Estimating Substitution Using Text Embeddings: Evidence from the Film Industry

Gideon Moore

November 12, 2024

Abstract

Using text descriptions of films in conjunction with weekly box office receipts, I develop a novel model of characteristic-space competition in the film industry. By exploiting plausably exogeneous variation in film release windows, I identify the impact of competitor characteristics on film revenue. As films become more similar, the impact of competition increases. Due to the film industry's thin profit margins and high fixed costs, replacing a competitor in the 10th percentile of similarity with one in the 90th percentile can reduce profit by as much as 47%.

1 Introduction

Demand estimation methods are central to the economist toolkit, informing antitrust regulation, optimal taxation, and trade policy. However, these approaches are often confined to markets in which goods have quantifiable characteristics, such as storage in the market for hard drives or wattage in the market for lightbulbs. Markets with difficult-to-quantify characteristics should not receive less attention simply because they are difficult-to-quantify; though it is hard to put a number on how funny a film is, we understand that a studio monopolizing "funny" films is likely to be bad for consumers. Thus, it is important economists develop tools to estimate demand in markets where traditional methods struggle to perform well.

Product-level demand systems such as the Almost Ideal Demand System (AIDS) can estimate rich substitution patterns without any characteristic data by looking directly at substitution between each pair of goods i and j (Deaton and Muellbauer 1980). However, these methods often require onerous quantities of data on consumer behavior; for example, estimation of AIDS requires sufficient variation to estimate parameters quadratic in the number of goods. Moreover, even if goods i and j are very similar, these methods infer nothing about good j from the substitution patterns of good i. A corollary of these limitations is that these methods require goods to have *already entered the market*, since without sales data on good i it is impossible to say anything about its substitution patterns. If a firm wishes to estimate substitution for a new product, these methods will have little to offer.

I propose a new method for quantifying product characteristics using text descriptions. By embedding text descriptions of a product in characteristic space, econometricians can identify which products are most similar. Thus, they can predict which products compete most closely with others and estimate how the presence of one product impacts the sales of another. Unlike product-based demand estimation methods, this method can inform research on products which have not yet entered the market, and can fit rich substitution patterns while tuning relatively few parameters.

While this method has broad applicability, I use the film industry as a case study. Film is a natural laboratory for several reasons. First, films have a broad consumer base, such that readers have an intuitive sense whether my method "makes sense"–even if someone hasn't seen *Spider-Man*, they likely have a sense whether it is more similar to *Man of Steel* or *The Notebook*, and can use this lens to judge my results. Second, film characteristics are often difficult to fit into a traditional econometric framework: while we can easily measure the wattage of a lightbulb, it is much harder to put a number on how funny or violent a film is–and harder still given the lack of clean data. Finally, the film industry has specific market characteristics which make demand estimation using my method straightforward. Uniform pricing shuts down the endogeneous pricing channel of competition, and long production pipelines force studios to choose characteristics without perfect knowledge of their competitors.

Using data from BoxOfficeGuru and TheMovieDB, I estimate the impact of competitor simi-

larity on film revenue in North America for each weekend from 2000-2019. I identify substitution patterns which are plausible in both magnitude and direction. As a film becomes more similar to its competitor, it becomes more and more detrimental to the competitor's revenue. Replacing a competitor in the 10th percentile of similarity with one in the 90th percentile reduces box-office revenue by 3.3%. Given films' thin profit margins, this is roughly a 47% decline in profit.

2 Literature

This paper contributes to three strands of literature. First, I implement a novel use for text embeddings, innovating on the text-as-data literature in economics. Second, given that my demand estimation method is based on position in latent characteristic space, I am also closely connected to existing work on spatial competition. Finally, given my film case study, I also contribute to the literature on the economics of the film industry.

2.1 Text as Data

This paper's primary contribution is to the literature on text as data, which is well summarized in Gentzkow, Kelly, and Taddy (2019). The statistical analysis of text has a long history; to my knowledge, the earlist example is Mendenhall (1887) using word frequency to predict the author of a mysterious text. This is an early example of *bag of words* methods, which treat text as a collection of words rather than a structured object. Other types of bag of words methods include term frequency-inverse document frequency (TF-IDF), which downweights words that appear in many documents, and Latent Dirichlet Allocation (LDA), which uses a generative model to infer topics from word concurrences.

As argued in Kenter and de Rijke (2015), bag of words and string comparison models struggle when working with short texts. If one paragraph includes the word "wizardry" while another includes the word "sorcery," a bag of words model will not capture the similarity between the two paragraphs. In large enough documents, it is more likely the two words will appear in similar contexts, but this is difficult to rely on when the sample text is only one or two sentences.

When working with short snippets of text like film descriptions, it is more profitable to pursue *text embedding methods* as pioneered by Google's word2vec architecture (Mikolov et al. 2013). Rather than counting individual words, word2vec identifies similar words using co-occurances in a training sample, then maps these words to a lower-dimensional space. This allows for the identification of similar texts even when they share few or no words, since it is able to identify that synonyms like "wizardry" and "sorcery" map to the same concept. I use OpenAI's frontier GPT-3 embeddings, which embeds not only words but entire paragraphs in order to capture a broader context.

The most similar paper to mine is Compiani, Morozov, and Seiler (2023) which uses text characteristics to estimate demand for technology products on Amazon. However, rather than using more traditional logit substitution, my model can estimate substitution patterns non-linearly and non-monotonically with distance between two products. I also benchmark my substitution patterns against "real world" user preference data from the MovieLens data, confirming my estimates capture actual substitution patterns. Finally, rather than Euclidean distance, my similarity measure is based on the cosine between the embedding vectors; this measure hews more closely to the industry standard, is more efficient to compute at scale, and avoids curse-of-dimensionality issues which often arise when taking Euclidean distances in high-dimensional space.

2.2 Spatial Competition

Though this paper does not focus on *physical space*, I also contribute to the literature on spatial competition in the tradition of Hotelling (1929) and Salop (1979), where firms interact based on relative position in space. By projecting descriptions into characteristic space, I model high-dimensional competition between products.

My closest predecessor studying competition in embedding space is Magnolfi, McClure, and Sorensen (2024), who use t-Stochastic Triplet Embeddings to estimate competition in the cereal industry. I take heavy inspiration from their model, however substitute frontier text embedding methods from OpenAI for their triplet-based embeddings. Both they and I follow Pinkse, Slade, and Brett (2002), who first estimate product cross-elasticities as a function of distance in characteristic

space.

2.3 Economics of Film

The final strand of literature where I wish to highlight my contribution is on the economics of the film industry.

I benefit greatly from research documenting the uniform pricing puzzle (Orbach and Einav 2007; Gil and Hartmann 2009; Ho et al. 2018). This literature extensively documents movie theaters' use of uniform pricing across heterogeneous films. While each paper suggests potential explanations for this phenomenon, my findings are agnostic to the true cause of uniform pricing; I simply use it as a convenient assumption to estimate demand.

I contribute to the ongoing literature founded by Prag and Casavant (1994) of determinents of box office demand. The authors find well-rated and well-advertised films generally sell more tickets, and conditional on those two factors other observables are generally insignificant. Elberse and Eliashberg (2003) model both the supply and the demand for films, finding that theatres' decision to screen a film at all is an important factor for revenue. De Vany and Walls (1996), De Vany and Walls (1997), De Vany and Walls (1999), and De Vany and Walls (2004) study several factors in box office revenue, including word-of-mouth, film retirement, and the film production function. Ravid and Basuroy (2004) examine the impact of film *content* on performance, finding that violence increases film profitability while sexual content does not.

To the best of my knowledge, this paper is the first to examine the *interplay* between competing films: how much is my revenue impacted by the existence of box office competitors?

3 The Film Industry

I use a case study of the film industry to highlight the value of my method. Text is distinctively important for the film industry because film characteristics are otherwise difficult to quantify; for example, both *Batman Begins* and *Ant-Man* are big-budget action films about animal-based

superheros, but even a quick glance at the descriptions makes clear they are not close substitutes.[1] This is particularly important given films are an "experience good"–people rely on the description to inform their purchasing decision, as most people will not form opinions through repeat purchases. Recognizing this, we expect firms put more effort into the text descriptions of their films than comparable firms might in more traditional goods markets, as these decisions are distinctively important to customer decision-making. Finally, the market for films has two distinctive features which make it a strong candidate for demand estimation: long production pipelines and a tradition of uniform pricing.

3.1 Production Timelines

Film production timelines are both long and secretive. Since studios must purchase a script, hire a cast, and shoot all *before* promoting a film, it is difficult for a film to change its characteristics in response to its competitors' characteristics. To provide suggestive evidence of this inelasticity, I highlight the existance of "twin movies": pairs of films released in close proximity with very similar characteristics. The most famous examples, *A Bug's Life* and *Antz*, both released in fall of 1998 and feature early CGI animation of ants rebelling against oppression.

Twin movies are a persistent feature of the film industry: the Wikipedia page for the phenomenon lists nearly 300 pairs of twin movies. Studios are generally inclined to avoid releasing a twin movie due to fear of excess competition. In 2016, French director Xavier Giannoli released his film *Marguerite*, inspired by the true story of a New York socialite turned failed opera singer, in the same year as 20th Century Fox's *Florence Foster Jenkins* retelling the same story. Giannoli told the *Independent*: "For me, it was terrible...I work a lot as a writer to find completely original stories. I don't want the audience to have the feeling, 'oh, I saw that!'" (Mottram 2016).

Given filmmakers' expressed distaste for twin movies, why do they continue to exist? Studios are seemingly unable to change their films' characteristics in response to their competitors' charac-

[1] *Batman Begins*: "Driven by tragedy, billionaire Bruce Wayne dedicates his life to uncovering and defeating the corruption that plagues his home, Gotham City. Unable to work within the system, he instead creates a new identity, a symbol of fear for the criminal underworld." *Ant-Man*: "Armed with the astonishing ability to shrink in scale but increase in strength, master thief Scott Lang must embrace his inner-hero and help his mentor, Doctor Hank Pym, protect the secret behind his spectacular Ant-Man suit from a new generation of towering threats."

teristics; at the point DreamWorks learns of *A Bug's Life*, it is too late for them to adjust *Antz* to be more distinct. Thus, the demand estimation coefficients are identified–they do not capture the strategic interplay of positioning and substitutability, but instead the direct impact of competition on film sales.

Given the above, my demand model assumes characteristics are exogeneous. If instead they were chosen as equilibrium objects, then the coefficients on proximity are no longer causal: the impact of locating near a close competitor becomes confounded with latent demand characteristics driving the *choice* to locate near a close competitor. However, the existance of twin movies suggests this is not a major concern.

One threat to this strategy would be if firms lack flexibility on their films' *content*, but can move the film's release date *forward or backward in time*. If I discover I am releasing my children's movie opposite *Frozen*, I may wish to postpone my release date. This would threaten the assumption of exogeneous characteristics among competitors. I think this is unlikely to be a major concern for a couple of reasons. First, there are certain prime release dates which are likely to be more profitable than others; for example, releasing a family film in November is much better than doing so in January, as you can capture the holiday season. Thus, if a film targeted a holiday release, moving its release date to avoid competition would potentially be quite costly since other weekends face lower demand. Second, the continued existence of twin movies further suggests this is not a margin firms seem to adjust significantly on: if studios could easily move release dates to avoid competition, we would expect to see fewer twin movies than we do.

3.2 Uniform Pricing

An ongoing puzzle within the film industry is the existence of a uniform pricing standard: within a timeslot, cinemas generally charge the same price for all films regardless of excess demand. As a salient example, "opening night" showings of films often sell out, yet theaters do not raise prices. Similarly, films late in their run often have significant excess capacity, yet theaters do not lower prices.

There are many explanations for this phenomenon; I am agnostic about which of these is the

true explanation. What is important is that ticket price *is* uniform across films. This pricing anomaly supports two simplifying assumptions which do not hold water in most other settings.

First, I argue *films do not compete on price.* Under more flexible pricing, it is possible two films in close proximity would lower prices to compete for the same audience, raising the quantity sold overall. However, since prices are fixed I assume that consumer decisions are driven entirely by product characteristics and idiosyncratic taste.

Second, I *use revenue as a proxy for quantity.* As described in Bond et al. (2021), this assumption is often perilous since a variable markup means that revenue may not reflect true production. However, since prices are fixed, I can interpret revenue as simply the quantity of tickets sold multiplied by some scalar price.

4 Model

4.1 Construction

Consumers face a discrete choice problem: which of the films on offer this weekend should they see? Let film i's attractiveness at time t be given by some δ_{it}. Consumers receive idiosyncratic demand shocks for each film, and choose the option which maximizes their utility. Due to the uniform pricing scheme discussed above, price *does not* enter into the consumer's decision; choices are driven entirely by film quality and idiosyncratic taste.

Aggregating over all consumers, we can express the demand for film i in week t as a function ϕ of the film's appeal δ_{it} and the appeal of its competitors δ_{-it}. Again due to the uniform pricing scheme, we can use revenues as a proxy for quantity, since quantity is simply a constant scalar transformation of revenue. Finally, demand for films varies idiosyncratically by weekend; for example, there is much more demand for films on Christmas Day than on a random weekend in sunny July. Thus, I include a fixed effect α_t to capture week-by-week variation. I can therefore write the demand for film i in week t as:

$$\ln(q_{it}) = \phi(\delta_{it}, \delta_{-it}) + \alpha_t$$

A film's appeal is composed of two parts. First, a fixed effect α_i captures the film's quality. For example, we would expect *Star Wars* to have a high α_i to capture the fact it is a popular film in a vacuum separate from its competitors. Second, films have an age fixed effect $\lambda_{t-r(i)}$, where $r(i)$ represents film i's release date. As an experience good, demand for a film is driven heavily by novelty; thus, the same film will experience much less demand in its second week than in its first. While it is reasonable to expect these λ to decrease with time, I do not impose this; instead, verifying this is true will be a test of model fit later on. Finally, I include an idiosyncratic shock ξ_{it} to capture unobserved demand drivers. Thus, the film's appeal is given by:

$$\delta_{it} = \alpha_i + \lambda_{t-r(i)} + \xi_{it}$$

Putting these terms into the demand equation, we have:

$$\ln(q_{it}) = \phi(\alpha_i + \lambda_{t-r(i)} + \xi_{it}, \alpha_{-i} + \lambda_{t-r(-i)} + \xi_{-it}) + \alpha_t$$

How should the demand for film i depend on its competitors? I put forward three competitor traits which should impact demand for film i:

- *Competitor Quality*: If competitor j has a high α_j, it should draw more demand away from film i. For example, *The Lord of the Rings* is broadly a "better" film than *Willow,* despite being similar on paper; thus, i would prefer to compete against *Willow* rather than *The Lord of the Rings*.
- *Competitor Age*: If competitor j is more recent, it should draw more demand away from film i. That is, i should broadly prefer to release against films in their fifth week of showing than in their first week.
- *Competitor Similarity*: If competitor j is more similar to film i, it should draw away more demand. For example, if i is *Star Wars*, it is much more enthusiastic to release against *The Notebook* than against *Star Trek*.

Let d_{ij} be a measure of similarity between films i and j. In this exercise I use the cosine distance

of the embedded text descriptions; however, this could be any measure of similarity. There is no reason to expect similarity to enter linearly; instead, I consider some flexible function of distance f.

To capture the three attributes above while also letting d_{ij} enter flexibly, I propose the following functional form:

$$\ln(q_{it}) = \underbrace{\alpha_i + \lambda_{t-r(i)} + \xi_{it}}_{\delta_{it}} + \sum_{j \neq i} f(d_{ij}) \cdot \underbrace{(\alpha_j + \lambda_{t-r(j)} + \xi_{jt})}_{\delta_{jt}} + \alpha_t$$

Demand for i is thus a linear function of its own appeal δ_{it}, a weekend fixed effect α_t, and its competitors' appeals δ_{jt} weighted by some function f of distance d_{ij}.

4.2 Estimation

As written, the model above is difficult to estimate. α_i appears not only in its own demand, but in competitors' as well; thus, we cannot use simple demeaning to absorb these fixed effects. Moreover, the f function interacts multiplicatively with the δ_{jt} terms, making matrix inversion ineffective. With some rearranging, however, this model becomes more tractable:[2]

$$\begin{aligned}
\ln(q_{it}) = \alpha_i &+ \sum_{j \neq i} f(d_{ij}) \alpha_j \\
&+ \lambda_{t-r(i)} \left(1 + \sum_{j \neq i}^{r(j)=r(i)} f(d_{ij}) \right) + \sum_{r(k) \neq r(i)} \left(\lambda_{t-r(k)} \left(\sum_{j}^{r(j)=r(k)} f(d_{ij}) \right) \right) \\
&+ \alpha_t \\
&+ \xi_{it} + \sum_{j \neq i} f(d_{ij}) \xi_{jt} \qquad \Big\} \text{ Mean 0 given } f(d_{ij}) \perp \xi_{jt}
\end{aligned}$$

Thus, conditional on the values of $f(d_{ij})$, each film's log quantity is a linear function of it and its competitors' fixed effects and ages plus a fixed effect for the relevant weekend plus mean-zero noise. Thus, the λ and α coefficients are estimable using OLS.

[2] Gory algebraic detail of this rearrangement is available in appendix A.

As an example, consider what the data matrix will look like in the following market. Film 1 released this week. Film 2 released this week as well, while film 3 released last week. I omit time fixed effects in this example since we focus on a single period; similarly, I omit time subscripts on the quantities. In this case, we will have the following data matrix where values in "film" columns are used to estimate α values while "age" columns are used to estimate λ values:

ln(Quantity)	Film 1	Film 2	Film 3	Age 0	Age 1
$\ln(q_1)$	1	$f(d_{12})$	$f(d_{13})$	$1 + f(d_{12})$	$f(d_{13})$
$\ln(q_2)$	$f(d_{12})$	1	$f(d_{23})$	$1 + f(d_{12})$	$f(d_{23})$
$\ln(q_3)$	$f(d_{13})$	$f(d_{23})$	1	$f(d_{13}) + f(d_{23})$	1

Regressing quantity on the given Xs identifies the coefficients for α_1, α_2, α_3, λ_0, and λ_1 via the second through sixth columns, respectively. These results make some amount of intuitive sense; if $\ln(q_1)$ performs below expectations, one explanation is that α_2 is large–represented by a large coefficient on the (negative) $f(d_{12})$.

Note all the above is taking f as given. Following Magnolfi, McClure, and Sorensen (2024), let f be a cubic polynomial in distance; that is:

$$f(d_{ij}) = \gamma_0 + \gamma_1 d_{ij} + \gamma_2 d_{ij}^2 + \gamma_3 d_{ij}^3$$

This functional form makes few assumptions about the relationship between distance and substitutability, allowing for flexible sign, monotonicity, intercept, level, and concavity. Thus, if the final result *does* exhibit those properties, we know they are coming from the data rather than baked into the model.

Since the λ and α coefficients in the model above interact multiplicatively with the γ coefficients in f, we cannot estimate γ at the same time that we estimate λ and α linearly. Instead, I will choose a γ coefficient vector and estimate λ and α conditional on that γ vector using OLS. We know the resulting model minimizes mean squared error *in the set of models using this* γ. By adjusting the γ vector to minimize mean squared error of these optimized models, I search across the lower envelope and know my final model minimizes mean squared error over the whole $(\gamma, \alpha, \lambda)$ space.

Note this optimization is very expensive, as it requires estimating thousands of α coefficients for each interation as well as computing potentially tens of millions of values of $f(d_{ij})$. To speed estimation, I first optimize γ using data from 2000 alone (the first "quality" year of my data). Conditional on this optimized γ, I then estimate α and λ on all releases from 2000-2019 (omitting 2020 onward due to the negative impact of the COVID-19 pandemic). This allows me to estimate the full model in a reasonable amount of time. In the future, I could potentially run the optimization over the whole dataset using the research computing cluster; however, optimizing on my personal computer is already quite time-consuming even on the small sample.

Traditional standard error methods will struggle here. Since the analytic standard errors on the α and λ coefficients do not factor in the uncertainty from the γ estimation, they will understate the true uncertainty. I should use bootstrapping to estimate the true standard errors on all three types of coefficients; however, again this is computationally infeasible on my consumer laptop. If I were to run this on the research computing cluster, I could potentially estimate uncertainty on all of these coefficients via bootstrapping.

4.3 Application

In some ways this model is a strange fit for "characteristic-space" competition, in that it includes film fixed effects α_i. One of the common virtues of using a characteristic-space model is that it allows us to estimate demand for goods which are not yet in the market by arguing based on similar characteristics. However, the econometrician cannot estimate α_i for a film which has not yet released.

While this is a limitation, I do not believe it to be as costly as it initially appears. First, as part of estimating the model, I will identify the distribution of α values. Thus, even with no information on a specific α_i, it is possible for me to estimate the distribution of outcomes given the possible α_i values, computing both expected values and the variance around these values.

Moreover, the *additive separability* of the α and λ terms grants certain counterfactuals extra relevance. Note in the model above α and λ never interact multiplicatively. Thus, given competitor similarity d_{ij} and ages $\lambda_{t-r(k)}$, it is possible to perform comparative statics agnostic to the α values.

For example, I can ask how demand for film i will compare if it releases in week 1 rather than week 2 conditional on its competitors. Since the terms containing α_i will be identical in these two scenarios, they drop out and leave a concrete estimate of the change in quantity which is not dependent on the α_i values.

5 Data

5.1 BoxOfficeGuru

I scrape weekend box office receipts from 1997-2024 from BoxOfficeGuru.com, a website maintained by Gitesh Pandya. Pandya is a film consultant specializing in releasing Indian films for the North American market. For each weekend, the website lists the top 10-20 films by North American box office revenue.

While BoxOfficeGuru was getting its footing in the late 1990s, their reports were less consistent; thus, I begin my sample in 2000. To avoid the negative shock of the COVID-19 pandemic, I end my sample in 2019. When Pandya is on vacation the website is not updated, so I have a few missing weekends in my sample interspersed thoughout my period of interest. I drop these weekends from the sample, as I do not have a good way to impute the missing data. This yields a grand total of 901 weekends for analysis.

5.2 TheMovieDB

I retrieve film characteristics from the API of TheMovieDB, a user-generated database of films. Specifically, I clean film names and years retrieved from BoxOfficeGuru and query the search API for the top result released in the relevant year. For each film, I can thus retrieve the description, genres, and original release language. Example film descriptions are visible in table 1.

TheMovieDB also lists the primary language of each film. I limit only to films originally released in English to avoid both non-English film descriptions and differential sales trends among non-English speaking consumers. In total, my sample contains 2,970 films over the 20 year span matched

#	Title	May 17 - 19	May 10 - 12	% Chg.	Theaters	Weeks	AVG	Cumulative	Distributor
1	John Wick: Chapter 3	$ 56,818,067			3,850	1	$ 14,758	$ 56,818,067	Lionsgate
2	Avengers: Endgame	29,973,505	63,299,904	-52.6	4,220	4	7,103	771,368,375	Disney
3	Pokemon Detective Pikachu	25,108,159	54,365,242	-53.8	4,248	2	5,911	94,295,005	Warner Bros.
4	A Dog's Journey	8,030,085			3,267	1	2,458	8,030,085	Universal
5	The Hustle	6,139,638	13,007,709	-52.8	3,077	2	1,995	23,204,362	UA
6	The Intruder	4,017,808	7,190,325	-44.1	2,231	3	1,801	28,050,949	Sony
7	Long Shot	3,341,917	6,271,532	-46.7	2,110	3	1,584	25,664,963	Lionsgate
8	The Sun Is Also a Star	2,511,530			2,073	1	1,212	2,511,530	Warner Bros.
9	Poms	2,180,698	5,361,937	-59.3	2,750	2	793	10,110,890	STX
10	Uglydolls	1,779,617	4,147,092	-57.1	2,030	3	877	17,433,285	STX
11	Breakthrough	1,081,480	2,575,263	-58.0	1,375	5	787	39,012,955	Fox
12	The Curse of La Llorona	890,020	1,851,722	-51.9	651	5	1,367	53,005,382	Warner Bros.
13	Captain Marvel	735,998	1,846,396	-60.1	726	11	1,014	425,152,517	Disney
14	Tolkien	731,445	2,200,537	-66.8	1,501	2	487	3,768,184	Fox Searchlight
15	Shazam!	660,710	1,033,186	-36.1	536	7	1,233	138,067,613	Warner Bros.
16	De De Pyaar De	425,934			104	1	4,096	425,934	Yash Raj
17	Dumbo	309,302	744,251	-58.4	415	8	745	111,521,409	Disney
18	Little	279,785	684,050	-59.1	314	6	891	40,195,795	Universal
19	The Biggest Little Farm	276,446	110,492	150.2	45	2	6,143	413,154	Neon
20	The White Crow	231,379	149,648	54.6	136	4	1,701	702,966	Sony Classics

Figure 1: BoxOfficeGuru Page for May 17-19, 2019

Film	Description
Frozen	Young princess Anna of Arendelle dreams about finding true love at her sister Elsa's coronation. Fate takes her on a dangerous journey in an attempt to end the eternal winter that has fallen over the kingdom. She's accompanied by ice delivery man Kristoff, his reindeer Sven, and snowman Olaf. On an adventure where she will find out what friendship, courage, family, and true love really means.
The Notebook	An epic love story centered around an older man who reads aloud to a woman with Alzheimer's. From a faded notebook, the old man's words bring to life the story about a couple who is separated by World War II, and is then passionately reunited, seven years later, after they have taken different paths.
Avengers: Endgame	After the devastating events of Avengers: Infinity War, the universe is in ruins due to the efforts of the Mad Titan, Thanos. With the help of remaining allies, the Avengers must assemble once more in order to undo Thanos' actions and restore order to the universe once and for all, no matter what consequences may be in store.
A Bug's Life	On behalf of "oppressed bugs everywhere," an inventive ant named Flik hires a troupe of warrior bugs to defend his bustling colony from a horde of freeloading grasshoppers led by the evil-minded Hopper.

Table 1: Example Film Descriptions from TheMovieDB

between BoxOfficeGuru and TheMovieDB.

6 Text Embeddings

I use OpenAI's embedding model `text-embedding-3-small` API to embed film descriptions. This produces a 1,536-dimensional vector for each film description. The OpenAI embedding procedure trains a transformer to predict adjacent texts based on a large corpus, such that vectors with similar cosine similarity are likely to have similar next words (Neelakantan et al. 2022; Kusupati et al. 2022). These embeddings are quite cheap; the bill for this project was less than $1. Thus, this method should be accessible to researchers with any level of resources even for projects of significant scale.

I prefer the OpenAI embeddings to other text-as-data methods for a few reasons. First, film descriptions are quite terse; most are around a paragraph, with a few as short as one or two sentences. Thus, traditional bag of words methods are likely to suffer, as word or bigram counts would be remarkably sparse. Because the OpenAI embedding model is trained on an enormous

Baseline Film	Nearest Neighbor #1	Nearest Neighbor #2
Frozen	*Frozen II*	*Mirror Mirror*
The Notebook	*Message in a Bottle*	*The Longest Ride*
Avengers: Endgame	*Avengers: Infinity War*	*Captain America: Civil War*
A Bug's Life	*The Ant Bully*	*Antz*

Table 2: Nearest Neighbors to Baseline Films by Cosine Similarity

corpus of text, it can infer when descriptions are similar even when they share *no* words–for example, a film about "wizardry" would be similar to one about "sorcery," which would not be true under a bag of words model (Brown et al. 2020).

Second, BERT embeddings tend to function at the sentence level (and word2vec embeddings function only at the word level), while the OpenAI embeddings are generally more context-aware. Given the film descriptions tend to be longer than a single sentence, I believe the OpenAI embeddings are likely to be more able to capture the full context of the film description.

6.1 Quality Check: Genre

As a check for the quality of the embeddings, I perform t-Stochastic Neighbor Embedding (t-SNE) on the embeddings to reduce them to two dimensions and emphasize clustering. For each film, I identify the first genre listed on the TheMovieDB page. When running t-SNE I limit to only the four most popular genres (action, drama, comedy, and horror) to keep the figure legible. Embeddings do cluster by genre, as visible in figure 2. Generally we see comedies visible on the left, while actions are on the right and horrors are at the top (admittedly, the nebulous "drama" genre is more diffuse).

6.2 Quality Check: Nearest Neighbors

Nearest neighbors to some popular films are visible in table 2. Just by inspection, it is clear this similarity score is capturing *something*: the closest films to *The Notebook* are other Nicholas Sparks book adaptations, and the closest films to *A Bug's Life* are *The Ant Bully* and *Antz*. Moreover,

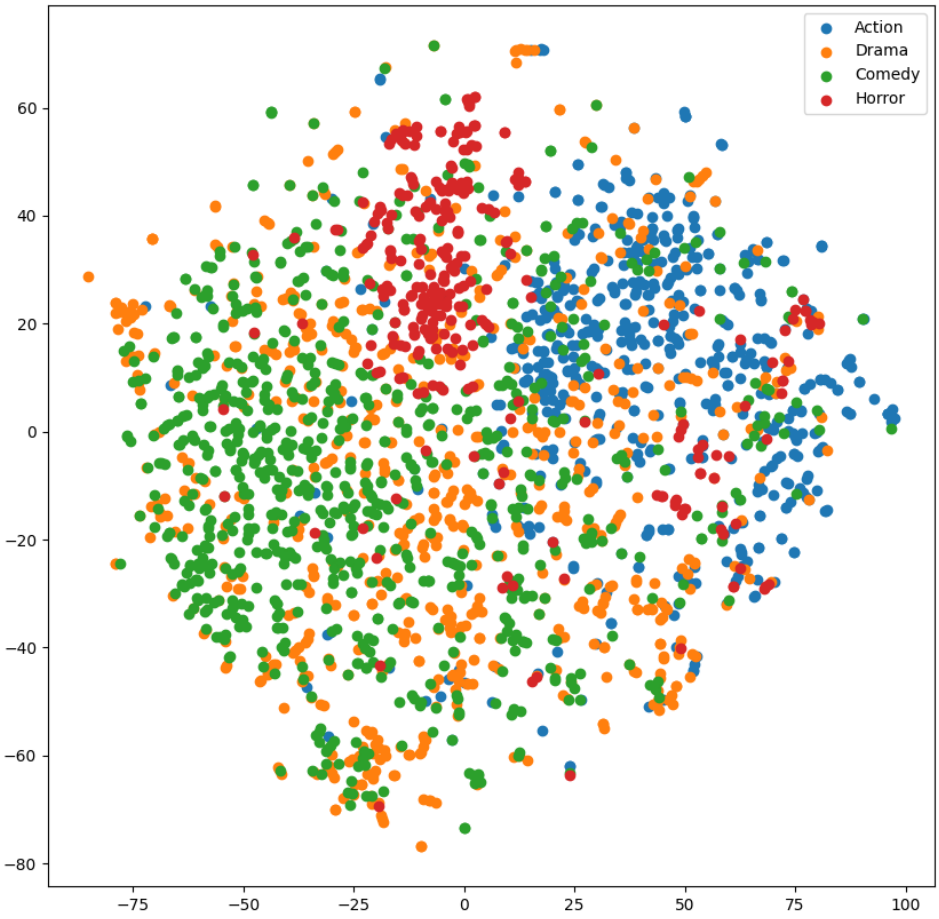

Figure 2: t-SNE of OpenAI Embeddings by Genre

the closest neighbors to *Frozen* and *Avengers: Endgame* are in the same franchise: *Frozen II* and *Avengers: Infinity War*, respectively. Thus, I feel comfortable using these embeddings to capture similarity between films.

7 Results

7.1 Age

First, I examine the age coefficients λ. The film in my sample with the longest run is *My Big Fat Greek Wedding,* which ran for 43 weeks. Estimating a weekly fixed effect for all 43 possible ages would absorb much of my power and create colinearity issues, as few films reach this age. Instead, I topcode age. Initially I ran the model with a topcode of 26 weeks; however, point estimates converged to be constant after only 9 while the standard errors become quite large. Given this inflection point, I impose a topcode of 9 weeks.

The estimated λ coefficients are visible in figure 3. As hypothesized, the age fixed effects decline with time such that new films are more popular than old ones. Recall that the dependent variable is *log* quantity; thus, the linear decline we observe in the λ coefficients implies an exponential decline in sales, as is expected in the film industry. Again, the model does not impose either the sign or the shape of the λ coefficients; the fact that they are monotonically decreasing at a linear rate suggests my model is well specified.

7.2 Quality Distribution

Now I check the distribution of α coefficients, visible in figure 4. The distribution is roughly symmetrical with a mean of 0.5. Again recall the log scale of the dependent variable; thus, this distribution implies that the distribution of quality in *levels* is quite skew, again as expected in the film industry.

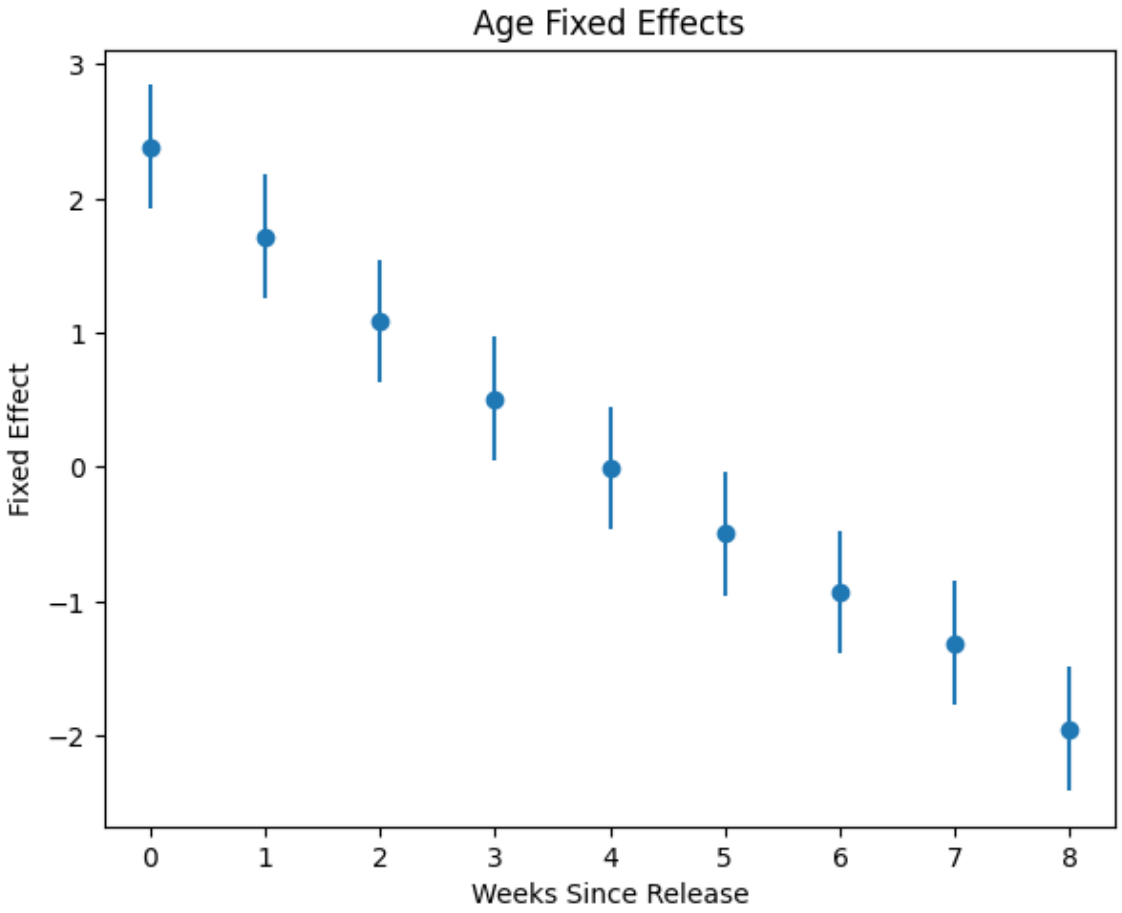

Figure 3: Age Coefficients

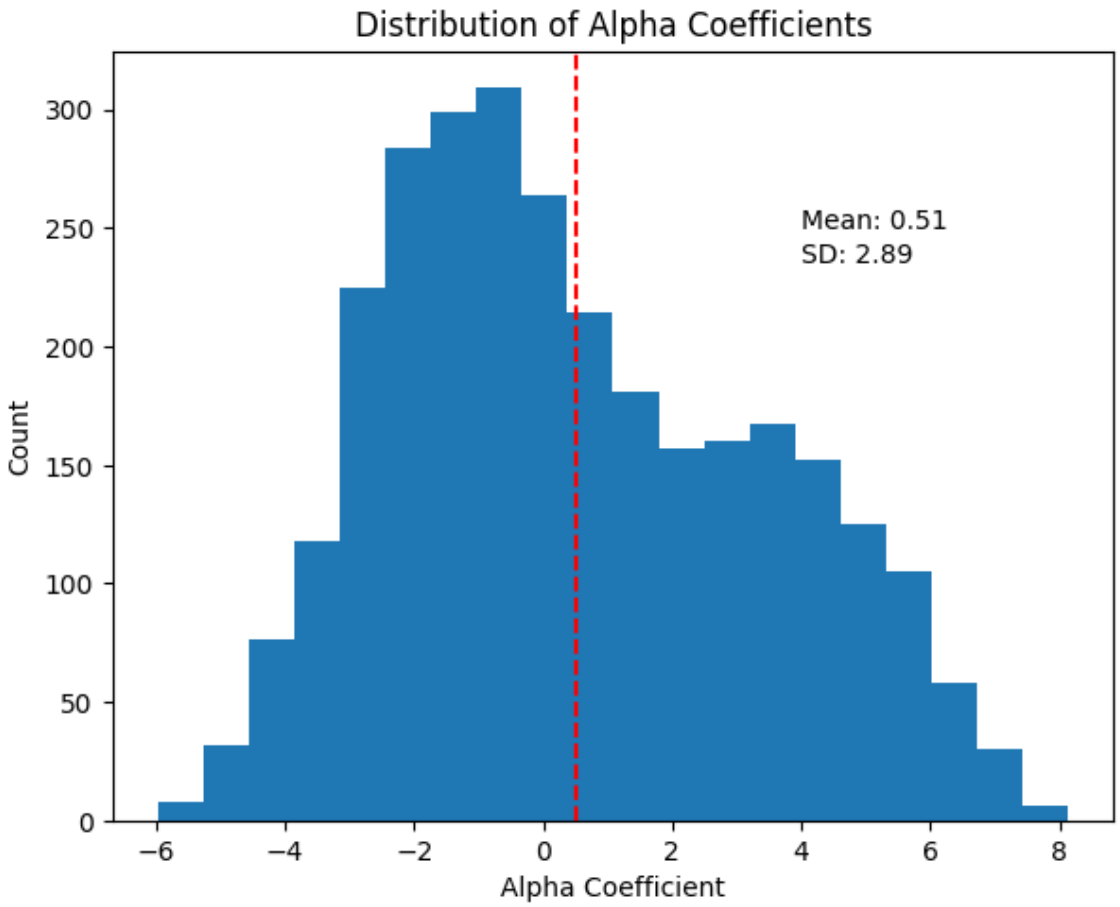

Figure 4: Distribution of α Coefficients

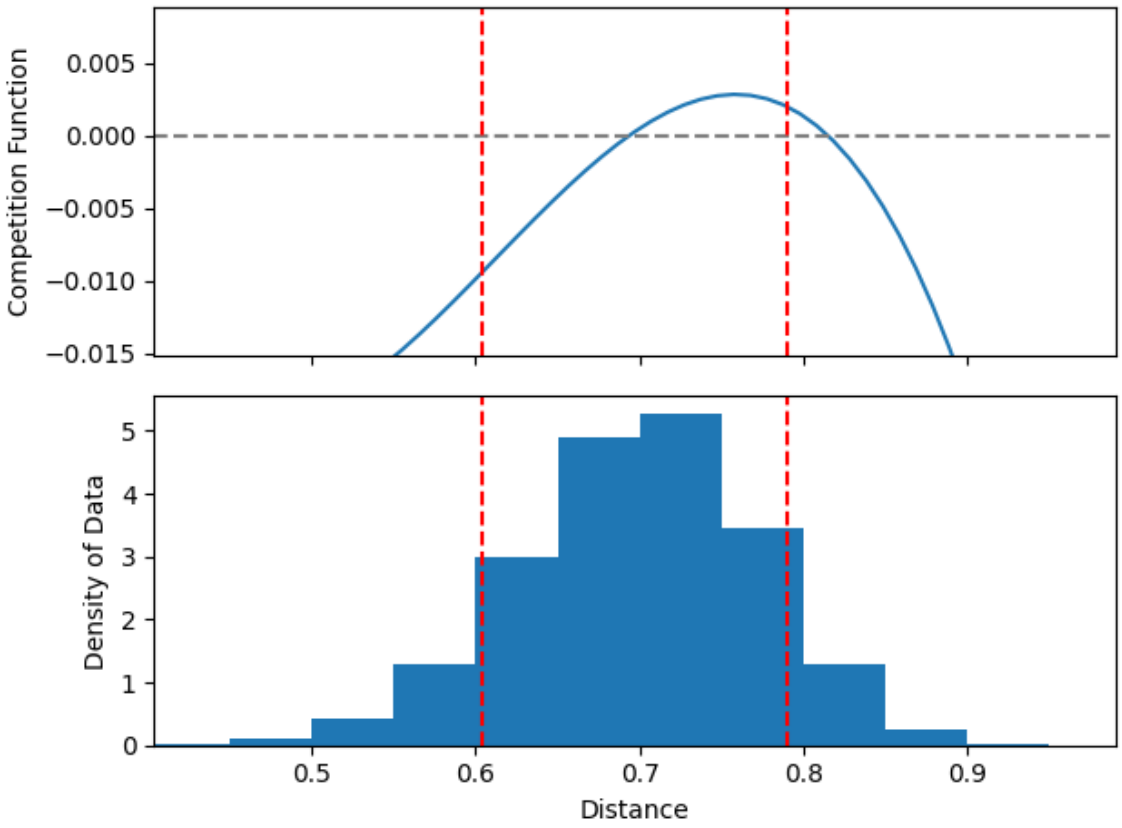

Figure 5: Influence by Distance

7.3 Elasticity

The impact of competition f is visible in figure 5. The curve is plotted over the density of the data; thus, the area over the top of the histogram should be most precisely estimated, while the edges are less precise. I have included the 10th and 90th percentiles of the data for reference; I would suggest focusing on the pattern within this range.

This curve exhibits a number of desirable characteristics on the support of the data:

- It is negative for similar films, suggesting that additional competition reduces a film's revenue.
- It is (roughly) monotonic and attenuating with distance, suggesting that as films become more distant their competitive impact declines.
- It approaches zero as distance becomes large, suggesting that films which are very dissimilar have no impact on each other.
- It is concave, suggesting that competition is most important when films are similar but less so when they are dissimilar.

The curve becomes positive for a portion of the domain. This may be an artifact of trying to fit a cubic to a function which is actually flat in this region since competition has ceased to matter (and thus this is actually 0). However, I can think of a couple reasons this phenomenon might be real:

- If two films are dissimilar enough, they may be complements rather than substitutes. For example, if a teenage child wants to see *Friday the 13th*, parents may drop off the child and take the younger sibling to see *Frozen*, when they would not have seen *Frozen* otherwise.
- Theater crowding may push people to see films they would not have otherwise seen. If a theater is showing *Friday the 13th,* this may crowd out an extra showing of *The Lego Movie*. This reduced supply of *The Lego Movie* drives people to see *Frozen* instead.

It is difficult to interpret the value of this function in a vacuum since it always interacts with δ_{it}. However, using the α and λ estimates above, I can compute some examples to help benchmark

the magnitude of these coefficients.

We know the average film's α_i is roughly 0.5. Similarly, the $\lambda_{r(i)-t}$ in the week of a film's premier is roughly 2.5. Thus, the average film's appeal δ_{it} in its first week is 3.

Suppose film i is attempting to estimate the impact of competitor j's premier, anticipating j is an average film. If j is in the 90th percentile of similarity to i, then i's log quantity would be reduced by $f(d_{ij}) \cdot \delta_{jt} = -0.01 \cdot (0.5 + 2.5) = -0.03$, or roughly 3%. On the other hand, if j is in the 10th percentile of similarity, i's log quantity would instead by *increased* by $f(d_{ij}) \cdot \delta_{jt} = 0.001 \cdot (0.5 + 2.5) = 0.003$, or roughly 0.3%. Thus, the presence of the wrong competitor relative to the right one could reasonably lower films' box office revenues by 3.3%.

To put this magnitude in context, Follows (2016b) and Follows (2016a) estimates that box office receipt makes up 42% of a film's revenue. Further (acknowledging the difficulties of managing so-called "Hollywood accounting"), he argues that the average film earns a profit of roughly 3.7%. If a film experiences a 3.3% decline in 42% of its revenue, its total revenue falls by 1.3%. The author documents that the vast majority of a film's cost is fixed rather than variable; thus, we can assume cost does not change given the presence of competitors. These figures together imply the profit margin in the presence of the closer competitor is only 1.95%, a 47.3% reduction in profitability.[3]

To be even more concrete, let us consider specific films. Ignoring the α coefficients for each of the following films, suppose Disney were releasing *Frozen* opposite gory historical action flick *Medieval*: a film in the bottom decile of cosine similarity to *Frozen*. If Disney were instead facing children's fantasy romp *Inkheart*–a film in the top decile of similarity to *Frozen*–they would expect a reduction in revenue similar to the one computed above as *Inkheart* pulls away customers *Medieval* would not.

8 Benchmarking: Collaborative Filtering

A skeptic might argue the patterns I see are merely coincidental: why would we believe films with similar descriptions necessarily appeal to the same audience? In this section, I use real film reviews

[3] $\frac{R-C}{C} = 0.033 \implies R = 1.033C \implies 0.987R = 1.0195C \implies \frac{0.987R-C}{C} = 0.0195 \implies \frac{0.037-0.0195}{0.037} = 47.3\%$

to validate my text-based model, showing that films with similar descriptions indeed appeal to similar consumers.

"Recommendation engines" are a common tool in the tech world used to predict what products a user will like. We can think of this project as harnessing a new type of recommendation engine, identifying films which are likely to be substitutable based on their descriptions. Algorithms in this space are broadly divided into two categories: collaborative filtering and content-based filtering. The text embeddings used above are a form of content-based filtering: using the "content" of the film (its text description) to predict its appeal. However, one way to benchmark the quality of this approach is to compare against a *collaborative* filtering algorithm. Unlike content-based filtering, collaborative filtering doesn't include any information about the content of a film; instead, it uses the preferences of other users to predict what a user will like. For example, if I like *The Dark Knight*, and most people who like *The Dark Knight* also like *Inception*, a collaborative filtering algorithm would predict I would like *Inception*.

Harper and Konstan (2016) provide a public-use dataset of real film reviews ("MovieLens"); it contains more than 25 million 1-to-5 star reviews of films by users. Moreover, films are tagged with their TheMovieDB ID; that is, the same ID I use to collect film descriptions. Thus, for any given pair of films, I can compute similarity under both the text-based and collaborative filtering methods. If the two measures are correlated, I can be more confident that the patterns I see in the text-based model are not coincidental. I validate my text-based model by comparing it against two frontier collaborative filtering algorithms: UV decomposition and topic-specific PageRank.

8.1 UV Decomposition

I implement UV-decomposition for matrix completion as described in Leskovec, Rajaraman, and Ullman (2020).[4] Suppose I have an $n \times m$ matrix of film reviews, with n users and m reviews. However, the matrix is generally sparse; most users have not reviewed most films. UV-decomposition decomposes this matrix into two thin matrices U and V such that UV' approximates the original

[4]The algorithm described in the textbook optimizes element-wise, which is quite slow. Based on my own vector calculus visible in appendix B, I've implemented *matrix-wide* optimization, which seems to work as well, but I have not seen a reference to this method elsewhere.

matrix. This is a cousin of singular value decomposition, however it relaxes the orthogonality constraint common in those approaches. By minimizing the sum of squared errors among projections in the *observed* entries of the matrix, we hope to project the *unobserved* entries of the matrix as well.

Note that UV decomposition relies on a contraction mapping: for a fixed V matrix, we choose the U matrix which minimizes the sum of squared errors. We then fix that U matrix and choose an optimal V matrix, iterating this process until convergence. This procedure is guaranteed to converge, though it may converge to a local minimum. Thus, we generally initialize the U and V matrices with random values, then run the algorithm multiple times to ensure we find something approximating the global minimum.

With this decomposition in hand, it is straightforward to assess whether two films are "similar": we simply compute the cosine similarity of the projected reviews. If two films have similar projected reviews, we can infer they appeal to similar audiences, and so are likely substitutable. The virtue of UV-decomposition is it works *even when the films have no reviewers in common*; that is, it can infer similarity between films which have never been reviewed by the same person.

Rather than blindly trusting the decomposition, we can validate it by comparing predicted reviews against actual reviews. I hold out 5% of the reviews as a test set, then train on the remaining 95%. If the predicted reviews for this held-out set are highly correlated with the actual reviews, we can be more confident in the UV decomposition.

As we might expect, the UV decomposition is better at predicting reviews for films with more reviews. To assess the size of this effect, I regress the true review on the predicted review for various thresholds of review count. If the threshold is low, we're including films with fewer reviews in our prediction set; if the threshold is high, we're looking only at hits like *Star Wars.* The results of these regressions are visible in figure 6.

For any value of the review count threshold, the coefficient on predicted reviews is positive and significant. This suggests that the UV decomposition is quite predictive of a film's appeal. Moreover, the coefficient is increasing in the review count threshold; that is, the UV decomposition is better at predicting reviews for films with more reviews. This is as expected; the more data we

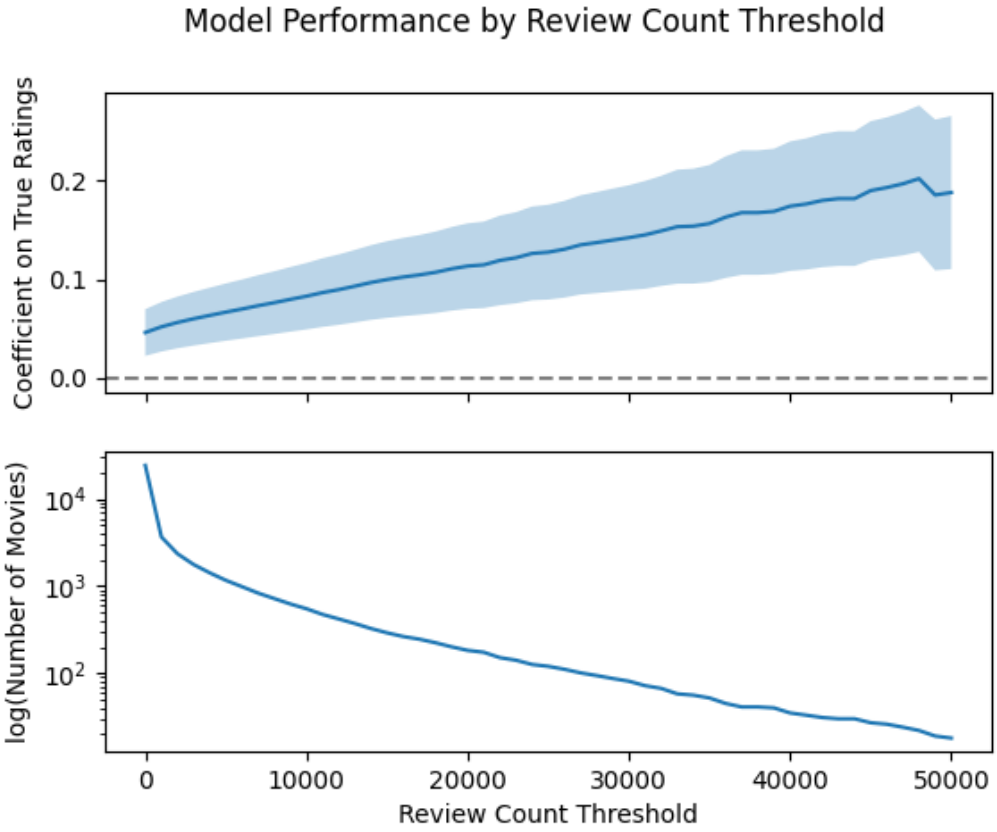

Figure 6: Actual vs. Predicted Reviews by Number of Reviews

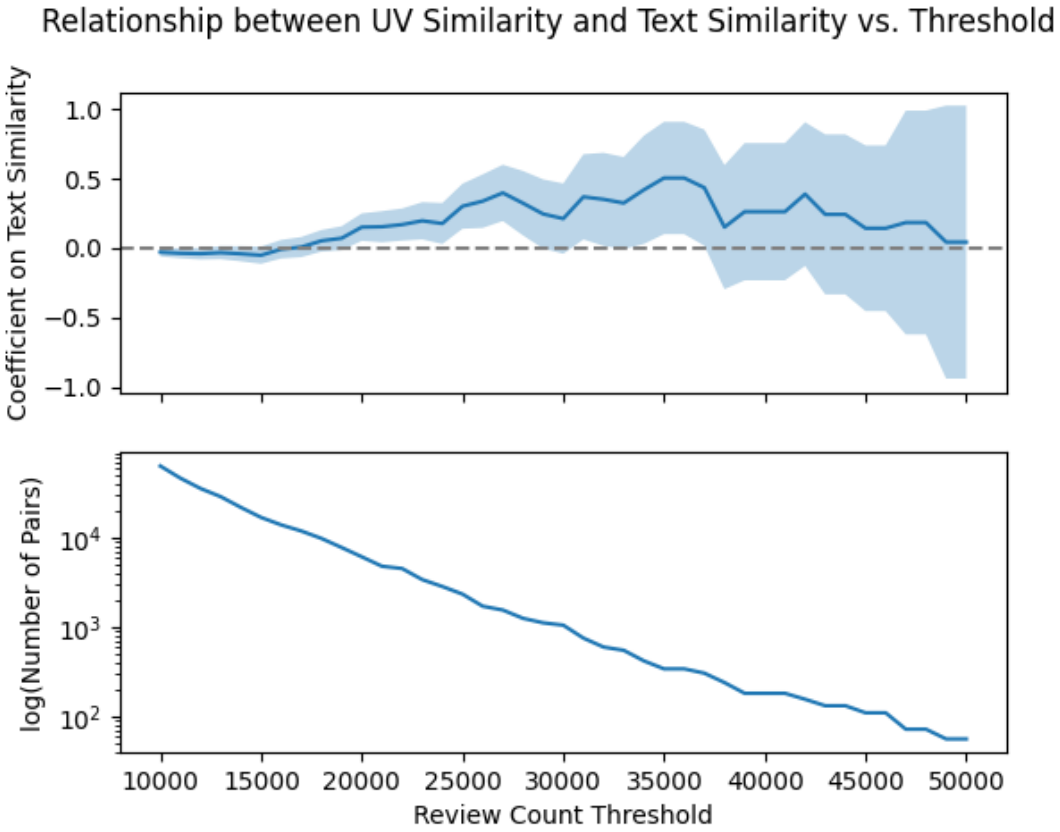

Figure 7: Coefficient of UV Decomposition on Text Similarity by Threshold

have on a film, the better we can predict its appeal.

These results affirm that the collaborative filtering estimates are likely to be quite predictive of a film's appeal. Thus, if my text-based model identifies the same pairs of films as substitutable, I can be more confident that the patterns I see reflect consumers' true preferences.

Assured that the UV decomposition results are valid, I attempt to predict the cosine similarity of the UV decomposition's predicted reviews using the cosine similarity of the text embeddings. Again we expect the quality of these predictions to vary with the volume of the data; thus, I plot the coefficient of this regression for various thresholds of review count as above. The results of these regressions are visible in figure 7.

For films with relatively few reviews, the coefficient on text similarity is insignificant; this is to be expected, as for these films UV decomposition struggles to capture anything of meaning. However, as the review count threshold increases, the coefficient on text similarity becomes increasingly positive and significant. This suggests that as UV decomposition becomes more accurate, it looks increasingly like text-based cosine similarity. This is a strong validation of my text-based model; the patterns I see in the text embeddings are not coincidental, but reflect consumers' true preferences. The quality of the match begins to decline when the review threshold becomes too high; however, this is not a fault with my method but with the decomposition, as at these high standards the number of films with sufficient views falls to less than 10, producing fewer than 100 pairs of films.

8.2 Topic-Specific PageRank

As a second validation, I harness the bipartite network nature of the reviews data. Using topic-specific PageRank (TSPR) as implemented by Sajani (2023), I compute a second review-based similarity between each pair of films. PageRank is an algorithm for detecting the most important nodes in a network; it was originally developed by Google to rank webpages. The algorithm works by randomly walking through the network, with a probability of jumping to a random node at each step. The importance of a node is the proportion of time the random walker spends at that node. TSPR is an extension of the algorithm which always "jumps" to a specific set of nodes rather than a random node.

For my application, I implement TSPR as follows. To find films similar to specific film x:

- The algorithm randomly walks over the bipartite network of users and films.
- At each step, the algorithm has a 15% chance of jumping back to x.
- If the algorithm does not jump to x, it takes a random step to a neighbor of its current node.
- Random draws are weighted by the number of stars in the review; e.g., a "5-star" review is five times as likely to be drawn as a "1-star" review.

Running this algorithm until convergence provides a measure of how central any given film y is

relative to film x. However, we are still not quite done, as some films are just *inherently* central; if everyone likes *Forrest Gump,* then it's not particularly informative that fans of film x *also* like *Forrest Gump*. Thus, I divide the TSPR score of y by y's *non*-topic specific PageRank score. This gives a measure of how much more central y is to x than it is to the average film. Given this is a ratio, it has propensity to explode if y's centrality is very small; thus, I prefer to take the log of this ratio for my analysis.

Once I have each film's TSPR score for each other film, I can compare these scores to the cosine similarity of the text embeddings. Note that the TSPR scores are not symmetric; that is, the TSPR score of film x to film y is not necessarily the same as the TSPR score of film y to film x. Thus, for each pair of films, I actually observe *two* measures of similarity; one from x to y and another from y to x; these pairs have the same text-based cosine similarity, but different TSPR values.

A common weakness of collaborative filtering algorithms like TSPR is that they require significant data before they become effective. Thus, I compute TSPR values only for films which receive at least 10,000 reviews in the MovieLens data. This leaves me with 258 films, or 66,564 pairs of films. To assess the importance of this margin, I then plot the coefficient of regressing TSPR on text similarity for thresholds 10,000 to 50,000. I also plot the number of film pairs included in the sample for each threshold value, to assess the trade-off of "more ratings per film" vs. "more films." The results of this plot are visible in figure 8.

This plot shows that the coefficient on text similarity for predicting TSPR is consistently and significantly positive for nearly all thresholds. Thus, the content of the text embeddings *does* seem to have predictive power for which films are similar based on revealed preference; that is, the patterns I see in the text-based model are not coincidental and reflect consumers' true preferences.

Moreover, while the number of pairs included remains large enough to have useful confidence intervals, the coefficient is increasing in the threshold value. If we think TSPR increases in quality with data, this suggests that as TSPR becomes higher quality, it looks increasingly like text-based cosine similarity. This exercise highlights an additional strength of my cosine similarity measure; while collaborative filtering needs tens of thousands of reviews per film to generate useful predictions, the embeddings-based method can generate useful predictions with only a few sentences of text.

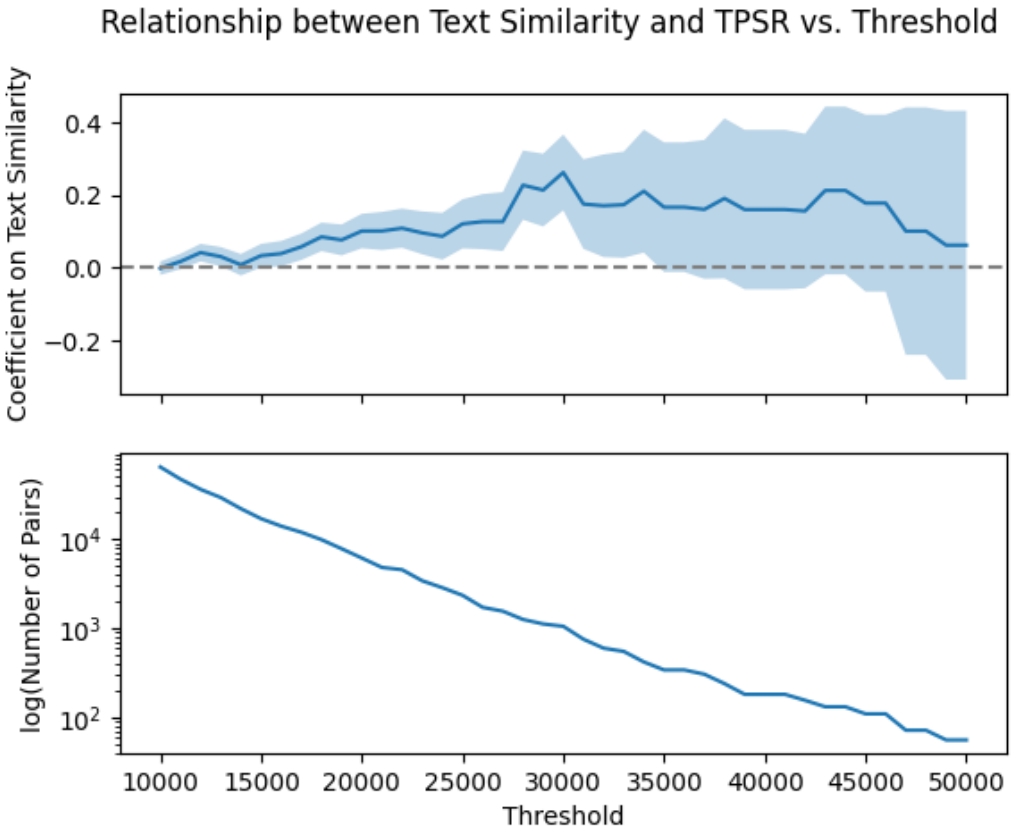

Figure 8: Coefficient of log(TSPR) on Text Similarity by Threshold

9 Conclusion

In summary, I have developed a new method for estimating competition in characteristic space. By developing a new model of competition in the film industry, I can harness boilerplate OpenAI embeddings to estimate impact of competition on a film's box office revenue.

Two frontier collaborative filtering algorithms, UV decomposition and topic-specific PageRank, confirm the validity of my text similarity measure. I find that the patterns I see in the text-based model are not coincidental, but reflect consumers' true revealed preferences.

Films with similar descriptions are indeed substitutes; moreover, substitutability falls as distance in characteristic space increases. My magnitudes are large but reasonable; changing one competitor film from the bottom decile of similarity to the top decile can reduce revenue by 3.3%, decreasing total profitability by 47%.

Acknowledgments

Thank you to Liran Einav and Shoshana Vasserman for their support both in and after Industrial Organization III. Thanks also to Janet Stefanov, Nick Scott-Hearn, and Lauren Harris for their feedback on this project, as well as the rest of the 2024 Industrial Organization III class.

References

Bond, Steve et al. (July 1, 2021). "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data". In: *Journal of Monetary Economics* 121, pp. 1–14. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2021.05.004. URL: https://www.sciencedirect.com/science/article/pii/S0304393221000544 (visited on 06/09/2024).

Brown, Tom B. et al. (May 28, 2020). *Language Models Are Few-Shot Learners*. arXiv.org. URL: https://arxiv.org/abs/2005.14165v4 (visited on 09/18/2024).

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler (Sept. 30, 2023). *Demand Estimation with Text and Image Data*. DOI: 10.2139/ssrn.4588941. URL: https://papers.ssrn.com/abstract=4588941 (visited on 03/06/2024). Pre-published.

De Vany, Arthur S. and W. David Walls (1996). "Bose-Einstein Dynamics and Adaptive Contracting in the Motion Picture Industry". In: *The Economic Journal* 106.439, pp. 1493–1514. ISSN: 0013-0133. DOI: 10.2307/2235197. JSTOR: 2235197. URL: https://www.jstor.org/stable/2235197 (visited on 06/09/2024).

— (1997). "The Market for Motion Pictures: Rank, Revenue, and Survival". In: *Economic Inquiry* 35.4, pp. 783–797. ISSN: 1465-7295. DOI: 10.1111/j.1465-7295.1997.tb01964.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1465-7295.1997.tb01964.x (visited on 06/09/2024).

— (Nov. 1, 1999). "Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office?" In: *Journal of Cultural Economics* 23.4, pp. 285–318. ISSN: 1573-6997. DOI: 10.1023/A:1007608125988. URL: https://doi.org/10.1023/A:1007608125988 (visited on 06/09/2024).

— (2004). "Motion Picture Profit, the Stable Paretian Hypothesis, and the Curse of the Superstar". In: *Journal of Economic Dynamics and Control* 28.6, pp. 1035–1057. ISSN: 0165-1889. URL: https://econpapers.repec.org/article/eeedyncon/v_3a28_3ay_3a2004_3ai_3a6_3ap_3a1035-1057.htm (visited on 06/09/2024).

Deaton, Angus and John Muellbauer (1980). "An Almost Ideal Demand System". In: *THE AMERICAN ECONOMIC REVIEW* 70.3.

Elberse, Anita and Jehoshua Eliashberg (Sept. 1, 2003). "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures". In: *Marketing Science* 22, pp. 329–354. DOI: 10.1287/mksc.22.3.329.17740.

Follows, Stephen (July 18, 2016a). *How Films Make Money Pt2: $30m-$100m Movies*. Stephen Follows. URL: https://stephenfollows.com/films-make-money-pt2-30m-100m-movies/ (visited on 09/20/2024).

Follows, Stephen (July 10, 2016b). *How Movies Make Money: $100m+ Hollywood Blockbusters*. Stephen Follows. URL: https://stephenfollows.com/how-movies-make-money-hollywood-blockbusters/ (visited on 09/17/2024).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (Sept. 2019). "Text as Data". In: *Journal of Economic Literature* 57.3, pp. 535–574. ISSN: 0022-0515. DOI: 10.1257/jel.20181020. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20181020 (visited on 06/07/2024).

Gil, Ricard and Wesley R. Hartmann (2009). "Empirical Analysis of Metering Price Discrimination: Evidence from Concession Sales at Movie Theaters". In: *Marketing Science* 28.6, pp. 1046–1062. ISSN: 0732-2399. JSTOR: 23884296. URL: https://www.jstor.org/stable/23884296 (visited on 06/08/2024).

Harper, F. Maxwell and Joseph A. Konstan (Jan. 7, 2016). "The MovieLens Datasets: History and Context". In: *ACM Transactions on Interactive Intelligent Systems* 5.4, pp. 1–19. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/2827872. URL: https://dl.acm.org/doi/10.1145/2827872 (visited on 09/05/2024).

Ho, Jason Y.C. et al. (June 1, 2018). "An Empirical Study of Uniform and Differential Pricing in the Movie Theatrical Market". In: *Journal of Marketing Research* 55.3, pp. 414–431. ISSN: 0022-2437. DOI: 10.1509/jmr.14.0632. URL: https://doi.org/10.1509/jmr.14.0632 (visited on 06/08/2024).

Hotelling, Harold (1929). "Stability in Competition". In: *The Economic Journal* 39.153, pp. 41–57. ISSN: 0013-0133. DOI: 10.2307/2224214. JSTOR: 2224214. URL: https://www.jstor.org/stable/2224214 (visited on 06/08/2024).

Kenter, Tom and Maarten de Rijke (Oct. 17, 2015). "Short Text Similarity with Word Embeddings". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. New York, NY, USA: Association for Computing Machinery, pp. 1411–1420. ISBN: 978-1-4503-3794-6. DOI: 10.1145/2806416.2806475. URL: https://dl.acm.org/doi/10.1145/2806416.2806475 (visited on 09/17/2024).

Kusupati, Aditya et al. (May 26, 2022). *Matryoshka Representation Learning*. arXiv.org. URL: https://arxiv.org/abs/2205.13147v4 (visited on 06/09/2024).

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman (Feb. 13, 2020). "Recommendation Systems". In: *Mining of Massive Datasets*. 3rd edition. New York, NY: Cambridge University Press, pp. 319–353. ISBN: 978-1-108-47634-8.

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen (2024). "Triplet Embeddings for Demand Estimation". In: *AEJ: Microeconomics* Forthcoming. ISSN: 1945-7669.

Mendenhall, T. C. (1887). "The Characteristic Curves of Composition". In: *Science* 9.214, pp. 237–249. ISSN: 0036-8075. JSTOR: 1764604. URL: https://www.jstor.org/stable/1764604 (visited on 06/07/2024).

Mikolov, Tomas et al. (Sept. 6, 2013). *Efficient Estimation of Word Representations in Vector Space*. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781 [cs]. URL: http://arxiv.org/abs/1301.3781 (visited on 06/07/2024). Pre-published.

Mottram, James (Mar. 4, 2016). "Florence Foster Jenkins: Why Film-Makers Are Suddenly Interested in the Tone Deaf American Socialite — The Independent". In: *The Independent. Culture*. URL: https://www.independent.co.uk/arts-entertainment/films/features/florence-foster-jenkins-why-filmmakers-are-suddenly-interested-in-the-american-socialite-a6911201.html (visited on 06/09/2024).

Neelakantan, Arvind et al. (Jan. 24, 2022). *Text and Code Embeddings by Contrastive Pre-Training*. arXiv.org. URL: https://arxiv.org/abs/2201.10005v1 (visited on 06/09/2024).

Orbach, Barak Y. and Liran Einav (June 1, 2007). "Uniform Prices for Differentiated Goods: The Case of the Movie-Theater Industry". In: *International Review of Law and Economics* 27.2, pp. 129–153. ISSN: 0144-8188. DOI: 10.1016/j.irle.2007.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0144818807000488 (visited on 06/08/2024).

Pinkse, Joris, Margaret E. Slade, and Craig Brett (2002). "Spatial Price Competition: A Semiparametric Approach". In: *Econometrica* 70.3, pp. 1111–1153. ISSN: 0012-9682. JSTOR: 2692309. URL: https://www.jstor.org/stable/2692309 (visited on 04/12/2024).

Prag, Jay and James Casavant (Sept. 1, 1994). "An Empirical Study of the Determinants of Revenues and Marketing Expenditures in the Motion Picture Industry". In: *Journal of Cultural*

Economics 18.3, pp. 217–235. ISSN: 1573-6997. DOI: 10.1007/BF01080227. URL: https://doi.org/10.1007/BF01080227 (visited on 06/09/2024).

Ravid, S. Abraham and Suman Basuroy (2004). "Managerial Objectives, the R-Rating Puzzle, and the Production of Violent Films". In: *The Journal of Business* 77.S2, S155–S192. ISSN: 0021-9398. DOI: 10.1086/381638. JSTOR: 10.1086/381638. URL: https://www.jstor.org/stable/10.1086/381638 (visited on 06/09/2024).

Sajani, Armin (July 1, 2023). *Fast-Pagerank: A Fast PageRank and Personalized PageRank Implementation*. Version 1.0.0. URL: https://github.com/asajadi/fast_pagerank (visited on 09/05/2024).

Salop, Steven C. (1979). "Monopolistic Competition with Outside Goods". In: *The Bell Journal of Economics* 10.1, pp. 141–156. ISSN: 0361-915X. DOI: 10.2307/3003323. JSTOR: 3003323. URL: https://www.jstor.org/stable/3003323 (visited on 06/08/2024).

A Rearranging the Model

Let us start with the expression of the model as given in section 4.1 ("Construction"):

$$\ln(q_{it}) = \alpha_i + \lambda_{t-r(i)} + \xi_{it} + \sum_{j \neq i} f(d_{ij})(\alpha_j + \lambda_{t-r(j)} + \xi_{jt}) + \alpha_t$$

From there, we can rearrange the right side one small step at a time:

$$= \alpha_i + \lambda_{t-r(i)} + \xi_{it} + \sum_{j \neq i} f(d_{ij})(\alpha_j + \lambda_{t-r(j)} + \xi_{jt}) + \alpha_t$$

$$= \alpha_i + \sum_{j \neq i} f(d_{ij})\alpha_j + \lambda_{t-r(i)} + \sum_{j \neq i} f(d_{ij})\lambda_{t-r(j)} + \xi_{it} + \sum_{j \neq i} f(d_{ij})\xi_{jt} + \alpha_t$$

$$= \alpha_i + \sum_{j \neq i} f(d_{ij})\alpha_j + \lambda_{t-r(i)} + \underbrace{\sum_{r(k)} \left(\sum_{j \neq i}^{r(j)=r(k)} f(d_{ij})\lambda_{t-r(k)} \right)}_{\text{Group competitors of same age}} + \xi_{it} + \sum_{j \neq i} f(d_{ij})\xi_{jt} + \alpha_t$$

$$= \alpha_i + \sum_{j \neq i} f(d_{ij})\alpha_j + \lambda_{t-r(i)} + \underbrace{\sum_{r(k)} \left(\lambda_{t-r(k)} \sum_{j \neq i}^{r(j)=r(k)} f(d_{ij}) \right)}_{\lambda_{t-r(k)} \text{ constant within group}} + \xi_{it} + \sum_{j \neq i} f(d_{ij})\xi_{jt} + \alpha_t$$

$$= \alpha_i + \sum_{j \neq i} f(d_{ij})\alpha_j + \lambda_{t-r(i)} + \underbrace{\lambda_{t-r(i)} \sum_{j \neq i}^{r(j)=r(i)} f(d_{ij}) + \sum_{r(k) \neq r(i)} \left(\lambda_{t-r(k)} \sum_{j}^{r(j)=r(k)} f(d_{ij}) \right)}_{\text{Break the term for } r(i), \text{ films of age equal to } i\text{'s, out of sum}} + \xi_{it} + \sum_{j \neq i} f(d_{ij})\xi_{jt} + \alpha_t$$

$$= \alpha_i + \sum_{j \neq i} f(d_{ij})\alpha_j + \underbrace{\lambda_{t-r(i)} \left(1 + \sum_{j \neq i}^{r(j)=r(i)} f(d_{ij}) \right)}_{\text{Group matching } \lambda_{t-r(i)}} + \sum_{r(k) \neq r(i)} \left(\lambda_{t-r(k)} \sum_{j}^{r(j)=r(k)} f(d_{ij}) \right) + \xi_{it} + \sum_{j \neq i} f(d_{ij})\xi_{jt} + \alpha_t$$

Describing the same steps above in words:

- First, we break up the sum and rearrange the equation so the α terms are together, the λ terms are together, and the ξ terms are together.
- Next, we break down the sum over competitor λ terms into two sums; the outer sum is over the age of the competitor, while the inner sum is over the competitors of that age.

- From there we pull out the element of the outer sum corresponding films of the same age as film i.
- Finally we can pull $\lambda_{t-r(i)}$ out of the combined sum of film i's age coefficient and its competitors' weighted age coefficients.

Thus, in the resulting equation, each α and λ parameter appears only once, multiplying an element that is constructed entirely from data conditional on f.

B Matrix-Wise Identification of UV Decomposition

Let M be an $m \times n$ matrix where most observations are missing. We wish to identify matrices U and V, of dimension $m \times d$ and $n \times d$ respectively, such that UV' approximates the observed entries of M, minimizing mean squared error.

I've found it's easiest to start in the *vector* case, then stack to get to the matrix case. Let us start by adjusting one $1 \times d$ row u of U at a time to minimize mean squared error. Let m represent the corresponding row of M impacted by changes in u. We wish to solve the following optimization problem:

$$\min_u\{(m - uV')(m - uV')'\} = \min_u\{mm' - mV'u' - uVm' + uVV'u'\}$$

The dimensions of this objective imply the result is a scalar; thus, we can take the gradient with respect to u and set it equal to zero to find the optimal u. The gradient is:

$$\nabla_u = \vec{0} = -2mV' + 2uVV'$$

Solving for the optimal u^*, this yields the following first-order condition:

$$u^* = mV'(VV')^{-1}$$

Entirely analogously, we can optimize a row of V with respect to the existing U to find:

$$v^* = (U'U)^{-1}U'm$$

From here, it is not hard to stack the vectors to get the matrix-wide optimization:

$$U^* = MV'(VV')^{-1} \qquad V^* = (U'U)^{-1}U'M$$

By generating random starting U and V, we can iterate this process until convergence to find (locally) optimal U and V matrices.

These matrix forms are particularly appealing since they highlight the procedure's implicit relation to linear regression. Philosophically, we are searching for d latent characteristics of each film (corresponding to rows of V) and the analogous coefficients on these characteristics which define each user's taste (corresponding to rows of U) such that the coefficient-characteristic pair justify the existing reviews. From there, we can interpolate missing films by simply multiplying the latent characteristics by the coefficients for each missing user. Note since the problem is symmetric we could just as easily say the latent values for users are the "characteristics" and the values for films are "coefficients," but I prefer to think of films as more easily defined by a finite set of characteristics than users.