Gender Difference in Movie Genre Preferences Factor Analysis on Ordinal Data

Jiayu Wu UID: 905054229
Department of Statistics
University of California, Los Angeles
jiayuwu@ucla.edu

Abstract

Factor analysis is the reconstruction of data through a linear combination of independent continuous latent factors. However, the assumption of continuity is often violated in application, for example, with ordinal response in survey datas. In this report, three approaches to perform factor analysis on ordinal data is discussed from the perspective of category quantification strategies, with both theoretical reasoning and empirical comparison. In empirical analysis, factor analysis is conducted on a survey dataset about young people's preferences for 10 movie genres. A hidden structure with two latent factors is identified and analyzed, a clear gender difference in movie preference is also discovered. Moreover, we argue that nonlinear factor analysis is the most plausible approach for ordinal data exploration and analytic, while it could be aided by polychoric factor analysis.
keywords: polychoric correlation, nonlinear FA, category quantification, survey data

1 Introduction

Factor analysis is an analytic data representation method to extract a small number of independent and interpretable factors from a high-dimensional observed dataset with complex structure. However, classical factor analysis is limited in use for its assumption on continuity and linearity. Thus, there are various generalizations of factor analysis to deal with diverse data types common in application.

Ordinal manifest variable is common in real-world problems. For example, in survey data participants are often asked to express their attitudes in scales; in recommender system problems, users typically express their interests in an item by rating with "stars", etc. Therefore, it is meaningful to establish a plausible procedure to apply factor analysis on real data with ordinal manifest variables. With such an objective, both theoretical grounds and empirical evidence should be fully examined.

In the first section of this report, the motivation for this project and the structure of the report is presented. In the second part, classical factor analysis method and the challenge of performing factor analysis on ordinal data is firstly reviewed, then three approaches to solve the problem are reasoned from the perspective of category quantification. In the empirical analysis section, we tackled a survey dataset about young people's preferences on 10 movie genre based on the methodology discussed

before, in order to identify underlying latent factors and analyze gender differences. In the conclusion part, limitations as well as results from this project is discussed.

2 Background and Methodology

2.1 Classical factor analysis

For an observed data matrix $Y_{n\times p}$ with p continuous manifest variables, classical factor analysis theory states that, it can be restructured as a linear combination of d continuous latent factors $\{z_1, z_2, ...z_d\}$ plus a noise term ϵ:

$$Y = WZ + \epsilon, \quad Z \sim N(0, I_d), \epsilon \sim N(0, \sigma I_p).$$

The latent factors Z and the weights W are often obtained from matrix factorization methods (like eigen decomposition), or learned with EM algorithm. Rotation of the factors (usually orthogonally) for greater interpretability is also common, yet it is often suspected of over-interpretation.

Factors analysis is widely applied and has various interpretations. From the perspective of generative learning, Z is unobserved latent variables follow a normal prior and weights W determines the manifest of those latent variables. From a data analytic view, it is a data representation and dimension reduction to extract an orthogonal basis that preserves the most of the information from the high-dimensional dataset by decomposing the correlation matrix. This method with various generalizations has been proven to be powerful in identification of unobserved while meaningful latent factors in application of sparse coding (Olshausen & Field, 1997), generative neural network (Han & Yang, 2016) and recommender system(Koren, Bell & Volinsky 2009), etc.

The use of classical factor analysis, however, is limited due to its strong assumptions: 1) both manifest variables $\{y_1, y_2, ...y_p\}$ and latent factors $\{z_1, z_2, ...z_d\}$ are continuous; 2) latent factors $\{z_1, z_2, ...z_d\}$ follows multivariate normal distribution; 3) the mapping from latent factors to manifest variables is linear. (Magidson & Vermunt, 2003; Han & Yang, 2016)

In this project, we attempt to deal with violation of the first assumption, which is commonplace in real-world dataset.

2.2 FA with ordinal variables

An ordinal variable is a categorical variable for which the possible values are ordered, which is prevalent in real datasets. It is also notable that, methodology for ordinal variable can always be generalized to deal with other categorical variable, because binary variable can be regarded as a specific case of ordinal variable, and categorical variable can be transformed into binary dummy variables.

The problem with ordinal variables in factor analysis is that it does not have a measure on association as the Pearson correlation on continuous variables, while a correlation measurement is required to extract uncorrelated latent factors.

In order to implement factor analysis on ordinal data, a natural reasoning is to monotonically map the discrete ordinal levels to a continuous space. The fact that the levels are monotonic in magnitude justifies this attempt. The process of linearizing categorical variables is often referred to as category quantification.

2.2.1 Naive approach

The first conceivable way of category quantification is to assume uniform distribution and treat the levels directly as values from a continuous distribution. This approach is straightforward and easy to apply, however, it can be problematic especially when the ordinal scales in the data are not likert-type.

It is because that the true distance between the neighboring ordinal scales is unknown. When using them directly as values from a continuous distribution, the underlying assumption is that the distance between any two neighboring scales is identical. This assumption may hold when the data has likert-type scales, but is often violated when variables are not measured in comparable scales or even in completely different sets of scales.

2.2.2 Polychoric approach

The next approach to be considered is to infer from the ordinal observation an underlying normally distributed continuous trait. The assumption is that the ordinal responses is made upon a continuously distributed trait and thresholds for categorical decision. Thus a measure of association between those continuous traits can be obtained, referred to as the polychoric correlation (tetrachoric correlation in the binary case). Then, classical factor analysis could be performed on the polychoric correlation.

Polychoric correlation was developed to measure raters' agreement (Drasgow, 1988). The principle of estimation is to find the thresholds on a multivariate normal distribution which maximize the likelihood of observing the empirical manifest, then the correlation is computed (Uebersax, 2006). This can also be generalized to cases where other latent continuous distributions other than multivariate normal is assumed.

It has the advantage of making ordinal variables with different measuring scales comparable, hence overcomes the aforesaid shortcoming of the naive approach. The cons could be that the assumption that each ordinal variable has an underlying continuous trait following some designated distribution does not always hold.

2.2.3 Nonlinear FA approach

As introduced before, the objective in factor analysis is to maximize the proportion of variance from the original data explained by a limited number of latent factors, while with ordinal data we intend to find and fix the distances between scales. In nonlinear factor analysis, the technique called optimal scaling is used to achieve those two goals at the same time. It can be understood as a process of simultaneous data representation and data transformation (Takane, 2005).

The techniques for deriving optimal quantification of categorical variables is discussed in homogeneity analysis (De Leeuw & Mair, 2007), which is applied in nonlinear FA (also referred to as Categorical PCA or CatPCA) (De Leeuw, 2011). The basic intuition is to firstly use binary dummy variables to reconstruct a categorial variable, then find a linearizing transformation of the binary matrix in a way that the correlation is maximized as possible. Thus, it becomes an optimization problem with rank constraint. As a specific case of categorical variable, optimal scaling on ordinal variable is accomplished with an additional level constraint to make the first column of categorical quantifications monotone.

The assumption of nonlinear FA is bilinearizability, which means that transformations of the variables can be found such that all bivariate regressions are exactly

linear (De Leeuw, 2005). This is an assumption weaker than assuming multivariate normality (De Leeuw, 1988) as with the polychoric approach. Nonlinear FA is an analytic-driven approach with the least model assumption in dealing with ordinal (categorical) variables, and thus can easily incorporate simultaneous analysis of multiple variables with different ordinal scales or even with different data type. A limitation of this method is that the rank constraint (the number of factors) must be determined in advance, and such a choice can be hard to justify because the criterion of the cumulative variance explained does not apply.

2.3 Discussion

In summary, ordinal manifest variables violates the continuity assumption in classical factor analysis, therefore, category quantification - to determine a scale distance measurement to linearize ordinal data - is required. Three approaches with decreasing strength of assumption are introduced to solve this problem. Theoretically, the nonlinear FA approach is the most assumption-free, hence an optimal choice for analytic-based data analysis, while it need the aid of related methods in choosing the number of constraints and cross-validating the interpretation.

3 Empirical Analysis

In this part, factor analysis is performed on a real-world survey dataset about young people's rating on movie genres. The aim is to reconstruct the data with 10 variables into a simple structure with latent factors susceptible of interpretation, and to explore gender difference in movie preferences. Based on the analysis, we also try to develop a plausible procedure for factor analysis with ordinal manifest variables.

3.1 Data preprocessing

The data used in this report is a survey dataset of participants' rating in 1-5 scale on 10 movie genres, related demographic features is also available. There are 1010 raw observations, 955 is left after eliminating missing values and dubious records.

The data is rechieved from kaggle in March, 2018. The original dataset is based on a 2013 survey on young people aged between 15-30 (mostly around 20). There are 150 survey questions in total, presented to 1010 participants in both electronic and written form. In the movie preferences part of the survey, participants are asked to rate 10 movie genres in a 1-5 scale from "Don't enjoy at all" to "Enjoy very much". The genres questioned includes: horror, thriller, comedy, romantic, Sci-Fi, war, fantasy, animated, documentary and action. The survey also provides demographic features such as age and gender.

In data preprocessing, records with missing value and with the same rating for every question are regarded as dubious and eliminated. Due to the magnitude of observations and the general aim to explore latent factors underlying manifest variables, such preprocessing increase the credibility of the data without too much harm to the completeness of information.

As a result, a dataset with 955 observations (by row) and 12 features is derived and summarized as the following:

Variable	Range	Mean	Median	skew	kurtosis
Age	[15,30]	20.43	20	1.52	0.09
Gender	{0"female", 1"male"}	0.41	0	-	-
Horror	1-2-3-4-5	2.75	3	0.20	-1.24
Thriller	1-2-3-4-5	3.37	4	-0.36	-0.82
Comedy	1-2-3-4-5	4.50	5	-1.63	2.45
Romantic	1-2-3-4-5	3.49	4	-0.34	-0.85
Sci-Fi	1-2-3-4-5	3.09	3	-0.04	-1.12
War	1-2-3-4-5	3.15	3	-0.06	-1.16
Fantasy	1-2-3-4-5	3.76	4	-0.55	-0.72
Animated	1-2-3-4-5	3.78	4	-0.67	-0.63
Documentary	1-2-3-4-5	3.64	4	-0.48	-0.58
Action	1-2-3-4-5	3.50	4	-0.40	-0.89
Table 1.					(955 obs.)

3.2 Pearson correlation v.s. polychoric correlation

In this section, we try to compare the association between 10 main features measured by Pearson correlation and by polychoric correlation. The former is justified as our data are collected under a likert-type scale survey design. Whereas, the later seems to be more plausible as in practice scale-level rating is hardly guaranteed, and the kurtosis and skewness of some variables (ex. "Comedy") indicates a violation of uniform assumption.

In Figure 1, the computed correlation matrices are displayed. For visual clarity, positive values are shaded in blue while negative ones in red, and the greater the absolute value of the correlation, the deeper the color. It can be observed that the colored patchs have very similar patterns, while the polychoric approach suggests a stronger association. This is further illustrated in Figure 2. which shows the differences in absolute value of elements from the two correlation matrix.

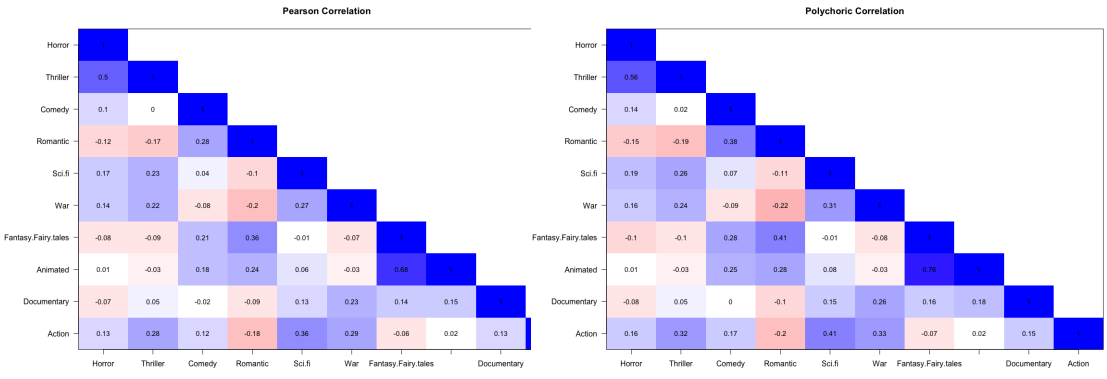

Figure 1.Pearson and polychoric correlation for movie preference data

From this comparison we find that Pearson and polychoric correlations behave similarly for this dataset, because the survey design of likert-type scales makes a uniform assumption justifiable. As a result, factor analysis based on those two correlation matrices is likely to give similar reconstruction of the original data

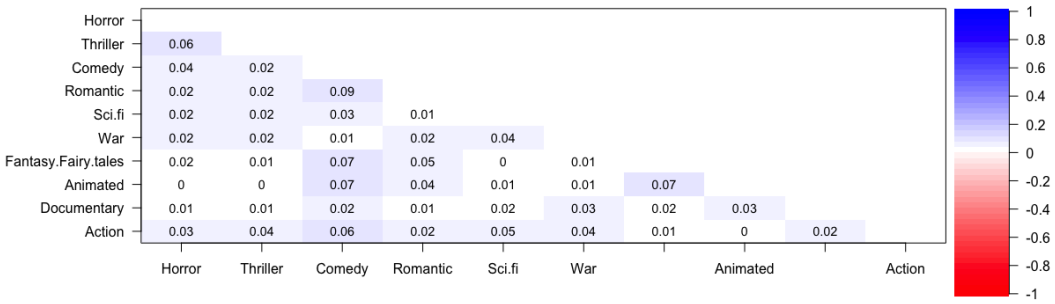

Figure 2. Difference in absolute values

structure. However, polychoric approach gives a stronger correlation measure, thus a larger proportion of variance could be addressed in factor analysis. This is confirmed by eigen decomposition in Table 2. where the eigenvalues of polychoric correlation matrix are bigger:

Corrleation	λ_1	λ_2	λ_3	λ_4	λ_5
Pearson	2.461	2.126	1.332	0.974	0.744
Polychoric	2.289	1.943	1.296	0.986	0.788
	λ_6	λ_7	λ_8	λ_9	λ_{10}
	0.669	0.612	0.517	0.343	0.222
	0.709	0.663	0.595	0.433	0.298

Table 2.

3.3 Classical FA with polychoric correlation

In this section, classical factor analysis with the polychoric correlation is performed.

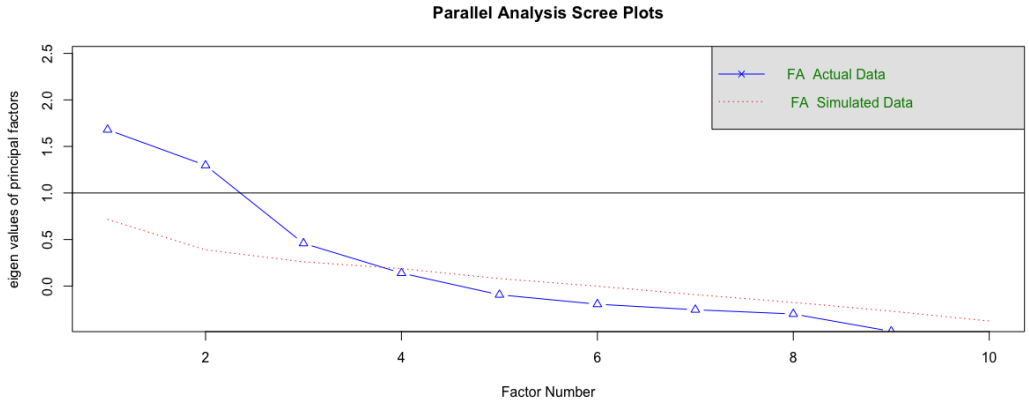

Figure 3. Scree plot of factors

Firstly, we decide on the number of factors to be chosen. By far, three factors seem to be a reasonable choice as the first three eigenvalues exceed 1 as in Table 2. The

scree plot in Figure 4 also shows a sharp break after the third eigenvalue. In addition, Figure 4 presents a comparison with the scree of a random data matrix of the same size marked in dashed line, which once again confirm the choice of three factors.

Therefore, a factor analysis with three latent factors is performed, with maximum likelihood estimation. 45.8% of variance is explained cumulatively, the mean item complexity is 1.3, indicating that three factors are sufficient. A detailed summary is presented on the left side in Table 3, while on the right side variables are sorted by the magnitude of the loadings on each factor and marked with "*" if the loading is greater than 0.3.

Variable	F_2	F_1	F_3	communalities	uniqueness	F_2	F_3	F_3
Horror		1.00		1.00	0.005	Fantasy*	Horror*	Action*
Thriller	-0.05	0.57	0.31	0.42	0.579	Animated*	Thriller*	Sci-Fi*
Comedy	0.33	0.14		0.13	0.874	Romantic*	Sci-Fi	War*
Romantic	0.42	-0.15	-0.27	0.28	0.722	Comedy*	War	Docu.*
Sci-Fi	0.02	0.19	0.54	0.33	0.672	Docu.	Action	Thriller*
War	-0.07	0.17	0.52	0.31	0.692	Sci-Fi	Comedy	Animated
Fantasy	0.93	-0.11		0.87	0.127	Horror	Animated	Comedy
Animated	0.81		0.09	0.67	0.334	Thriller	Docu.	Horror
Documentary	0.16	-0.08	0.37	0.17	0.834	Action	Fantasy	Fantasy
Action	-0.05	0.16	0.63	0.42	0.581	War	Romantic	Romantic
Variance	0.18	0.15	0.13					
Cum. var	0.18	0.33	0.46					

Table 3.

F_2 accounts for 18% of total variance. It is primarily defined by the variables "Fantasy", "Animated" and "Romantic", it could be interpreted as a factor of the preference for storyline and emotional conveyance. F_1 explained 15% of total variance, the variables "Horror" and "Thriller" have high positive loadings, while "Fantasy" and "Romantic" load negatively on this factor, which indicates that F_1 may be a latent factor of the preference for excitement from movie. The proportion of variance explained by the third factor F_3 is 13%, the variables "Action", "Sci-Fi" and "War" have high loadings on this factor. A possible interpretation of F_3 is the preference for the scene and special effects of a movie. In Figure 4., the interpretation of three factors is visually displayed, all loadings with an absolute value greater than 0.3 are represented as an edge.

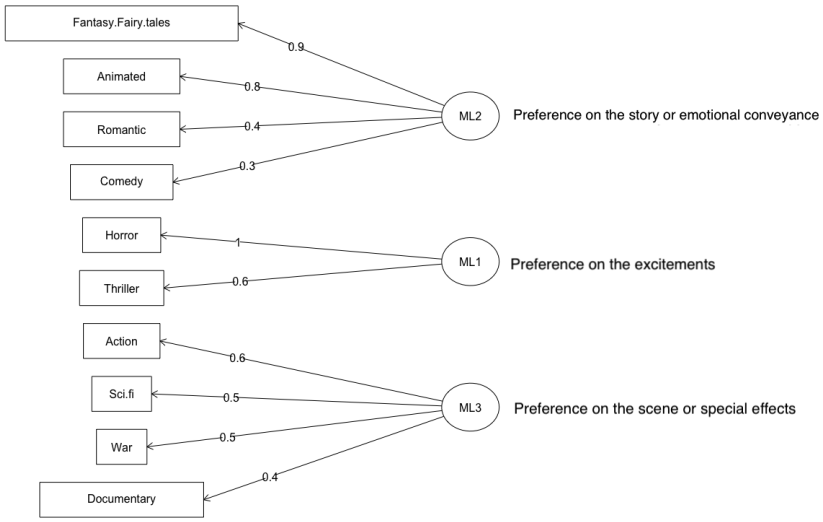

Figure 4. Graphical presentation of factors interpretaion

A powerful visual aid for exploratory factor analysis is the biplot, as displayed in Figure 5, which shows both the observations and factors in a plot. In order to get some understanding of gender difference in the preferences for movie genre, the observation in the biplot is marked in orange if the participant is female, otherwise in blue. It can be observed that the gender difference is more obvious on F_3 and F_2. Observations from male participants cluster on the positive side along the F_3 axis, which indicates a preference on movie genre with spectacular scene and special effects, like action movie or Sci-fi movie, while females do not seem to have general propensity on this factor. Whereas, female participants give more responses reflected on the positive side along the F_2 axis, and tend to rate lower on movie genres that lacks emotional conveyance, while males give evenly distributed responses along this axis.

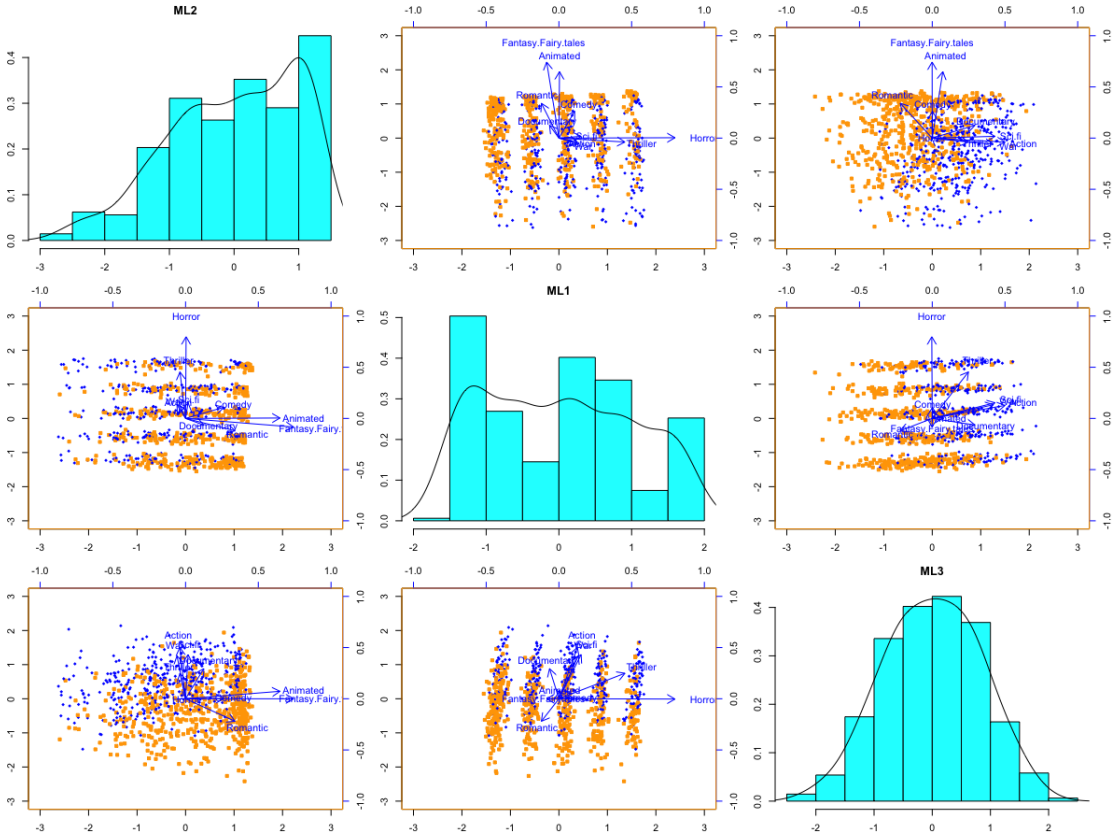

Figure 5. FA biplot with observations by gender

Another important implication of Figure 5. is that it seems that the latent factors F_2 and F_1 are not normally distributed, which is a violation of the assumption of classical factor analysis. Moreover, the previous interpretation of latent factors entails ambiguity, as the interpretation of F_1 and F_3 give rise to overlaps. Therefore, it is necessary to further explore the structure of the data in the nonlinear factor analysis approach, in order to avoid false specification of the latent factors and over-interpretation of the pattern.

3.4 Nonlinear FA with optimal scaling

In this section, we apply on the dataset the most assumption-free way to perform factor analysis on ordinal data: nonlinear factor analysis.

The first challenge is to decide on the number of factors for analysis. In nonlinear factor analysis, category quantification is attempted at the same time as maximizing the variance, so the proportions of variance explained (typically measured with eigenvalues in the classical method) change with the rank constraints, making it necessary and difficult to dertermine in advance the number of factors, i.e, the rank constraints.

To start with, three factors as in the polychoric approach is tried, however the resulting third factor seems to be redundant as showed on the left side in Figure 6, where the third factor is dominated by a single variable.

Because eigenvalues from nonlinear factor analysis change with different rank constraint, the scree plots with three factors and with ten factors respectively are printed in Figure 7 for a better view, a sharp break after the second eigenvalue can be identified from both plots. Therefore, we should try to proceed with two factors. In fact, the graphical presentation on the right side of Figure 6. indicates better interpretability with two factors.

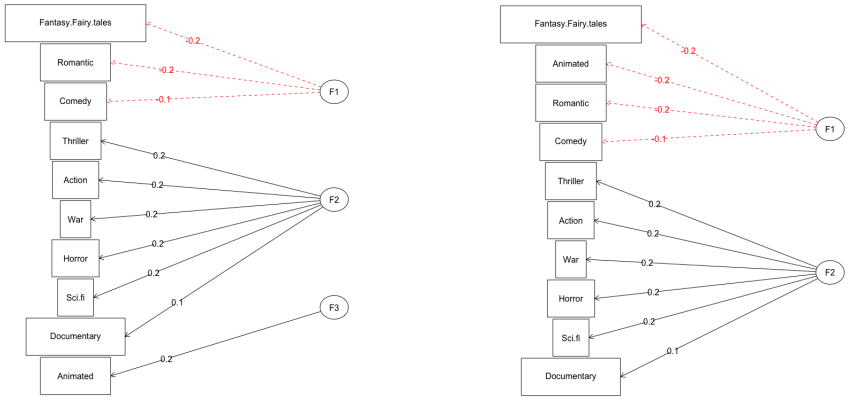

Figure 6. Comparison between two and three factors for nonlinear FA

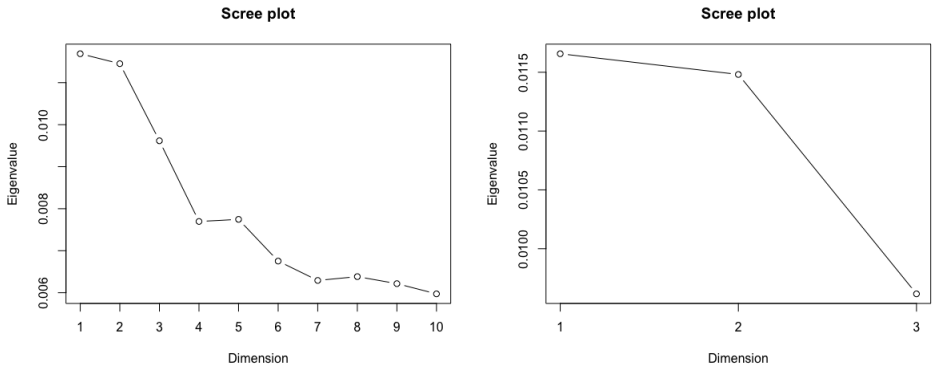

Figure 7. Scree plots for nonlinear FA with different constraints

The loading matrix is reported in Table 4. $\tilde{F}_1$ can be interpreted as the preference for story and emotional conveyance, which is analogously to F_2 from the polychoric approach, yet in the opposite direction. Whereas, $\tilde{F}_2$ seem to synthesize F_1 and F_3 as one single factor of the preference for grand scenes and excitement from the movie. This structure is not only simpler than that derived in the polychoric approach previously, but also less ambiguous in interpretation.

Variable	$\tilde{F}_1$	$\tilde{F}_2$	communalities	uniqueness
Horror	0.04	0.16	0.029	0.97
Thriller	0.07	0.19	0.042	0.96
Comedy	-0.14	0.03	0.021	0.98
Romantic	-0.15	-0.12	0.038	0.96
Sci-Fi	0.04	0.16	0.028	0.97
War	0.07	0.17	0.034	0.97
Fantasy	-0.20	-0.05	0.042	0.96
Animated	-0.16	-0.03	0.026	0.97
Documentary	-0.06	0.10	0.012	0.99
Action	0.05	0.18	0.036	0.96
Variance	0.41	0.59		

Table 4.

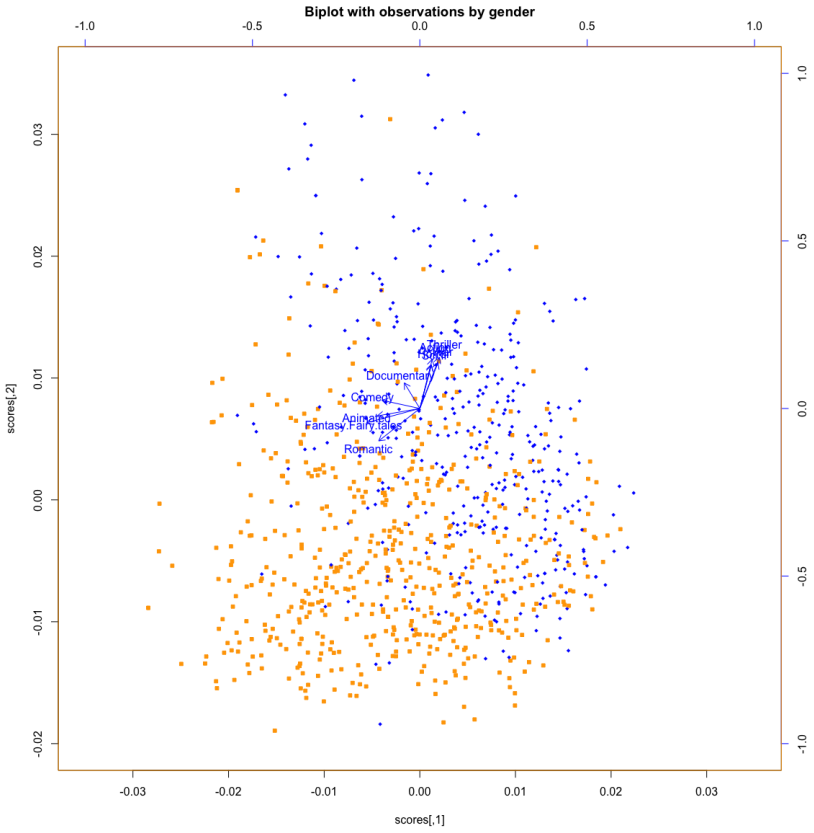

Figure 8. Nonlinear FA biplot

The biplot by gender in Figure 8. suggests that male participants typically have a preference for movie genres with grand and exciting scenes while less interested in the emotional resonance from movies, as the observations in blue cluster to the upper right and align with the variables "Thriller", "Horror", "Action", "War" and "Sci-Fi". Females, on the other hand, do not show a preference as clear-cut, whereas there is also a general tendency towards movie genres with more profound sentimental content such as "Fantasy", "Romantic" and "Animated". From the loading plot in Figure 9, we may obtain a more clear view of the genres that are close to each other, and it is notable that this distance clustering on the dimensions defined with two latent factors is in line with commonsense. It is also remarkable that the two latent factors seem to be normally distributed as showed in the histograms.

Compared with classical factor analysis with polychoric correlation, nonlinear FA gives a more accountable result. For one thing, the reconstruction of data is more reliable, as weaker assumption is required and the resulting latent factors are normally distributed. For the other thing, the data representation is more interpretable, since a simpler structure with less factors is derived with a reasonable interpretation that reveals hidden patterns of the data.

3.5 Discussion

Based on nonlinear factor analysis, we derive two latent factors of the preference for storyline or emotional conveyance and for scene or excitement from the original data. These two factors are roughly normally distributed and can be synthesized into the manifest variables with nonlinear transformation plus noise term. With this factorization, we also reveals gender difference in movie preference and clustering of closely-related movie genres.

Moreover, we demonstrate the advantages of nonlinear factor analysis on ordinal data that it requires less assumption, hence gives more genuine data representation with a simpler, more interpretable structure. Nevertheless, it could be aided by polychoric factor analysis. Firstly, the polychoric correlation gives a overview of the association between variables before deriving factors. Secondly, the computation as well as intuition of polychoric factor analysis is simple and straightforward, therefore, it offers a reference beforehand and a validation afterwards, on the choice of the number of factors and the interpretation of the latent structure.

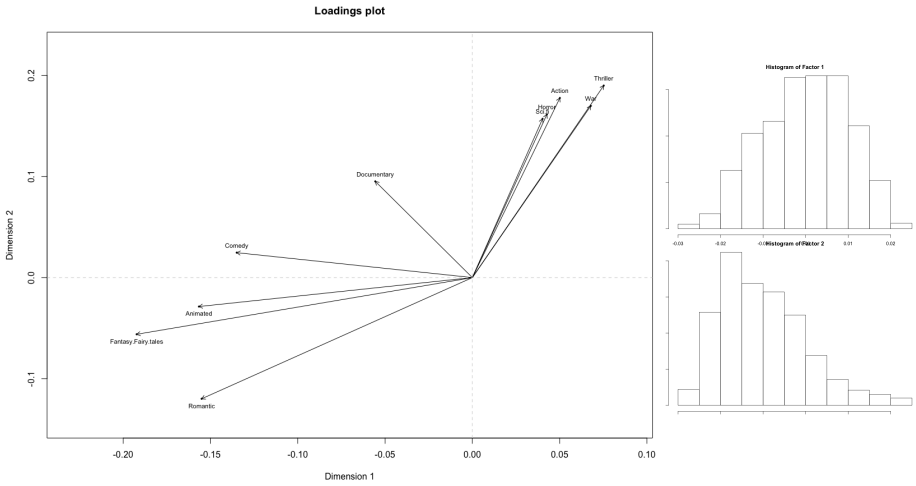

Figure 9. Loading plot and factor histogram

4 Conclusion

In conclusion, this report deals with the application of factor analysis on ordinal data. The theoretical challenge in category quantification and three corresponding solutions are first discussed, polychoric factor analysis and nonlinear factor analysis is then performed on real data to analyze young people's movie preferences.

The analysis identified a factor of the preference for the storyline and emotional conveyance and a factor of the scene and sensational excitement. Males typically have a preference for exciting movie scenes while less interested in emotional resonance. Whereas, females has a general propensity for profound storyline and sentimental content. It can also be concluded that, "Thriller", "Horror", "Action", "War", "Sci-Fi" are closely-related genres, as the opposites to "Fantasy", "Romantic" and "Animated".

Nonlinear factor analysis is demonstrated to be the optimal approach with the minimum assumption and the greatest effectiveness in recognizing hidden structure that is accountable as well as simple. It is also shown that its application could be aided by the more straightforward polychoric approach for factors selection and interpretation.

For future extension on this project, there are several directions to be considered. For one thing, with such a large dataset, bootstrap could be used to cross-validate the genuinety of the revealed hidden structure; and model-based approach with a similar intuition, such as item response theory, can be compared. For the other thing, other variables from the same survey dataset could be included to study, for example, correlation between movie preferences and music preferences. In addition, the same methodology should be tried on more complex datasets like those with higher dimension and missing values, to address a realistic need in analyzing the ratings of movies or other specific items.

References

De Leeuw, J., & Mair, P. (2007). Homogeneity Analysis in R: The package homals.

De Leeuw, J. (2011). Nonlinear principal component analysis and related techniques.

De Leeuw, J. (2005). Multivariate analysis with optimal scaling.

Drasgow F. (1988). Polychoric and polyserial correlations. In Kotz L, Johnson NL (Eds.), Encyclopedia of Statistical Sciences. Vol. 7 (pp. 69-74). New York: Wiley.

Han, T., Lu, Y., Zhu, S. C., & Wu, Y. N. (2017). Alternating Back-Propagation for Generator Network. In AAAI (Vol. 3, p. 13).

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer 42(8):30?37.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision Research 37(23):3311?3325.

Magidson, J., & Vermunt, J. K. (2003). Comparing latent class factor analysis with traditional factor analysis for datamining. Statistical data mining and knowledge discovery, 373-383.

Uebersax, J. S. (2006). Introduction to the tetrachoric and polychoric correlation coefficients. Obtenido de http://www. john-uebersax. com/stat/tetra. htm.[Links].