance between the way concepts are paired together in the semantic space and the notion of similarity put forward by Tversky (1977), not in geometrical sense, but as a function of common and uncommon characteristics. Estimating the distributions shows that it makes sense to think of the underlying decision making dimensions in terms of latent classes,

indirectly inferred by looking at the multitude of tags. A number of topics can be clearly labelled by the high probability terms. Some point to the more classical genre and actor dimensions, others enhance genre into a style. The exercise also manifests that information and content seekers are particular subgroups of users, not driven by the same motives as others. While the strength of the latent class approach is clearly emphasized, the exercise was meant as a first exploration to discover patterns given a multitude of tags, a lot of them having a very low frequency of occurrence. One of the aims was to select the most relevant tags that can serve as explanatory variables, driving choice behaviour for movies. Relevance surpasses frequency, but relates to concepts or motives being important also to certain subgroups or niches.

Individual choice is fundamental in the study of economics. What is of interest is the question whether agents will decide to opt for a movie or not and what are the driving forces that steer the decision making process. Chapter 6, by shedding light on that question, can for that reason be considered as one of the core parts of the thesis. The estimation model put forward is a Bayesian Latent Class Regression Model. The <UserID, MovieID, Rating> triplet from the MovieLens dataset was converted into a decision model of 470 users opting for a set of 2.214 movies. The choice of a movie is approximated by the user having rated the movie. While certainly incomplete to capture the entire consumption history of a person, it can still be considered as one the most elaborate survey methods to track longitudinal film watching behaviour. The choice of LCLR is motivated by the fact that logistic regression follows naturally from Random Utility Theory and by allowing an easy integration of Bayesian latent class concepts, paves the way to explore typical preference patters. Conditional on their presence, heterogeity in consumption is introduced. From the side of recommender systems, the probabilistic latent class models by Hofmann & Puzicha (1999) and by Kagie et al. (2009) were singled out as promising approaches. Not only because of their methodological strength when used as a way to reduce dimensionality, but because like LDA, the features are probabilistically attached to different latent classes. It therefore captures the uncertain relationship of a person versus the product features. One could state that in the latent class logistic regression model, the world of big data analytic tools and the discrete choice modelling of economic

theory, such as put forward by McFadden (1980), reach each other. While forming an inspiration, the presented model is distinct from the probabilistic recommender approaches. There, the dependent and thus prediction variable is rating, here it is choice. This is an important distinction, as it embodies a view, sketched at the beginning of the thesis, that the decision making process is fundamentally dynamic, where users act on beliefs, make a choice, and then undergo the experience which is then evaluated. Choice and rating are seen as separated by time since evaluation comes in at the moment of belief revision. Moreover, whether or not a movie is watched is the prime concern of the industry. It makes the dependent variable to be binary, not a 5 scale categorical, as is often the case in recommender systems. The objective of recommender research is often reduced to maximizing prediction accuracy, largely ignoring the meaning of the segments that are generated by the application of their techniques, which offer valuable information to target consumers. On the other hand, the recommender approach is uncensored large scale, benefiting from all the advantages of big data. Here, the sampled data are large compared to survey data used in mainstream movie economics literature, yet small compared to those of real world applications.

The tag based estimated latent class regression model for movies provided some interesting results. While obviously sparse, the tag features appeared to serve as good explanatory variables. The mean absolute errors for the suggested models were smaller compared to those generally observed in the recommender literature. Looking at tags as signalling the underlying motives behind movie choice appears as such promising. The model outperforms a benchmark using only genre, although the genre model didn't perform badly, certainly when considering that the information is easy to capture. Anchoring the tags to the movie objects is a complex exercise, tracing back a movie that has once been given a tag in its multi-linguistic appearance. However, to get insight into the consumers' motives, the tag based approach outlined a number of clear segments corresponding to decision dimensions of varying kinds. Globally, the detected segments are a mixture of the main decision factors put forward in economic theory and recommender literature. Some preference classes are not remote from the genre classification, such as action or fun

seeking groups, however add more specific terms that can be used for precise profiling. The information variables, while overall important, are the true driving force in the decision for a number of subgroups. Indeed, for certain segments, quality certifiers, such as Oscars or the film being mentioned on a list have a significant positive impact on their choice behaviour. It agrees with the variables singled out in the movie economics literature review, reflecting a vision of experience goods inducing prior uncertainty. For another group, the information elements are important, when joined with content elements, such as political or documentary. The content dimension, also put forward in the LDA analysis, appears to be overall significant, while largely ignored in empirical movie studies. It is clear that some of the keywords are not merely projecting elements of connotation to an object, but unite individuals valuing those features to a lesser or larger extent. Here, the method used shows its true merits. Features are not important when standing in isolation, but in their co-occurrence. A consumer class can share with others their appreciation for action movies, but at the same time being very different in their liking for engaged content. The method allows this fine grained approach. Some individuals do indicate niche preference while others reveal more omnivore taste. Sometimes characteristics are positively appreciated by all, but more pronounced by some. Other features, such as documentary, gay, nudity are more divisive, receiving positive values in a number of preference profiles, while judged negative in another. The findings explain to a certain extent why empirical econometric models, based on the representative consumer hypothesis, remain inconclusive on a number of variables. Elements being highly influential to some matter less to others, and dependent on your sample, this will result in different statistical findings. The observations in this thesis manifestly support the view of consumer preferences being intrinsically heterogeneous. Adding latent classes clearly improved model fit as indicated by the BIC criterion. The largest gain was extracted when going from one segment to a limited number of subgroups. However, the larger the predefined number of classes, the more that niche categories turn visible.

Overall, Bayesian Latent Class models exhibit clear advantages when studying choice patterns for creative goods. The latent classes introduce heterogeneity in the choice

profiles of consumers. It takes the shape of "typical preference profiles", meaning that those attached to them act in a similar way when confronted with the same feature set. It sets diversity of taste in between that of content based systems, in which one individual agrees with one profile and economic theory, where one profile represents the behaviour of all. The significance of the latent class parameters therefore directly challenges preference homogeneity. Important also is that features are probabilistically attached to all classes, indicating a higher or lower, positive or negative impact on the chance to opt for a movie. It supports the view that creative goods are a bundling of characteristics, triggered and compared by consumers in their decision making process. The Bayesian approach is a necessary one to express the uncertainty of the object-feature relationship as well as that of the user versus its class. Typical profiles can as such also be interpreted as consideration sets, abstract prototypical classes in which agents seek resemblance when confronted with a new product. When dealing with novelty goods, and persistency of taste herein, one can only fall back on similarity, and more specifically on feature similarity, through which intertemporal choice stability takes place.

While the notion of class comparison encompasses that of persistency, the LDA and LCLR are static methods, assuming stability of taste over the investigated period. However, adhering the vision of movies as an experience good suggests dynamics. The time dimension is innate to the process of experiencing a creative product, bearing on prior beliefs that are sequentially revised. The learning aspect induces cultural capital being formed, making that preference profiles alter. Recommender literature largely by passes or transcends the issue of shifting taste. Only the SVD++ model of Koren & Bell (2011) recognizes the temporal effect, which they address by including time drifting parameters. The regular updates of real life recommender systems softens the problem on the side of computer science. Economic theory traditionally faces paradigmatic difficulties when confronted with varying taste patterns. They are integrated through human capital formation or addiction models. Both proved valuable and at the same time insufficient to represent the nature of the dynamic patterns of cultural consumption. Capitalizing on the vision that intertemporal persistency in taste can be captured through feature loyalty, a extra dimension was

added to latent class modelling. Introducing a Markov structure, through its transition probabilities, the likelihood of being loyal to a class or displaying switching behaviour can be estimated. The previous chapter offers a trial of this approach, dividing the considered period in two distinct time spans and limiting the number of tags and classes to guarantee identification. While being experimental, the results demonstrate the potential of this line of inquiry for future research. Estimation results indicate both stability and instability when looking at the loyalty of users to classes, the nature of the behaviour being highly dependent of the class one belong to. Consumer groups attaching high value to information elements continue to do so in the next period. Also the content group displays a high level of persistency in their behaviour. The fun and action seeking movie watchers show the highest probability to switch to another segment. Adding rating into the analysis completes this work. When supporting the hypothesis that individuals change beliefs, one would expect the transition probabilities to be dependent on movie evaluation observed during the previous period. This is indeed the case, the impact of rating is significant, but not always monotonous. For the information and content seeking group, the probability of staying increases with the value of past rating, going to 95 percent when the average rating level in previous period was 5. Here also, the other classes display a significant but not necessarily homogeneous relationship between staying and the previous evaluation levels. More research will be needed to certify the strength of this approach.

Adding the last element completes a line of research aimed at investigating what type of model is best suited to represent choice behaviour for movies. It takes from economics the ideas that creative products are experience goods that can be conceived of as bundles of comparable characteristics. Those features can be compared, not in terms of geometric distance, but as a function of characteristics in common. From computer sciences, the work adopts a bottom-up methodology so as to read patterns in the data, open access data sets provided by the GroupLens research team, and the conviction that the rich information sets coming about in users' interaction with the internet provides valuable insight into the motives of film consumers. Probabilistic latent class models bridge both worlds, allowing us to infer the unobserved

consumer heterogeneity from a limited number of observed variables. An additional Markov structure, installing dynamics, provides a full underpinning to the notion of experience goods as an intrinsically intertemporal concept representing a process of prior and ex post judgement. In this, the novelty aspect is captured through feature similarity; heterogeneity by introducing typical preference profiles and dynamics by means of loyalty or disloyalty to them. That way, the Bayesian latent class approach offers a valuable framework to deal with a category of products that proved difficult to pin-down in traditional econometric models.

Research questions in retrospect

1. Conceptualisation of creative goods

The definition of experience goods ought to be comprehensive, encompassing the notion erlebnes, referring to immediate joy, and the intertemporal aspects of belief formation captured by the word erfahrung. Unlike the positions stated in the influential paper by Holbrook and Hirschman, the divide between the cognitive and emotional is abandoned. Instead, the concept of contiguity is given a central position, pointing to the frequency with which events, signals or symbolic components are paired together to become evocative at a later stage. The essence of creative products is the novelty aspect, the continuous composition and recomposition of decisive features. They are assessed against prototypes or representative classes. Each decision situation demands a judgement over a new set of connected features that are not necessarily compared in a quantitative way, but in terms of presence or absence. Here, the Lancasterian approach based on the utility concept is confronted with the contrast model of similarity or measures of proximity.

2. Transfer of ideas from computer science to economics

Recommender systems research is a branch specialised in prediction algorithms for online products, mainly creative goods. Research on movie recommendations takes a very prominent position and ought to be considered by economics researchers studying the topic. Their approach is bottup-up, starting from big data sets, built up through the online interaction by the users. The data quality is lower in reference to the standards of econometrics, but nevertheless, the datasets are a valuable source for empirical cultural economics currently relying on aggregated data. The recom-

mender theory approach is performance based, largely discarding the premisses that are the core of economic consumer theory. However, their methods embody the ideas of contiguity. In content based systems, past consumption translates into future prediction through feature similarity. Collaborative filtering approaches work with typical preference profiles or peer groups where similarity is based on co-occurrence of items revealed in past consumption patterns. Heterogeneity in preference, a critical point in micro econometrics, is absolute in content based systems and a hidden premise in collaborative algorithms. Distance measures dominate the comparison of features or item collections. They are reflected in the machine learning techniques, ruling research on the topic and gradually gaining popularity in consumer studies.

3. Using online social information to discover relevant movie features

Tags are user generated keywords or small phrases annotating an object. Freely added by individuals to categorize, they seem at first glance a promising candidate to get insight into the main features an individual relates to a creative product. At the same time, there are a lot of counterarguments against their use. Being free expressions, they cover linguistically a broad spectrum and tag quality is blurred by the presence of synonyms, homonyms and spelling differences. However, collective tagging behaviour is shown not to follow a chaotic path but rather to converge to a stable equilibrium. More importantly, user added tags appear to be one of the few information sources to gain insight into the critical dimensions that are at stake in consumers' decision making for movies. Examination of a large set of tags added by MovieLens users, inform us that both the information variables, at the forefront in economic theory, as well as the more technical features prominent in recommender systems, play an important role. Some dimensions are underexposed in both research areas, such as the content variable. Also genre remains a prominent classifier, be it in conjunction with other features. The study of tags strengthens the view of creative products as multi-featured objects where features can be of both subjective and technical nature.

4. Segmentations of consumer choice patterns based on tags

Based on Movielens data of rating and tagging and using latent class techniques, consumers were divided by their typical preference pattern. Tags were attached to movies as sole features. The generated consumer segments based on tag valuation give rise to a number of meaningful classes. Notable is that the information features, dominating cultural economics literature, the technical features of content based recommender systems and the underexposed script variables all translate into typical patterns. Segmentation based on tags clearly demonstrates how certain features, albeit important to all, can be deterministic in the choice decision for some, secondary to others. Some tags work in a unifying way, others cause a clear split of preferences. The analysis shows the limitation of the genre variable. Genre is but one decision variable, while the boundaries of classifications are clearly broader and changing over time. Segmenting consumers based on tags outperforms that based on genre, not only in terms of predictive statistics but even more so in term of labelling of segments. Using tag based latent classes, heterogeneity is introduced in a meaningful way, contributing to micro econometrics, where adhering the representative consumer hypothesis leads to ambiguous estimation results. Apart from the scientific value added, segmenting consumers based on tags offers a useful methodology to the movie sector, in service of targeted communication and strategies of price discrimination. Results are not the result of stated preferences, but come from information available to the business.

5. The value added of Bayesian latent class approach

The Bayesian paradigm is well suited for the study of experience goods. The uncertain nature of consumers towards a good as well as the varying attitudes of individuals towards their "typical profile" can be fully expressed in a mixture model. It allows the individuals' distributions to be inferred as posteriors based on the observation of their choice patterns. Segments are probabilistically related to the tags, a relationship which can be estimated, and studied over time, opening perspectives for further research on preference shifts or reversals. Prediction follows naturally from the user's posterior attachment to a class and the features characterising the

object. The probabilistic latent class approach proved its value at three levels. First as exploration devise: topic models allowed to detect patterns in large sets of unstructured linguistic data. Secondly, latent class logistic regression allowed to test heterogeneity in consumer patterns, in a way that is consistent with random utility models, where at the same time tag similarity is not based on distance measures but on feature similarity in line with the contrast model. Finally, latent Markov models allow combining the studying of heterogeneity with dynamics by estimating the probability of consumers changing segments.

6. The paradox of novelty goods and taste dynamics

It is an open question if anything meaningful can be said about preference dynamics for creative goods if each good is new. A creative product can be seen as unrelated to a previous one, making each decision situation unique, thus making the concept of persistency empty. Yet, there is a firm conviction that taste is somehow stable. The paradox can be lifted when thinking of creative goods in term of composites of features that are assessed at each period in reference to a dominating personal category. Through latent class techniques, those categories can be inferred. Stability in taste or taste shifts can be studied as loyalty to the segments or departure from them. It borrows from the practice of content based recommender systems that predicting the future ratings from the past occurs through similarity in features. Being a hidden supposition in recommender theory, through a dynamic latent class approach and in combination with online consumer data, it is argued in this thesis that the nature of dynamic taste patterns can be studied in a meaningful way.

References

Abercrombie, N. (1996), *Television and Society*, Polity Press Book, Polity Press, Cambridge.

Adomavicius, G. & Tuzhilin, A. (2005), 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', *IEEE Trans. on Knowl. and Data Eng.* **17**(6), 734–749.

Adomavicius, G. & Tuzhilin, A. (2011), Context-aware recommender systems, *in* 'Recommender systems handbook', Springer US, Boston, MA, pp. 217–253.

Agarwal, D. & Chen, B.-C. (2010), flda: Matrix factorization through Latent Dirichlet Allocation, *in* 'Proceedings of the Third ACM International Conference on Web Search and Data Mining', WSDM '10, ACM, New York, NY, USA, pp. 91–100.

Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle, *in* 'Proceedings of the 2nd International Symposium on Information Theory', Akademiai Kiado, Budapest, pp. 267–281.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Aldous, D. J. (1985), Exchangeability and related topics, *in* P. L. Hennequin, ed., 'École d'Été de Probabilités de Saint-Flour XIII — 1983', Springer, Berlin, Heidelberg, pp. 1–198.

Alspector, J., Kolcz, A. & Karunanithi, N. (1998), Comparing feature-based and clique-based user models for movie selection, *in* 'Proceedings of the Third ACM Conference on Digital Libraries', DL '98, ACM, New York, NY, USA, pp. 11–18.

Amez, L. (2003), *Experimental Learning in the Arts*, number 34, Discussion papers in business economics, London Metropolitan University.

Amez, L. (2010), 'Mapping the field of arts and economics'. The 16th International Conference of ACEI held at the Copenhagen Business School, Denmark, June 10-12 2010.

Anderson, E. (1998), 'Customer satisfaction and word of mouth', *Journal of Service Research* **1**(1), 5–17.

Anderson, J. R. & Matessa, M. (1990), A rational analysis of categorization, *in* 'Proceedings of the Seventh International Conference (1990) on Machine Learning', Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 76–84.

Armantier, O., Levy-Garboua, L., Owen, C. & Placido, L. (2015), 'Discovering preferences: A theoretical framework and an experiment'. Accessed: 2016-12-27. **URL:** *http://www.saet.uiowa.edu/papers/2015/ALOP_May19_2015.pdf*

Bagella, M. & Becchetti, L. (1999), 'The determinants of motion picture box office performance: Evidence from movies produced in Italy', *Journal of Cultural Economics* **23**(4), 237–256.

Baker, W. & Faulkner, R. (1991), 'Role as resource in the Hollywood film industry', *American Journal of Sociology* **97**(2), 279–309.

Balby Marinho, L., Hotho, A., Jaschke, R., Nanopoulos, A., Rendle, S., Schmidt-Thieme, L., Stumme, G. & Symeonidis, P. (2011), Social tagging recommender systems, *in* 'Recommender systems handbook', Springer US, Boston, MA, pp. 615–644.

Baltas, G. & Doyle, P. (2001), 'Random utility models in marketing research: a survey', *Journal of Business Research* **51**(2), 115–125.

Banfield, J. D. & Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**(3), 803–821.

Basu, C., Hirsh, H. & Cohen, W. (1998), Recommendation as classification: Using social and content-based information in recommendation, *in* 'Proceedings of the

Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence', AAAI '98/IAAI '98, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 714–720.

Basuroy, S., Chatterjee, S. & Ravid, S. (2003), 'How critical are critical reviews? The box office effects of film critics, star power, and budgets', *Journal of Marketing* **67**(4), 103–117.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970), 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *The Annals of Mathematical Statistics* **41**(1), 164–171.

Becher, T. & Trowler, P. (2001), *Academic tribes and territories: intellectual enquiry and the culture of disciplines*, Society for Research into Higher Education & Open University Press, Buckingham, UK.

Becker, G. (1965), 'A theory of the allocation of time', *Economic Journal* **75**(299), 493–517.

Becker, G., Grossman, M. & Murphy, K. (1994), 'An empirical-analysis of cigarette addiction', *American Economic Review* **84**(3), 396–418.

Ben-Akiva, M., Mcfadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., De Palma, A., Gopinath, D., Karlstrom, A. & Munizaga, M. A. (2002), 'Hybrid choice models: Progress and challenges', *Marketing Letters* **13**(3), 163–175.

Bergman, L. R. & Magnusson, D. (1997), 'A person-oriented approach in research on developmental psychopathology', *Development and Psychopathology* **9**(2), 291–319.

Bernardo, J. J. & Blin, J. M. (1977), 'A programming model of consumer choice among multi-attributed brands', *Journal of Consumer Research* **4**(2), 111.

Bianchi, M. (2002), 'Novelty, preferences, and fashion: when goods are unsettling', *Journal of Economic Behavior & Organization* **47**(1), 1–18.

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992), 'A theory of fads, fashion, custom, and cultural-change as informational cascades', *Journal of Political Economy* **100**(5), 992–1026.

Billsus, D. & Pazzani, M. J. (1998), Learning collaborative information filters, *in* 'Proceedings of the Fifteenth International Conference on Machine Learning', ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 46–54.

Blaug, M. (2001), 'Where are we now on cultural economics?', *Journal of Economic Surveys* **15**(2), 123–143.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *J. Mach. Learn. Res.* **3**, 993–1022.

Bohnenkamp, B., Knapp, A.-K., Hennig-Thurau, T. & Schauerte, R. (2015), 'When does it make sense to do it again? An empirical investigation of contingency factors of movie remakes', *Journal of Cultural Economics* **39**(1), 15–41.

Bordwell, D., Staiger, J. & Thompson, K. (1985), *The Classical Hollywood Cinema: Film Style & Mode of Production to 1960*, Columbia University Press, New York.

Bordwell, D. & Thompson, K. (2008), *Film art: an introduction*, McGraw Hill, Boston.

Bozdogan, H. (1987), 'Model selection and Akaike's information criterion (aic): The general theory and its analytical extensions', *Psychometrika* **52**(3), 345–370.

Breese, J. S., Heckerman, D. & Kadie, C. (1998), Empirical analysis of predictive algorithms for collaborative filtering, *in* 'Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence', UAI'98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 43–52.

Burke, R. (2002), 'Hybrid recommender systems: Survey and experiments', *User Modeling and User-Adapted Interaction* **12**(4), 331–370.

Burnham, K. P. & Anderson, D. R. (2002), *Model selection and multimodel inference: a practical information-theoretic approach*, Springer, New York.

Cameron, S. (1999), 'Rational addiction and the demand for cinema', *Applied Economics Letters* **6**(9), 617–620.

Caru, A. & Cova, B. (2007), *Consuming Experience*, Routledge, London and New York.

Cavanaugh, J. E. (2012), 'Lecture v: The Bayesian information criterion'. Accessed: 2016-12-27.
URL: *http://myweb.uiowa.edu/cavaaugh/ms_lec_5_ho.pdf*

Caves, R. (2000), *Creative Industries: Contracts Between Art and Commerce*, Harvard University Press, Cambridge, Massachusetts and London, UK.

Chandler, D. (1997), 'An introduction to genre theory'. Accessed: 2016-12-27.
URL: *http://visual-memory.co.uk/daniel/Documents/intgenre/chandler_genre_theory.pdf*

Christakou, C., Lefakis, L., Vrettos, S. & Stafylopatis, A. (2005), A movie recommender system based on semi-supervised clustering, *in* 'International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)', Vol. 2, pp. 897–903.

Chung, K. H. & Cox, R. A. K. (1994), 'A stochastic model of superstardom: An application of the Yule distribution', *The Review of Economics and Statistics* **76**(4), 771–775.

Cohn, D. & Hofmann, T. (2000), The missing link: A probabilistic model of document content and hypertext connectivity, *in* 'Proceedings of the 13th International Conference on Neural Information Processing Systems', NIPS'00, MIT Press, Cambridge, MA, USA, pp. 409–415.

Collins, A., Hand, C. & Snell, M. C. (2002), 'What makes a blockbuster? Economic analysis of film success in the United Kingdom', *Managerial and Decision Economics* **23**(6), 343–354.

Collins, L. M. & Lanza, S. T. (2010), *Latent Class and Latent Transition Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Collins, L. M. & Wugalter, S. E. (1992), 'Latent class models for stage-sequential dynamic latent variables', *Multivariate Behavioral Research* **27**(1), 131–157.

Corts, K. (2001), 'The strategic effects of vertical market structure: Common agency and divisionalization in the US motion picture industry', *Journal of Economics & Management Strategy* **10**(4), 509–528.

Courty, P. (2011), 'Unpriced quality', *Economics Letters* **111**(1), 13–15.

Cox, D. & Snell, E. (1989), *Analysis of Binary Data, Second Edition*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, Boca Raton, London, New York, Washington DC.

Crane, D. (1972), *Invisible Colleges; Diffusion of Knowledge in Scientific Communities*, University of Chicago Press.

Cui, J., Liu, H., He, J., Li, P., Du, X. & Wang, P. (2011), 'Tagclus: a random walk-based method for tag clustering', *Knowledge and Information Systems* **27**(2), 193–225.

Dattolo, A., Eynard, D. & Mazzola, L. (2011), An integrated approach to discover tag semantics, *in* 'Proceedings of the 2011 ACM Symposium on Applied Computing', SAC '11, ACM, New York, NY, USA, pp. 814–820.

de Solla Price, D. (1963), *Little Science, Big Science - and Beyond*, Columbia University Press.

De Vany, A. (2004), *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry*, Contemporary political economy series, Routledge, London and New York.

De Vany, A. (2005), Contracting with stars when "nobody knows anything", *in* 'Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry', Routledge, London and New York, pp. 231–254.

De Vany, A. S. & Eckert, R. (1991), 'Motion picture antitrust: the Paramount cases revisited', *Research in Law and Economics* (14), 51–112.

De Vany, A. & Walls, W. (1996), 'Bose-Einstein dynamics and adaptive contracting in the motion picture industry', *Economic Journal* **106**(439), 1493–1514.

De Vany, A. & Walls, W. (1997), 'The market for motion pictures: Rank, revenue, and survival', *Economic Inquiry* **35**(4), 783–797.

De Vany, A. & Walls, W. (2002), 'Does Hollywood make too many R-rated movies? Risk, stochastic dominance, and the illusion of expectation', *Journal of Business* **75**(3), 425–451.

De Vany, A. & Walls, W. (2004), 'Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar', *Journal of Economic Dynamics & Control* **28**(6), 1035–1057.

De Vany, A. & Walls, W. D. (1999), 'Uncertainty in the movie industry: Does star power reduce the terror of the box office?', *Journal of Cultural Economics* **23**(4), 285–318.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of The American Society For Information Science* **41**(6), 391–407.

DeFillippi, R. J. & Arthur, M. B. (1998), 'Paradox in project-based enterprise: The case of film making', *California Management Review* **40**(2), 125–139.

Dellarocas, C., Zhang, X. M. & Awad, N. F. (2007), 'Exploring the value of online product reviews in forecasting sales: The case of motion pictures', *Journal of Interactive Marketing* **21**(4), 23–45.

Delmestri, G., Montanari, F. & Usai, A. (2005), 'Reputation and strength of ties in predicting commercial success and artistic merit of independents in the Italian feature film industry', *Journal of Management Studies* **42**(5), 975–1002.

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

DeSilva, I. (1998), Consumer selection of motion pictures, *in* 'The motion picture mega industry', Allyn and Bacon, Boston, MA, pp. 144–171.

Desrosiers, C. & Karypis, G. (2011), A comprehensive survey of neighborhood-based recommender methods, *in* 'Recommender Systems Handbook', Springer US, Boston, MA, pp. 107–144.

Deuchert, E., Adjamah, K. & Pauly, F. (2005), 'For oscar glory or oscar money?', *Journal of Cultural Economics* **29**(3), 159–176.

Deutsch, S., Schrammel, J. & Tscheligi, M. (2011), Comparing different layouts of tag clouds: Findings on visual perception, *in* 'Proceedings of the Second IFIP WG 13.7 Conference on Human-computer Interaction and Visualization', HCIV'09, Springer-Verlag, Berlin, Heidelberg, pp. 23–37.

Dewey, J. (1934), *Art as Experience*, Perigee Books, New York.

Dhar, R., Nowlis, S. & Sherman, S. (2000), 'Trying hard or hardly trying: An analysis of context effects in choice', *Journal of Consumer Psychology* **9**(4), 189–200.

Dhar, R. & Simonson, I. (1992), 'The effect of the focus of comparison on consumer preferences', *Journal of Marketing Research* **29**(4), 430–440.

Eggenberger, F. & Polya, G. (1923), 'Uber die statistik verketteter vorgange', *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift fur Angewandte Mathematik und Mechanik* **3**(4), 279–289.

Einav, L. (2003), 'Not all rivals look alike: estimating an equilibrium model of the release date timing game'. Accessed: 2016-12-27.
URL: *ftp://ftp.cemfi.es/pdf/papers/wshop/einav.pdf*

Einav, L. (2007), 'Seasonality in the U.S. motion picture industry', *The RAND Journal of Economics* **38**(1), 127–145.

Einav, L. (2010), 'Not all rivals look alike: Estimating an equilibrium model of the release date timing game', *Economic Inquiry* **48**(2), 369–390.

Ekstrand, M. D., Riedl, J. T. & Konstan, J. A. (2011), 'Collaborative filtering recommender systems', *Foundations and Trends in Human-Computer Interaction* **4**(2), 81–173.

Elberse, A. & Eliashberg, J. (2003), 'Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures', *Marketing Science* **22**(3), 329–354.

Eliashberg, J., Hui, S. K. & Zhang, Z. J. (2014), 'Assessing box office performance using movie scripts: A kernel-based approach', *IEEE Transactions on Knowledge and Data Engineering* **26**(11), 2639–2648.

Eliashberg, J., Jonker, J., Sawhney, M. & Wierenga, B. (2000), 'MOVIEMOD: An implementable decision-support system for prerelease market evaluation of motion pictures', *Marketing Science* **19**(3), 226–243.

Eliashberg, J. & Shugan, S. (1997), 'Film critics: Influencers or predictors?', *Journal of Marketing* **61**(2), 68–78.

Faber, R. & Oguinn, T. (1984), 'Effect of media advertising and other sources on movie selection', *Journalism quarterly* **61**(2), 371–377.

Fama, E. (1963), 'Mandelbrot and the stable Paretian hypothesis', *Journal of Business* **36**(4), 420–429.

Fama, E. F. (1965), 'The behavior of stock-market prices', *The Journal of Business* **38**(1), 34–105.

Faulkner, R. & Anderson, A. (1987), 'Short-term projects and emergent careers - Evidence from Hollywood', *American Journal of Sociology* **92**(4), 879–909.

Feinerer, I., Hornik, K. & Meyer, D. (2008), 'Text mining infrastructure in R', *Journal of Statistical Software* **25**(1), 1–54.

Filson, D., Switzer, D. & Besocke, P. (2005), 'At the movies: The economics of exhibition contracts', *Economic Inquiry* **43**(2), 354–369.

Frank, B. (1994), 'Optimal timing of movie releases in ancillary markets: The case of video releases', *Journal of Cultural Economics* **18**(2), 125–133.

Fu, W. & Govindaraju, A. (2010), 'Explaining global box-office tastes in Hollywood films: Homogenization of national audiences' movie selections', *Communication Research* **37**(2), 215–238.

Gantner, Z., Drumond, L., Freudenthaler, C., Rendle, S. & Schmidt-Thieme, L. (2010), Learning attribute-to-feature mappings for cold-start recommendations, *in* 'Proceedings of the 2010 IEEE International Conference on Data Mining', ICDM '10, IEEE Computer Society, Washington, DC, USA, pp. 176–185.

Garboua, L. L. & Montmarquette, C. (1996), 'A microeconometric study of theatre demand', *Journal of Cultural Economics* **20**(1), 25–50.

Gelman, A., Hwang, J. & Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', *Statistics and Computing* **24**(6), 997–1016.
URL: *http://dx.doi.org/10.1007/s11222-013-9416-2*

Gemmell, J., Shepitsen, A., Mobasher, B. & Burke, R. (2008), Personalizing navigation in folksonomies using hierarchical tag clustering, *in* I.-Y. Song, J. Eder & T. M. Nguyen, eds, 'Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings', Springer, Berlin, Heidelberg, pp. 196–205.

Gentile, C., Spiller, N. & Noci, G. (2007), 'How to sustain the customer experience: An overview of experience components that co-create value with the customer', *European Management Journal* **25**(5), 395–410.

Glancy, H. (1992), 'MGM Film Grosses, 1924-1948: The Eddie Mannix Ledger', *Historical Journal of Film Radio and Television* **12**(2), 127–144.

Glancy, H. (1995), 'Warner-Bros film grosses, 1921-51: The William Schaefer Ledger', *Historical Journal of Film Radio and Television* **15**(1), 55–73.

Goettler, R. & Leslie, P. (2005), 'Cofinancing to manage risk in the motion picture industry', *Journal of Economics & Management Strategy* **14**(2), 231–261.

Goetzmann, W. N., Ravid, S. A. & Sverdlove, R. (2013), 'The pricing of soft and hard information: economic lessons from screenplay sales', *Journal of Cultural Economics* **37**(2), 271–307.

Goldberg, D., Nichols, D., Oki, B. M. & Terry, D. (1992), 'Using collaborative filtering to weave an information tapestry', *Commun. ACM* **35**(12), 61–70.

Golder, S. A. & Huberman, B. A. (2006), 'Usage patterns of collaborative tagging systems', *J. Inf. Sci.* **32**(2), 198–208.

Goldman, W. (1983), *Adventures in the Screen Trade, A Personal View of Hollywood and Screenwriting*, Warner Books, New York.

Gomery, D. (1992), *Shared pleasures: a history of movie presentation in the United States*, Wisconsin studies in film, University of Wisconsin Press, Madison.

Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J. & Riedl, J. (1999), Combining collaborative filtering with personal agents for better recommendations, *in* 'Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence', AAAI '99/IAAI '99, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 439–446.

Goodman, L. A. (1974*a*), 'The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A modified latent structure approach', *American Journal of Sociology* **79**(5), 1179–1259.

Goodman, L. A. (1974*b*), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**, 215–231.

Gorman, W. M. (1953), 'Community preference fields', *Econometrica* **21**(1), 63–80.

Greene, W. H. & Hensher, D. A. (2003), 'A latent class model for discrete choice analysis: contrasts with mixed logit', *Transportation Research Part B: Methodological* **37**(8), 681–698.

Griffiths, T. L. & Steyvers, M. (2004), 'Finding scientific topics', *Proceedings of the National academy of Sciences* **101**(suppl 1), 5228–5235.

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B. (2007), 'Topics in semantic representation', *Psychological Review* **114**(2), 211–244.

Grun, B. & Hornik, K. (2011), 'topicmodels: An R package for fitting topic models', *Journal of Statistical Software* **40**(1), 1–30.

Gunawardana, A. & Meek, C. (2009), A unified approach to building hybrid recommender systems, *in* 'Proceedings of the Third ACM Conference on Recommender Systems', RecSys '09, ACM, New York, NY, USA, pp. 117–124.

Guy, M. & Tonkin, E. (2006), 'Folksonomies tidying up tags?', *D-Lib Magazine* **12**(1).

Haberman, S. J. (1974), 'Log-linear models for frequency tables with ordered classifications', *Biometrics* **30**(4), 589–600.

Hand, C. (2001), 'Increasing returns to information: further evidence from the UK film market', *Applied Economics Letters* **8**(6), 419–421.

Hanssen, F. A. (2000), 'The block booking of films reexamined', *The Journal of Law & Economics* **43**(2), 395–426.

Harper, F. M. & Konstan, J. A. (2015), 'The movielens datasets: History and context', *ACM Trans. Interact. Intell. Syst.* **5**(4), 19:1–19:19.

Hennig-Thurau, T., Houston, M. B. & Heitjans, T. (2009), 'Conceptualizing and measuring the monetary value of brand extensions: The case of motion pictures', *Journal of Marketing* **73**(6), 167–183.

Hennig-Thurau, T., Houston, M. B. & Walsh, G. (2007), 'Determinants of motion picture box office and profitability: an interrelationship approach', *Review of Managerial Science* **1**(1), 65–92.

Herlocker, J. L., Konstan, J. A., Borchers, A. & Riedl, J. (1999), An algorithmic framework for performing collaborative filtering, *in* 'Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '99, ACM, New York, NY, USA, pp. 230–237.
URL: *http://doi.acm.org/10.1145/312624.312682*

Hirsch, P. M. (1972), 'Processing fads and fashions: An organization-set analysis of cultural industry systems', *American Journal of Sociology* **77**(4), 639–659.

Hodge, R. & Kress, G. (1988), *Social Semiotics*, Cornell University Press, Ithaca, NY.

Hofmann, T. (1999), Probabilistic latent semantic indexing, *in* 'Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', SIGIR '99, ACM, New York, NY, USA, pp. 50–57.

Hofmann, T. (2004), 'Latent semantic models for collaborative filtering', *ACM Trans. Inf. Syst.* **22**(1), 89–115.

Hofmann, T. & Puzicha, J. (1999), Latent class models for collaborative filtering, *in* 'Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2', IJCAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 688–693.

Hofstede, G. (1980), *Culture's Consequences: International Differences in Work-Related Values*, Cross-Cultural Research and Methodology Series, Sage Publications, Thousand Oaks, London, New Delhi.

Hofstede, G. & Bond, M. (1988), 'The confucius connection - From cultural roots to economic-growth', *Organizational Dynamics* **16**(4), 5–21.

Holbrook, M. (1987), Perception et representation esthetiques du consommateur, *in* 'Economie et Culture', La Documentation Française, Paris, pp. 147–155.

Holbrook, M. (1999), 'Popular appeal versus expert judgements of motion pictures', *Journal of Consumer Research* **26**(2), 144–155.

Holbrook, M. (2005), 'The role of ordinary evaluations in the market for popular culture: Do consumers have "good taste"?', *Marketing Letters* **16**(2), 75–86.

Holbrook, M. B. (1993), 'Nostalgia and consumption preferences: Some emerging patterns of consumer tastes', *Journal of Consumer Research* **20**(2), 245–256.

Holbrook, M. B. & Addis, M. (2007), 'Taste versus the market: An extension of research on the consumption of popular culture', *Journal of Consumer Research* **34**(3), 415–424.

Holbrook, M. B. & Hirschman, E. C. (1982), 'The experiential aspects of consumption: Consumer fantasies, feelings, and fun', *Journal of Consumer Research* **9**(2), 132–140.

Hornik, K. (2007), 'The snowball package'. Accessed: 2016-12-27.
URL: *http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/packages/Snowball.pdf*

Hornik, K. & Grun, B. (2014), 'movmf: An r package for fitting mixtures of von Mises-Fisher distributions', *Journal of Statistical Software* **58**(1), 1–31.

Hosany, S. & Witham, M. (2010), 'Dimensions of cruisers' experiences, satisfaction, and intention to recommend', *Journal of Travel Research* **49**(3), 351–364.

Ijiri, Y. & Simon, H. (1977), *Skew distributions and the sizes of business firms*, Studies in mathematical and managerial economics, North-Holland, Amsterdam.

Jayakar, K. & Waterman, D. (2000), 'The economics of American theatrical movie exports: An empirical analysis', *Journal of Media Economics* **13**(3), 153–169.

Jin, R., Chai, J. & Si, L. (2004), An automatic weighting scheme for collaborative filtering, *in* K. Jarvelin, J. Allen, P. Bruza & M. Sanderson, eds, 'Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 337–344.

Jovanovic, B. (1987), 'Micro shocks and aggregate risk', *The Quarterly Journal of Economics* **102**(2), 395–409.

Kagie, M., van der Loos, M. & van Wezel, M. (2009), 'Including item characteristics in the probabilistic latent semantic analysis model for collaborative filtering', *AI Commun.* **22**(4), 249–265.

Kahneman, D. & Snell, J. (1992), 'Predicting a changing taste: Do people know what they will like?', *Journal of Behavioral Decision Making* **5**(3), 187–200.

Karniouchina, E. V. (2011), 'Impact of star and movie buzz on motion picture distribution and box office revenue', *International Journal of Research in Marketing* **28**(1), 62–74.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**(430), 773–795.

Kenney, R. W. & Klein, B. (1983), 'The economics of block booking', *The Journal of Law & Economics* **26**(3), 497–540.

Kihlstrom, R., Mirman, L. & Postlewaite, A. (1984), Experimental consumption and the Rothschild effect, *in* 'Bayesian Models in Economic Theory', North Holland, Amsterdam, pp. 279–302.

King, T. (2007), 'Does film criticism affect box office earnings? evidence from movies released in the U.S. in 2003', *Journal of Cultural Economics* **31**(3), 171–186.

Kipp, M. E. & Campbell, D. G. (2006), 'Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices', *Proceedings of the American Society for Information Science and Technology* **43**(1), 1–18.

Kivetz, R., Netzer, O. & Schrift, R. (2008), 'The synthesis of preference: Bridging behavioral decision research and marketing science', *Journal of Consumer Psychology* **18**(3), 179–186.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R. & Riedl, J. (1997), 'Grouplens: Applying collaborative filtering to usenet news', *Commun. ACM* **40**(3), 77–87.

Koren, Y. & Bell, R. (2011), Advances in collaborative filtering, *in* 'Recommender Systems Handbook', Springer US, Boston, MA, pp. 145–186.

Krestel, R., Fankhauser, P. & Nejdl, W. (2009), Latent Dirichlet Allocation for tag recommendation, *in* 'Proceedings of the Third ACM Conference on Recommender Systems', RecSys '09, ACM, New York, NY, USA, pp. 61–68.

Krishnan, A. (1998), *Of Wealth Power and Law: the Origin of Scaling in Economics*, Racah Institute of Physics, The Hebrew University of Jerusalem.

Krishnan, A. (2009), *What are academic disciplines*, number 03/09, NCRM working paper series.

Lampel, J., Lant, T. & Shamsie, J. (2000), 'Balancing act: Learning from organizing practices in cultural industries', *Organization Science* **11**(3), 263–269.

Lancaster, K. J. (1966), 'A new approach to consumer theory', *Journal of Political Economy* **74**(2), 132–157.

LaSalle, D. & Britton, T. (2003), *Priceless: Turning Ordinary Products Into Extraordinary Experiences*, Harvard Business School Publishing, Boston Massachusetts.

Lazarsfeld, P. & Henry, N. (1968), *Latent structure analysis*, Houghton, Mifflin.

Lekakos, G. & Caravelas, P. (2008), 'A hybrid approach for movie recommendation', *Multimedia Tools and Applications* **36**(1), 55–70.

Li, H.-z., Hu, X.-g., Lin, Y.-j., He, W. & Pan, J.-h. (2016), 'A social tag clustering method based on common co-occurrence group similarity', *Frontiers of Information Technology & Electronic Engineering* **17**(2), 122–134.

Liang, H., Xu, Y., Li, Y., Nayak, R. & Tao, X. (2010), Connecting users and items with weighted tags for personalized item recommendations, *in* 'Proceedings of the 21st ACM Conference on Hypertext and Hypermedia', HT '10, ACM, New York, NY, USA, pp. 51–60.

Lievrouw, L. (2014), Reconciling structure and process in the study of scholarly communication, *in* 'Scholarly Communication and Bibliometrics', Sage Publications, Newbury Park, CA, pp. 59–69.

Litman, B. (1983), 'Predicting success of theatrical movies - an empirical-study', *Journal of Popular Culture* **16**(4), 159–175.

Litman, B. R. & Kohl, L. S. (1989), 'Predicting financial success of motion pictures: The '80s experience', *Journal of Media Economics* **2**(2), 35–50.

Liu, Y. (2006), 'Word of mouth for movies: Its dynamics and impact on box office revenue', *Journal of Marketing* **70**(3), 74–89.

Lops, P., de Gemmis, M. & Semeraro, G. (2011), Content-based recommender systems: State of the art and trends, *in* 'Recommender Systems Handbook', Springer US, Boston, MA, pp. 73–105.

Luce, R. (1959), *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York.

Luehrman, T. A. & Teichner, W. A. (1992), *Arundel Partners: The Sequel Project*, Harvard Business Review.

Macgregor, G. & McCulloch, E. (2006), 'Collaborative tagging as a knowledge organisation and resource discovery tool', *Library Review* **55**(5), 291–300.

Mandelbrot, B. (1963*a*), 'New Methods in Statistical Economics', *Journal of Political Economy* **71**(5), 421–440.

Mandelbrot, B. (1963*b*), 'The variation of certain speculative prices', *Journal of Business* **36**(4), 394–419.

Mandelbrot, B. B. (1997), *Fractals and Scaling in Finance*, Springer-Verlag, New York.

Manski, C. F. (1977), 'The structure of random utility models', *Theory and Decision* **8**(3), 229–254.

Marlow, C., Naaman, M., Boyd, D. & Davis, M. (2006), Ht06, tagging paper, taxonomy, flickr, academic article, to read, *in* 'Proceedings of the Seventeenth Conference on Hypertext and Hypermedia', HYPERTEXT '06, ACM, New York, NY, USA, pp. 31–40.

Marschak, J. (1960), Binary-choice constraints and random utility indicators, *in* 'Mathematical models in the social sciences, Proceedings of the first Stanford symposium, Stanford mathematical studies in the social sciences', Stanford University Press, Stanford, California, pp. 312–329.

McCutcheon, A. (1987), *Latent Class Analysis*, number 64 *in* 'Quantitative Applications in the Social Sciences', Sage Publications, Thousand Oaks, London, New Delhi.

McFadden, D. (1973), Conditional logit analysis of qualitative choice behavior, *in* 'Frontiers in Econometrics', Academic Press, New York, pp. 105–142.

McFadden, D. (1980), Econometric models of probabilistic choice, *in* 'Structural analysis of discrete data with econometric applications', MIT Press, Cambridge MA, pp. 198–272.

McFadden, D. (2014), The new science of pleasure: consumer choice behavior and the measurement of well-being, *in* 'Handbook of Choice Modelling', Edward Elgar Publishing, Inc., Cheltenham, UK, pp. 7–48.

McKenzie, J. (2012), 'The economics of movies: A literature survey', *Journal of Economic Surveys* **26**(1), 42–70.

Melville, P., Mooney, R. J. & Nagarajan, R. (2002), Content-boosted collaborative filtering for improved recommendations, *in* 'Eighteenth National Conference on Artificial Intelligence', American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 187–192.

Menard, S. (2000), 'Coefficients of determination for multiple logistic regression analysis', *The American Statistician* **54**(1), 17–24.

Mezias, J. & Mezias, S. (2000), 'Resource partitioning, the founding of specialist firms, and innovation: The American feature film industry, 1912-1929', *Organization Science* **11**(3), 306–322.

Milicevic, A. K., Nanopoulos, A. & Ivanovic, M. (2010), 'Social tagging in recommender systems: A survey of the state-of-the-art and possible extensions', *Artif. Intell. Rev.* **33**(3), 187–209.

Millen, D. R., Yang, M., Whittaker, S. & Feinberg, J. (2007), Social bookmarking and exploratory search, *in* L. J. Bannon, I. Wagner, C. Gutwin, R. H. R. Harper & K. Schmidt, eds, 'ECSCW 2007: Proceedings of the 10th European Conference on Computer-Supported Cooperative Work, Limerick, Ireland, 24-28 September 2007', Springer, London, pp. 21–40.

Miller, D. & Shamsie, J. (1996), 'The resource-based view of the firm in two environments: The Hollywood film studios from 1936 to 1965', *Academy Of Management Journal* **39**(3), 519–543.

Moon, S., Bergey, P. K. & Iacobucci, D. (2010), 'Dynamic effects among movie ratings, movie revenues, and viewer satisfaction', *Journal of Marketing* **74**(1), 108–121.

Morris, S. A. & Van der Veer Martens, B. (2008), 'Mapping research specialties', *Annual review of information science and technology* **42**(1), 213–295.

Moul, C. C. (2007), 'Measuring word of mouth's impact on theatrical movie admissions', *Journal of Economics & Management Strategy* **16**(4), 859–892.

Nasery, M., Elahi, M. & Cremonesi, P. (2015), Polimovie: a feature-based dataset for recommender systems, *in* 'ACM RecSys Workshop on Crowdsourcing and Human Computation for Recommender Systems (CrawdRec)', Vol. 3, ACM, pp. 25–30.

Neale, S. (1980), *Genre*, BFI books, British Film Institute, London.

Neelamegham, R. & Chintagunta, P. (1999), 'A Bayesian model to forecast new product performance in domestic and international markets', *Marketing Science* **18**(2), 115–136.

Nelson, P. (1970), 'Information and consumer behavior', *Journal of Political Economy* **78**(2), 311–329.

Nelson, R., Donihue, M., Waldman, D. & Wheaton, C. (2001), 'What's an oscar worth?', *Economic Inquiry* **39**(1), 1–16.

Nosofsky, R. (1986), 'Attention, similarity, and the identification-categorization relationship', *Journal of Experimental Psychology-General* **115**(1), 39–57.

O'Connell, A. (2006), *Logistic Regression Models for Ordinal Response Variables*, number 146 *in* 'Quantitative Applications in the Social Sciences', Sage Publications, Thousand Oaks, London, New Delhi.

Oh, H., Fiore, A. M. & Jeoung, M. (2007), 'Measuring experience economy concepts: Tourism applications', *Journal of Travel Research* **46**(2), 119–132.

Oh, J. (2001), 'International trade in film and the self-sufficiency ratio', *Journal of Media Economics* **14**(1), 31–44.

Orbach, B. Y. & Einav, L. (2007), 'Uniform prices for differentiated goods: The case of the movie-theater industry', *International Review Of Law And Economics* **27**(2), 129–153.

Paas, L. J., Bijmolt, T. H. & Vermunt, J. K. (2007), 'Acquisition patterns of financial products: A longitudinal investigation', *Journal of Economic Psychology* **28**(2), 229 – 241.

Pang, B. & Lee, L. (2008), 'Opinion mining and sentiment analysis', *Found. Trends Inf. Retr.* **2**(1-2), 1–135.

Peltoniemi, M. (2015), 'Cultural industries: Product-market characteristics, management challenges and industry dynamics', *International Journal of Management Reviews* **17**(1), 41–68.

Pentreath, N. (2015), *Machine Learning with Spark*, Packt Publishing, Birmingham, Mumbai.

Pine, B. & Gilmore, J. (1999), *The Experience Economy: Work is Theatre & Every Business a Stage*, Harvard Business Publishing, Harvard Business School Press.

Pitsilis, G. & Wang, W. (2015), 'Harnessing the power of social bookmarking for improving tag-based recommendations', *Computers in Human Behavior* **50**, 239–251.

Porteous, I., Asuncion, A. & Welling, M. (2010), Bayesian matrix factorization with side information and Dirichlet process mixtures, *in* 'Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence', AAAI'10, AAAI Press, California, USA, pp. 563–568.

Prag, J. & Casavant, J. (1994), 'An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry', *Journal of Cultural Economics* **18**(3), 217–235.

Quintarelli, E. (2005), 'Folksonomies: power to the people'. Accessed: 2016-12-27. **URL:** *http://www.iskoi.org/doc/folksonomies.htm*

Ravid, S. A. (1999), 'Information, blockbusters, and stars: A study of the film industry', *The Journal of Business* **72**(4), 463–492.

Ravid, S. & Basuroy, S. (2004), 'Managerial objectives, the r-rating puzzle, and the production of violent films', *Journal of Business* **77**(suppl. 2), S155–S192.

Reinstein, D. & Snyder, C. (2005), 'The influence of expert reviews on consumer demand for experience goods: A case study of movie critics', *Journal of Industrial Economics* **53**(1), 27–51.

Resnick, P. & Varian, H. R. (1997), 'Recommender systems', *Commun. ACM* **40**(3), 56–58.

Resnik, P. (1995), Using information content to evaluate semantic similarity in a taxonomy, *in* 'Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1', IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 448–453.

Ricci, F., Rokach, L. & Shapira, B. (2011), Introduction to recommender systems handbook, *in* 'Recommender systems handbook', Springer US, Boston, MA, pp. 1–35.

Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B. (2011), *Recommender systems handbook*, Springer, Boston, MA.

Robins, J. A. (1993), 'Organization as strategy: Restructuring production in the film industry', *Strategic Management Journal* **14**(S1), 103–118.

Rosen, S. (1981), 'The economics of superstars', *American Economic Review* **71**(5), 845–858.

Ruscio, J. & Ruscio, A. M. (2008), 'Categories and dimensions', *Current Directions in Psychological Science* **17**(3), 203–207.

Salter, J. & Antonopoulos, N. (2006), 'Cinemascreen recommender agent: Combining collaborative and content-based filtering', *IEEE Intelligent Systems* **21**(1), 35–41.

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. (2000), Analysis of recommendation algorithms for e-commerce, *in* 'Proceedings of the 2nd ACM Conference on Electronic Commerce', EC '00, ACM, New York, NY, USA, pp. 158–167.

Sawhney, M. S. & Eliashberg, J. (1996), 'A parsimonious model for forecasting gross box-office revenues of motion pictures', *Marketing Science* **15**(2), 113–131.

Schmitt, B. (1999), *Experiential Marketing: How to Get Customers to Sense, Feel, Think, Act, Relate*, Free Press, New York.

Schmitt, B. (2011), 'Experience marketing: Concepts, frameworks and consumer insights', *Foundations and Trends in Marketing* **5**(2), 55–112.

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

Sedgwick, J. (2007), A Shacklean approach to the demand for movies, *in* 'The Evolution of Consumption: Theories and Practices', Vol. 10 of *Advances in Austrian Economics*, Emerald Group Publishing Limited, pp. 77–91.

Sedgwick, J. & Pokorny, M. (1998), 'The risk environment of film making: Warner Bros in the inter-war years', *Explorations in Economic History* **35**(2), 196–220.

Sedgwick, J. & Pokorny, M. (2010), 'Consumers as risk takers: Evidence from the film industry during the 1930s', *Business History* **52**(1), 74–99.

Sen, S., Harper, F. M., LaPitz, A. & Riedl, J. (2007), The quest for quality tags, *in* 'Proceedings of the 2007 International ACM Conference on Supporting Group Work', GROUP '07, ACM, New York, NY, USA, pp. 361–370.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M. & Riedl, J. (2006), Tagging, communities, vocabulary, evolution, *in* 'Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work', CSCW '06, ACM, New York, NY, USA, pp. 181–190.

Sen, S., Vig, J. & Riedl, J. (2009*a*), Learning to recognize valuable tags, *in* 'Proceedings of the 14th International Conference on Intelligent User Interfaces', IUI '09, ACM, New York, NY, USA, pp. 87–96.

Sen, S., Vig, J. & Riedl, J. (2009*b*), Tagommenders: Connecting users to items through tags, *in* '18th International World Wide Web Conference', pp. 671–671.

Sharma, R., Nigam, S. & Jain, R. (2014), 'Opinion mining of movie reviews at document level', *International Journal on Information Theory (IJIT)* **3**(3), 13–21.

Shaw, C. & Ivens, J. (2002), *Building Great Customer Experiences*, Palgrave Macmillan, Hampshire, New York.

Shepitsen, A., Gemmell, J., Mobasher, B. & Burke, R. (2008), Personalized recommendation in social tagging systems using hierarchical clustering, *in* 'Proceedings of the 2008 ACM Conference on Recommender Systems', RecSys '08, ACM, New York, NY, USA, pp. 259–266.

Shirky, C. (2005), 'Ontology is overrated: Categories, links, and tags'. Accessed: 2016-12-27.
URL: *http://www.shirky.com/writings/ontology_overrated.html*

Simonson, I., Carmon, Z., Dhar, R., Drolet, A. & Nowlis, S. M. (2001), 'Consumer research: In search of identity', *Annual Review of Psychology* **52**(1), 249–275.

Simonton, D. K. (2009), 'Cinematic success criteria and their Predictors: The art and business of the film industry', *Psychology & Marketing* **26**(5), 400–420.

Sinha, R. (2005), 'A cognitive analysis of tagging'. Accessed: 2016-12-27.
URL: *https://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/*

Sisto, A. & Zanola, R. (2010), 'Cinema attendance in Europe', *Applied Economics Letters* **17**(5), 515–517.

Small, H. (1973), 'Co-citation in the scientific literature: A new measure of the relationship between two documents', *Journal of the American Society for Information Science* **24**(4), 265–269.

Smith, S. & Smith, V. (1986), 'Successful movies - a preliminary empirical-analysis', *Applied Economics* **18**(5), 501–507.

Stam, R. (2000), *Film Theory: An Anthology*, Wiley, Malden, Massachusetts and Oxford, UK.

Stigler, G. & Becker, G. (1977), 'De gustibus non est disputandum', *American Economic Review* **67**(2), 76–90.

Stoyanovich, J., Yahia, S. A., Marlow, C. & Yu, C. (2008), A study of the benefit of leveraging tagging behavior to model users' interests in del.icio.us, *in* 'In AAAI Spring Symposium on Social Information Processing'.

Sunada, M. (2010), 'Vertical integration in the Japanese movie industry', *Journal of Industry, Competition and Trade* **10**(2), 135–150.

Symeonidis, P. (2008), Content-based dimensionality reduction for recommender systems, *in* 'Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007', Springer, Berlin, Heidelberg, pp. 619–626.

Szomszor, M., Cattuto, C., Alani, H., O Hara, K., Baldassarri, A., Loreto, V. & Servedio, V. D. (2007), Folksonomies, the semantic web, and movie recommendation. 4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0, 3-7th, June 2007.

Thompson, K. (1985), *Exporting entertainment: America in the world film market, 1907-1934*, BFI books, British Film Institute, London.

Throsby, D. (2001), *Economics and Culture*, Cambridge University Press, Cambridge, UK.

Tolson, A. (1996), *MEDIAtions: Text and Discourse in Media Studies*, Hodder Arnold Publication, London.

Tso-Sutter, K. H. L., Marinho, L. B. & Schmidt-Thieme, L. (2008), Tag-aware recommender systems by fusion of collaborative filtering algorithms, *in* 'Proceedings

of the 2008 ACM Symposium on Applied Computing', SAC '08, ACM, New York, NY, USA, pp. 1995–1999.

Tversky, A. (1977), 'Features of similarity', *Psychological Review* **84**(4), 327–352.

Tversky, A. & Kahneman, D. (1974), 'Judgment under uncertainty - Heuristics and biases', *Science* **185**(4157), 1124–1131.

Uebersax, J. (2010), 'Latent structure models for the analysis of rater agreement and multiple diagnostic tests'. Accessed: 2016-12-27.
URL: *http://www.john-uebersax.com/stat/lsm.htm*

Uhrmacher, P. B. (2009), 'Toward a theory of aesthetic learning experiences', *Curriculum Inquiry* **39**(5), 613–636.

Vander Wal, T. (2007), 'Folksonomy coinage and definition'. Accessed: 2016-12-27.
URL: *http://vanderwal.net/folksonomy.html*

Varian, H. (1998), 'Markets for information goods'. Accessed: 2016-12-27.
URL: *http://people.ischool.berkeley.edu/ hal/Papers/japan/*

Vermunt, J. (2008), Latent class models in longitudinal research, *in* 'Handbook of Longitudinal Research: Design, Measurement, and Analysis', Elsevier, Burlington, MA, pp. 373–385.

Vermunt, J. K., Langeheine, R. & Bockenholt, U. (1999), 'Discrete-time discrete-state latent markov models with time-constant and time-varying covariates', *Journal of Educational and Behavioral Statistics* **24**(2), 179–207.

Vermunt, J. & Magidson, J. (2005), *Technical Guide for Latent GOLD 4.0: Basic and Advanced*, Belmont, Massachusetts: Statistical Innovations Inc.

Vig, J., Sen, S. & Riedl, J. (2011), Navigating the tag genome, *in* 'Proceedings of the 16th International Conference on Intelligent User Interfaces', IUI '11, ACM, New York, NY, USA, pp. 93–102.

Wallace, W. T., Seigerman, A. & Holbrook, M. B. (1993), 'The role of actors and actresses in the success of films: How much is a movie star worth?', *Journal of Cultural Economics* **17**(1), 1–27.

Weinstein, M. (1998), 'Profit-sharing contracts in Hollywood: Evolution and analysis', *Journal of Legal Studies* **27**(1), 67–112.

Westbrook, R. (1987), 'Product-consumption-based affective responses and postpurchase processes', *Journal of Marketing Research* **24**(3), 258–270.

Wetzker, R., Umbrath, W. & Said, A. (2009), A hybrid approach to item recommendation in folksonomies, *in* 'Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval', ESAIR '09, ACM, New York, NY, USA, pp. 25–29.

Wilde, L. L. (1980), 'The economics of consumer information acquisition', *The Journal of Business* **53**(3), S143–S158.

Wildman, S. & Siwek, S. (1988), *International Trade in Films and Television Programs*, AEI trade in services series, Ballinger, Cambridge, Massachusetts.

Zufryden, F. (1996), 'Linking advertising to box office performance of new film releases - A marketing planning model', *Journal of Advertising Research* **36**(4), 29–41.

Glossary

Bass diffusion model Description of the adaptation process of a new product in terms of innovators and imitators

Bayesian information criterion Index used to compare competing models composed out of likelihood value and number of estimation parameters in the model

Bibliometrics Quantitative analysis of science information retrieval, communication and dissemination

Collaborative filtering Technique for automated recommendation coming from filtering information based on user or item proximity

Contiguity A series of events, stimuli and responses occurring in proximity

Contingent valuation Technique to infer the value for non-market sources from stated preferences

Contrast model Model of feature similarity initiated by Amos Tversky

Credence good Product where attribute claims are uncertain at a pre-purchase stage and where, unlike experience goods, uncertainty remains after consumption

Experience goods Product where attribute claims are uncertain at a pre-purchase stage and can only be lifted after the product is consumed

Folkosonomy Bottum-up taxonomy emerging from the collective addition of tags

Information good Product where the value is determined largely by the information on a carrier rather than by the carrier itself

k-neirest neighbour Classification method taking the k closest items as determined by a distance metric

Latent Class Analysis Statistical technique to identify unobservable subgroups or segments from observed, usually categorical variables, through a pattern of conditional probabilities

Latent Dirichlet Allocation Method to discover latent classes, mainly used to discover topics in textual corpora, by inferring from a generative prior statistical model placing a Dirichlet prior on the topic distribution

Latent Semantic Indexing Statistical method to index documents by allocating them to hidden topics calculated through matrix decomposition

Logistic Regression Regression between a categorical dependent, often a translation of a qualitative response, and a set of explanatory variables which can be discrete and/or continuous

Long tail distribution Distribution characterised by a high portion of events having a low frequency of occurrence and a low portion of events having a high frequency of occurrence

Maltin rating Movie rating system assembled in Leonard Maltin's Movie Guide

Markov Model Stochastic model where the transition probability of the next period state is dependent on the value of the present state only

MPAA rating System of labelling awarded by the Motion Picture Association of America based on the film's esteemed suitability for certain audiences

Novelty good Product that changes in attribute composition at each issue

Random Utility Model Mathematical model assuming discrete choice and decomposing an agent's utility into a deterministic component and a stochastic unobserved error component.

Rational Addiction Model Description of addictive behaviour in a setting of rational utility maximizing forward looking agents

Recommender System Software and techniques to suggest an item to a user mostly based on predicted rating and mainly applied on creative products distributed online

Tag User generated keyword to annotate an object

Valence Index of consumer evaluation of online reviews often expressed in terms of proportions of positive versus negative messages

Word of mouth Person to person distribution of information

List of Acronyms

ACM Association for Computing Machinery

AIC Akaike Information Criterion

AWE Approximate Weight of Evidence

BIC Bayesian Information Criterion

CAIC Consistent Akaike Information Criterion

CF Collaborative Filtering

DW Arthur De Vany and David Walls

EM Expectation Maximization

GMM Generalized Method of Moments

IMDB Internet Movie Database

IIA Independence of Irrelevant Alternatives

IO Industrial Organisation

KL Kullback-Leibler

L Likelihood

LL Log Likelihood

LCA Latent Class Analysis

LCLR Latent Class Logistic Regression

LCMM Latent Class Markov Model

LDA Latent Dirichlet Allocation

LSI latent Semantic Indexing

MPAA Motion Picture Association of America

MAE Mean Absolute Error

MSE Mean Squared Error

OLS Ordinary Least Squares

PG Parental Guidance

PLSI Probabilistic Latent Semantic Indexing

PLC Probabilistic Latent Class

PLC-CF Probabilistic Latent Class Collaborative Filtering

RMSE Root Mean Squared Error

RUM Random Utility Model

SQL Structured Query Language

SVD Singular Value Decomposition

UTC Coordinated Universal Time

WOM Word of Mouth

List of Movie Websites

Flixter: www.flixster.com Movie related social network site also providing movie information such as top box office and DVD rental. Users can form subgroups by connecting to other users, can rate movies and actors and receive movie showtimes and news.

Internet Movie Database: www.imdb.com Online database owned by Amazon gathering and exhibiting information on movies, television programs and games. They link movies to a number of characteristics such as plot, biography, cast and director. The information is sourced by users who can also add ratings that are aggregated into a movie score. The portal shows lists of top rated movies and TV shows, movie news and trailers. The database is the prime source of information for researchers adding movie features and making genre divisions.

MovieLens: www.movielens.org Recommender System created by GroupLens, a research lab of the Department of Computer Science and Engineering at the University of Minnesota. The site helps users find movies they like based on taste profiles inferred through their rating behaviour, using collaborative filtering algorithms. The MovieLens interface shows a number of movie lists, including top picks, recent releases, favourites from last year and new additions and maintains a user tagging system. The rating and tagging user data are made available, in an anonymised form, as open data for research purposes, making them one of the prime sources for research on recommender systems. It groups 4 databases, 100k, 1M, 10M and 20M referring to the number of ratings in the samples and successively released in the period 1998 till 2016. The early databases contains user information, the latter databases add tagging information.

Netflix: www.netflix.com Netflix is a private company supplying streaming media and video-on-demand and acting as content producer. Their initial business was that of DVD sales and rental. They maintained a personalized video-recommendation system based on ratings and reviews by their customers and organised an open competition for the best collaborative filtering algorithm based on a dataset of over 100.000.000 ratings provided by 480.189 users. That database was later used for scientific research.

Rotten Tomatoes: www.rottentomatoes.com Private company owned by Flixter, assembling reviews on movies and television programs. Critics scores are aggregated into the Tomatometer and public opinions taken from the user community are united in an audience score. They provide movie news and information on top box office, DVD and streaming movies.

List of non-movie Recommender Systems

del.icio.us Social bookmarking system where users can add freely chosen tags. Based on folksonomies, hot lists and recent pages are selected

Usenet Platform to exchange news organised in newsgroups

Jester Recommender system for jokes

Last.fm Music streaming website. Based on past choice, users are connected to taste groups

CiteULike Management service for references of scientific literature, combined with a filing system employing tags. Based on shared libraries, topic groups are formed.

Flickr Photo and video management, sharing and tagging service

List of Used Software

Latent Gold Commercial software sold by Statistical Innovations and developed mainly by J.Vermunt and J. Magidson. The package specialises uniquely in latent class analysis: clustering, factor analysis, regression and Markov models, for which it provided a Gui. Apart from that, it offers an Advanced/Syntax add-on. All user information can be found on www.statisticalinnovations.com.

R-package TM This is a software package in R developed by I. Feinerer for textmining purposes. It contains the commands to perform the traditional operations in natural language processing, such as tokenizing and stemming, and for that relies on earlier CRAN-packages such as NLP and Snowball. The software was used to generate a term-document matrix.
Commands can be found in https://cran.r-project.org/web/packages/tm/tm.pdf.

R-package Topic Models This is a software package in R developed by B. Grun and K. Hornik offering code to fit LDA models with the VEM algorithm and with Gibbs sampling. It needs input of a document-term matrix for which the code is provided by R-package TM. Commands can be found in
https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

VOSviewer: http://www.vosviewer.com Vosviewer is software, developed by L. Waltman and N.J. Van Eck, to create and visualise bibliometric networks based on co-citation, co-word occurrence or bibliographic coupling. The software was used to generate figures 1.1 and 1.2.

List of Figures

List of Tables