Can I See Beyond What You See? Blending Machine Learning and Econometrics to Discover Household TV Viewing Preferences

Zhuolun Li, Robert J. Kauffman, Bing Tian Dai
Living Analytics Research Centre, School of Information Systems, Singapore Management University
{zlli, rkauffman, btdai}@smu.edu.sg

Abstract

This research blends machine learning-based discovery of preference patterns that uses natural language processing for TV viewing data with explanatory modeling that uses econometrics, as a basis for understanding TV viewing preferences at the household-level. We employ a dataset of about 1.1 million observations that was collected via set-top box technology that tracked household-level consumption of the content of its channel subscription package. The data describe the details of what households watched on TV, including the channels and shows, start times and durations, and overall viewing times for content from different digital entertainment genres. This research demonstrates the efficacy of our machine learning and explanatory econometrics approach, and presents insights on consumer behavior and content bundling that are useful for firm strategy in digital entertainment services.

1. Introduction

The medium of television has the power to reach an extraordinarily large audience. According to media market research firm Nielsen [14], the number of U.S. households with TV access in the 2015-2016 viewing season was 116.4 million. The U.S. market is the largest in the world, with an estimated TV subscription revenue base of USD 105.3 billion projected for 2019. Not far behind is China, with TV subscription revenue of USD 24.1 billion [15]. Meanwhile, TV service providers have been experiencing major business disruption: essentially a "videoquake" according to PwC consultants, Bothum and Vollmer [3]. In the present era of digital entertainment, more people are choosing to stream video via *over-the-top services*, in which content is delivered through the Internet and via mobile phones, as opposed to more traditional cable and satellite TV subscriptions. Meanwhile, households seem to demand fewer channels, more personalized choices, and lower monthly bills. As a result, service providers are challenged to create more finely segmented and affordable, ever smaller program bundles to maintain their pay-TV subscription rates and revenues.

How can they respond? The key to designing more personalized program bundles with higher ROI is to understand and quantify household-level TV viewing preferences. It is possible to use observable demographics to associate household characteristics with viewing preferences [11]. Household information such as race, ethnicity, dwelling types and income level are useful to infer viewing preferences for channel bundle subscriptions and subsequent program consumption – but not enough. The details of household preferences are unobservable, and service providers don't know who in a household is watching at any given time: they can't see what the members of a household can easily observe about their preferences for TV viewing.

Or can they? Due to IT advances, service providers in the digital entertainment industry can now monitor household TV viewing behavior more fully [8]. The mechanism that makes this possible is *set-top box and identity chip-based return path data* (RPD) [4,5]. This enables service providers to know exactly when and what content households are consuming. This permits a provider to observe a household's choice of channels and programming genres over others to map out its *viewing choice set*. This also supports the quantification of viewing preferences, by sampling RPD streams and performing big data analytics to extract knowledge from its patterns.

We ask three questions: (1) How can the blended or "fusion analytics" application of machine learning and econometric methods create insights to help a digital entertainment service provider to improve its informedness about consumer viewing preferences? (2) What specific results and insights are we able to obtain from the dataset that we utilized in this research? And (3) what can be learned for program bundling design to improve service ROI?

2. Some Ways to Discover Preferences

Our research is related to consumer segmentation in Marketing, choice models in Microeconometrics, and machine learning and natural language processing-related text mining in Computer Science.

2.1. Consumer Segmentation in Marketing

Marketing segmentation views a heterogeneous market as being composed of smaller homogeneous markets. Firms can increase their profitability with

URI: http://hdl.handle.net/10125/41630
ISBN: 978-0-9981331-0-2

market segmentation, a tenet of the classic price discrimination model. Various approaches to segmentation have been applied. Among *predictive approache*s, cluster-wise regression and mixture models have been important.

The heterogeneity of consumer preferences in market segmentation has focused on how to group consumers into homogeneous segments. This study examines heterogeneous consumer preferences at the consumer level instead of the segment level. *Within-segment heterogeneity* is possible in our work. Our unit of analysis for the consumption of digital entertainment is the *household level*. Household heterogeneity can be identified from demographic characteristics and unobservable preferences extracted by machine learning methods.

2.2. Discrete Choice Models in Microeconometrics

Analysis of consumer choice involves modeling discrete dependent variables for purchase decisions. In these models, a consumer is faced with a set of alternatives and makes a utility-maximizing choice. The choice reveals the underlying preferences.

Discrete choice models apply to situations where one alternative is chosen among a set of mutually-exclusive alternatives because these models assume alternatives are perfect substitutes. This is far from perfect since it is common for consumers to simultaneously choose several alternatives from their choice set.

We will relax the *perfect-substitutes assumption* by adopting a *translated non-linear utility model* [1]. It can handle situations when a consumer selects multiple alternatives simultaneously. Satiated consumption of each alternative (diminishing marginal utility) can be included. This study is different from previous research: our approach with Bayesian econometrics can estimate sample preferences and individual preferences. We also model consumer heterogeneity instead of assuming that household heterogeneity follows a normal distribution with a constant mean [9].

2.2. Machine Learning, Data Mining, and CS

Machine learning in Computer Science offers numerous useful methods for the discovery of knowledge, content and patterns in complex and large datasets [13]. Among the many methods, topic modeling and *latent Dirichlet allocation* (LDA) from natural language processing [2] support discovery of unobserved groups of observations in data. LDA is effective in unsupervised text mining for topic discovery involving *unobservable latent factors* [7].

For our context, LDA can support joint estimation of *TV program genres* to view (as topics related to what is selected) and household consumption. LDA also can extract information on preferences because it can identify preferred program viewing genres, due to commonality of the words used in describing the content that households watch. Big data, sparse matrices, and different time periods all can be handled as well. And these things turn out to be perfect for us in this research on TV viewing.

3. Research Context, Data and Variables

We applied our blended analytics approach to discover preferences in digital entertainment with a large dataset on household TV viewing. The dataset for household-level TV viewing was provided by a digital entertainment firm. We use the viewing histories of 197,186 households over a 7-month period from December 2012 to June 2013.

The household characteristics were captured by a set of demographic variables, including nationality, dwelling type, and region of residence. *Nationality* identifies demand differences for local and international households. *Dwelling type* and *region of residence* proxy for household income levels. To understand household-level TV viewing interest, we collected the textual descriptions of channels and programs. We used these digitized contents to extract the *hidden topics* underlying the program content consumed by households. The topics were extrapolated to household preferences for different genres. We used observable demographics and the extracted preferences in modeling household differences.

For each TV viewing session, we acquired the timestamps and duration from our RPD data, and the genres, channels and programs they watched. A household's consumption of TV contents was aggregated into the genre-level to avoid revealing details of programs and channels. The firm classified TV content into genres defined by the digital entertainment firm, involving programming related to news, children, movies, and so on.

We also gained access to household-level subscription data for genres and channels. Since the subscription fee for a household was only charged monthly, we further aggregated the household TV viewing data at the *household-genre-month* level. Our final dataset has 1.1 million observations.

4. Modeling TV Viewing Preferences

4.1. An LDA Model to Mine TV Set-Top Box Data

A topic model is appropriate for modeling household and genre-level descriptions, to discover the probability distribution of each genre over a set of latent topics, relative to household-level TV viewers. Most important for this research is that the heterogeneity of household preferences can be determined based on the TV programs that the members of a

household view. Electronic Program Guide (EPG) descriptions to which we had data access to summarize the programs in textual form, so consumption preferences can be discovered through text analytics for such descriptions. To carry out the analyses, we first applied this approach to obtain household consumption preferences for the programs that they watched. This enabled us to do the same thing at the genre level to get genre descriptions.

Household-level consumption preferences and genres occur in probability distributions over a set of latent topics. So a household's preferences for a certain genre can be modeled in terms of the similarity between two corresponding probability distributions.[1] Their similarities are compared to indicate each household's preferences for each genre.

4.2. Theory-Based Econometrics for TV Viewing

In Microeconomics, consumers select goods and services from a choice set with a budget constraint. They must allocate their budget to yield the greatest value through a process called *utility maximization* in classical demand theory. Instead of predicting households' TV viewing based on a set of variables, we will explain it using this theoretical perspective. Service providers must understand the *household utility function* for digital entertainment to create genre and channel bundles, and recommend programs that are profitable.

To do this, we adapted the *translated non-linear utility model* [9] to represent household-level TV viewing behavior. The utility for household *i* to view consumption of content from the different genres is:

$$\overline{U_j}(\boldsymbol{x_i}) = \sum_j \psi_{ij}(x_{ij} + \gamma_j)^{\alpha_j}, \quad (1)$$

where $\boldsymbol{x_i}$ is the consumption vector of household *i* for different genres. ψ_{ij} is the *baseline utility* for household *i* for genre *j*, and α_j represents a household's consumption satiation.[2] High baseline utility and low satiation of a genre will lead to a *variety-avoidance pattern*: consumption of almost one specific genre. In contrast, low baseline value and high satiation for a genre will result in a *variety-seeking pattern*: consumption of a variety of genres.[3]

Household heterogeneity is handled by using a multivariate regression specification:

$$\boldsymbol{\psi_i} = \Delta' \boldsymbol{z_i} + \boldsymbol{u_i}, \;\; \boldsymbol{u_i} \sim N(0, V) \quad (2)$$

where $\boldsymbol{\psi_i}$ is the baseline utility vector for household *i* and $\boldsymbol{z_i}$ includes household i' s demographics (region of residence, nationality, dwelling type) and unobservable genre preferences extracted by LDA. The error term $\boldsymbol{u_i}$ follows a multivariate normal distribution with V as the variance-covariance matrix. A household's budget constraint is:

$$\sum_j p_j x_j \le E, \quad (3)$$

where p_j is the price of genre *j* and E is a household's total expenditure on TV viewing.

Solving the utility maximization problem for Eq. 1-3 yields a likelihood function:

$$P\left(x_{ip}^* > 0 \; and \; x_{iq}^* = 0; \; p = 2, \dots, n; \; q = n+1, \dots, m\right)$$
$$= \int_{-\infty}^{h_{im}} \cdots \int_{-\infty}^{h_{i,n+1}} \phi\left(h_{i2}, \dots, h_{in}, v_{i,n+1}, \dots, v_{im} \middle| 0, \Omega\right) \cdot |J| dv_{i,n+1} \cdots dv_{im}, \quad (4)$$

where $\boldsymbol{x_i^*}$ is the vector of household i' optimal consumption. (See [9] for details.) This m-vector $\boldsymbol{x_i^*}$ includes n non-zero components and $(m-n)$ zero components. The other variables involved in the likelihood function are defined as:

$$h_{ij} = \left[ln\left(\psi_{i1}\alpha_1\left(x_{i1}^* + \gamma_j\right)^{\alpha_1 - 1}\right) - \ln(p_1)\right] - \left[\ln\left(\psi_{ij}\alpha_j\left(x_{ij}^* + \gamma_j\right)^{\alpha_j - 1}\right) - \ln(p_j)\right], \quad j = 2, \dots, m \quad (5)$$

$$v_{iq} = \varepsilon_q - \varepsilon_1, \;\; \varepsilon_1 \; and \; \varepsilon_q \sim N(0,1); \; q = n+1, \dots, m \quad (6)$$

$$J_{rc} = \frac{\partial h_{i,r+1}}{\partial x_{i,c+1}^*}, \;\; r, c = 1, \dots, n-1 \quad (7)$$

These equations make our estimation possible.

5. Results and Discussion

We next will present the distributional characteristics of household preferences for TV viewing extracted by LDA, follow this with our interpretation of their economic and marketing impacts. We will especially comment on the differences in content demand based on the empirical preferences that we distilled from our econometric analysis.

5.1. Machine Learning Results Based on LDA

For topic modeling in this research, we set the number of topics at 20, which is large enough to capture the different topics of the TV programs. We accessed the topics by focusing on words that contribute most to the identification of a topic. The topics are described in terms of categories based on domain knowledge and observed patterns in the data:

[1] *Kullbach-Leibler* (*KL*) *divergence* from information theory is often used as a similarity assessment or information loss gauge between two probability distributions, one representing *ground truth* and the other representing *theory or model-based estimates* [10].

[2] For Eq. 1 to be a valid utility function, ψ_{ij} must be positive and α_j must be in the unit interval.

[3] γ_j controls the utility translation and ensures a corner solution [1], so household *i* only watches one genre.

- *Linguistic*. There was a concentration of household consumption for TV programs in non-English languages. This suggests an ethnic or linguistic topic for each non-English language.
- *News*. Some households only consumed news programs in the early morning or late night. The topics were local, Asia and business news.
- *Documentary*. These were extracted from documentary programs.
- *Sports*. Household consumption of sports programming also exhibited concentration. Sports topics capture a specific kind of sport.
- *Kids*. There were several kids topics, focused on cartoons and animation, targeted at different age group. There was a finer grouping of kids programs still all were under the same genre.
- *Variety*. Households tended to consume the same programs each day, especially variety shows.

With these topics, the model we built can give a useful reading on household TV viewing patterns, and was a basis for extracting their viewing preferences. Table 1 summarizes the descriptive statistics of household preferences for the different genres.

Table 1. Extracted preferences: descriptive stats

VARIABLE	MEAN	STD. DEV.
Chinese	8.274	3.078
Education	4.088	1.901
Entertainment	4.859	1.860
Infotainment	10.580	2.939
International	3.401	1.177
Kids	5.199	2.553
Lifestyle	4.263	2.019
Movie	4.072	1.921
News	7.590	2.470
Sports	11.892	2.238
Notes. 163,231 households; 1.1 million obs. Extracted from LDA model results.		

The topic model extracted these statistics from the TV viewing histories of 1.1 million households. Households, in general, seem to have had the highest preferences for the *Sports* genre, suggesting a large proportion of them were sports enthusiasts. We also observed that the *Infotainment* and *Chinese* genres were highly preferred. The digital entertainment firm packaged ethnic programs together to achieve this. These also show the popularity of ethnic content among the households. In contrast, households had the lowest preferences for the *International* genre.

To segment the household population, it is important for the digital entertainment firm to understand the correlation of household preferences for different genres. The correlations are shown in Table 2.

Household preferences for different genres are not independent, as might be expected. For instance, *Entertainment* is highly correlated with *Education*, *Kids*, *Lifestyle*, and *Movie*, at the level of 0.66, 0.71, 0.76, and 0.63, respectively. This reveals the pattern of *composite preferences*, indicating that different members of a household probably have distinct preferences for a subset of genres. (See Table 3 for the composite preference component details.)

Table 2. Correlations: extracted genre preferences

	Ch	*Ed*	*En*	*If*	*It*	*K*	*L*	*M*	*N*	*S*
Ch	1									
Ed	.26	1								
En	.52	.66	1							
If	.79	.04	.32	1						
It	-.02	-.17	-.26	-.02	1					
K	.14	.58	.71	.01	-.62	1				
L	.31	.95	.76	.16	-.16	.58	1			
M	.32	.74	.63	.20	-.16	.54	.75	1		
N	.48	.28	.24	.16	-.03	.18	.29	.06	1	
S	.27	-.01	.06	.37	.15	-.06	.01	-.36	.31	1
Notes. *Ch = Chinese; Ed = Education; En = Entertainment; If = Infotainment, It = Intl.; K = Kids; L = Lifestyle; M = Movies; N = News; S = Sports*.										

Table 3. Principal component eigenvectors

COMPONENT	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*
Chinese	.27	.49	.01	-.29	-.26	.09
Education	.42	-.13	.24	.26	.11	-.37
Entertainment	.43	.02	-.04	-.06	.18	.62
Infotainment	.17	.51	-.04	-.50	.12	-.23
International	-.17	.22	.75	.17	.09	.44
Kids	.37	-.23	-.41	.06	.11	.37
Lifestyle	.44	-.08	.23	.19	.15	-.25
Movie	.39	-.19	.31	-.26	-.11	-.13
News	.18	.32	-.12	.53	-.68	.01
Sports	.01	.49	-.20	.42	.60	-.11
Notes. The eigenvalues for the 10 principal components are presented in Appendix Table A1.						

To capture the composite preferences, we applied *principal component analysis* (PCA). PCA enabled us to identify and use the most significant composite preferences, without much loss of information embedded in the variance of the data. We found that components *C1* to *C6* explained more than 95% of the variance in the dataset. The remaining 10 components are likely to be noise that doesn't help with the explanation. Thus, we focused on the first 6 principal components. For example, *C1* captures a composite preference (*Education*, *Entertainment*, *Kids*, *Lifestyle*, and *Movies*). The application of PCA also addresses the issue of multicollinearity in preferences and helped us obtain stable estimates in the econometric analysis.

5.2. Econometric Results from a Bayesian Model

We estimated the model using the *Markov chain*

Monte Carlo (MCMC) *method*. We applied the *Metropolis-Hastings algorithm* [12] for the posterior distribution of model parameters. We drew 500 sample households at random for the Bayesian estimation. Table 4 shows the marginal utility and satiation.

Table 4. Marginal utility and satiation estimates

GENRE	$\overline{\beta_j}$ (SE)	δ_j (SE)
Chinese	Fixed	-0.966 (0.019)
Education	2.373 (0.154)	-0.991 (0.009)
Entertainment	1.198 (0.162)	-0.951 (0.028)
Infotainment	-0.503 (0.224)	-0.970 (0.027)
International	2.684 (0.155)	-0.997 (0.003)
Kids	0.658 (0.193)	-0.725 (0.031)
Lifestyle	-0.169 (0.144)	-0.967 (0.026)
Movie	0.320 (0.197)	-0.840 (0.045)
News	-1.416 (0.105)	-0.948 (0.042)
Sports	1.653 (0.173)	-0.450 (0.051)

Notes. For identification, we fixed the *Chinese* genre baseline utility so $\gamma_{ij} = 1.0$. ψ_{ij} and α_j in Eq. 1 were reparametrized: $\beta_{ij} = \ln(\alpha_j \psi_{ij})$ and $\delta_j = \alpha_j - 1$.

β_{ij} captures the marginal utility of household i to consume genre j when its current consumption is zero ($x_j = 0$). Our results show that, on average, the *International* genre provided a household the highest marginal utility (2.684), followed by *Education* (2.373) and *Sports* (1.653). In contrast, the *News* genre provided the lowest marginal utility.

Higher values of the $\boldsymbol{\delta_j}$ parameter indicate less satiation for the consumption of genre j. There were large differences among different genres in terms of satiation and viewing fatigue. Among the 10 genres, the *Sports*, *Kids*, and *Movies* genres did not easily satiate viewers, while they were more easily overloaded with the other genres. Although the *International* genre had the highest marginal utility, it also had the highest fatigue effect. The combined estimates of marginal utility and satiation suggest that the *Sports* genre had the highest utility for continuous consumption. Our results suggest the importance of taking into account marginal utility and satiation to design appropriate genre and viewing bundles.

We also found that the marginal utilities of the different genres exhibited large preference variances across households. (See Appendix Table B1.) This implies there were groups of households in our sample that obtained unusually high utility from consuming genre-specific programming. For such households, it was appropriate for the digital entertainment firm to adopt a *no-bundling strategy*. How might they have been able to accomplish that? By applying an *à la carte* content selection approach, the digital services provider could have guided households with specific preferences to consume higher-quality content, extracting more consumer surplus than traditional bundling alone would allow. Also, others have noted that bundling works well when consumers have a choice to buy the content separately and not just the bundle, which has the potential to create undesirable consumer responses [6].

Estimates of the household heterogeneity model are presented in Table 5. The extracted preferences are more useful than the observed demographics in statistical significance terms. We focused on the effects of the extracted preferences on the marginal utilities of the different genres. (For estimates on household demographics, see Appendix Table 2.)

Table 5. Household heterogeneity model estimates

GENRE	*C1*	C2	C3	C4	C5	C6
Chinese	Fixed					
Education	-.03 (.08)	1.02 (.11)	-.41 (.16)	-1.36 (.16)	1.42 (.18)	1.16 (.28)
Entertainment	-.07 (.10)	.78 (.13)	-.46 (.19)	-.97 (.21)	1.39 (.21)	-.18 (.37)
Infotainment	.21 (.08)	-.12 (.10)	-.24 (.16)	-.27 (.14)	-.29 (.21)	.75 (.26)
International	.38 (.07)	.81 (.09)	-.43 (.14)	-.84 (.13)	.91 (.15)	.07 (.24)
Kids	.19 (.09)	1.88 (.14)	.69 (.19)	-.92 (.22)	1.05 (.22)	-.05 (.32)
Lifestyle	-.32 (.11)	1.15 (.13)	-.62 (.20)	-1.41 (.20)	1.58 (.24)	1.05 (.36)
Movie	-.29 (.12)	.88 (.15)	-.30 (.25)	-1.23 (.23)	1.27 (.31)	.25 (.44)
News	.10 (.10)	.97 (.15)	-.61 (.21)	-1.41 (.27)	2.67 (.25)	-.39 (.34)
Sports	.03 (.10)	.67 (.12)	.02 (.17)	-1.68 (.18)	.74 (.20)	.35 (.34)

Notes. We fixed the *Chinese* genre as before.

6. Conclusion

Discovering consumer preferences is an important issue for the digital entertainment sector, and industry in general. In this research, we proposed a blended approach to bridge model-based data mining and theory-based econometric modeling. We implemented and tested our blended approach with a dataset on TV viewing for set-top box-delivered digital entertainment. Our approach was effective in extracting household preferences as a basis for designing firm-level bundling and recommendation strategies based on new knowledge that neither individual research approach would have been able to produce alone.

There are several limitations. Our observation period is short at 7 months, disallowing a time-trend assessment. Our model views households as having a household utility function. So our model is a *unitary model of the household* in which budget constraints and the demand of different household members are pooled. Future research can use the *non-unitary model of the household* [16] to allow within-household differences (e.g., husband and wife's incomes).

Acknowledgments. This research with a corporate sponsor, requiring the anonymization of all data. No

personally-identifiable information about customers has been disclosed. It was supported by the Singapore National Research Foundation under the International Research Centre @ Singapore Funding Initiative, administered by the Interactive Digital Media Programme Office (IDMPO).

References

[1] Bhat, C.R. A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. *Trans. Res. Pt. B: Method.*, 39, 2005, 679-707.

[2] Blei, D.M., Ng, A.Y., Jordon, M.I Latent Dirichlet allocation. *J. Mach. Learn.* 3, 2003, 993-1022.

[3] Bothum, D., Vollmer, C. 2016 entertainment and media industry trends. PwC, New York, NY, 2016.

[4] Chang, R.M., Ghosh P., Jung G., Kauffman R.J., Zhang P. Do household cable TV viewing patterns demonstrate efficiency and concentration? In *7th China Summer Wkshp. Inf. Mgt.*, Tianjin, China, 2013.

[5] Chang, R.M., Kauffman, R.J., Son, I. Consumer micro-behavior and TV viewership patterns: data analytics for the two-way set-top box. *Proc. 14th Intl. Conf. Elec. Comm*., ACM, New York, NY, 2012.

[6] Derdenger, T., Kumar, V. The dynamic effects of bundling as a product strategy. *Mktg. Sci.,* 32(6), 2013, 827-859.

[7] Griffiths, T.L., Steyvers, M. Finding scientific topics. *Proc. Nat. Acad. Sci.*, 101, 2004, 5228–5235.

[8] Hoang, A.P., Kauffman, R.J. Experience me! the impact of content sampling strategies on the marketing of digital goods. *Proc. 49th Hawaii Intl. Conf. Sys. Sci.*, ACM Press, NY, 2016.

[9] Kim, J., Allenby, G., Rossi P. Modeling consumer demand for variety. *Mktg. Sci.*, 21(3), 2002, 229-250.

[10] Kullbach, S. The Kullbach-Leibler distance. *Amer. Stat.*, 41(4), 1987, 340-341.

[11] Li, J., Guo, Z., Kauffman, R.J. Recovering household preferences for digital entertainment. In. *Proc. 49th Hawaii Conf. Sys. Sci.*, ACM Press, New York, NY, 2015.

[12] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(1), 1953, 1087-1092.

[13] Mohri, M., Rostamizadeh, A., Talwalkar, A. *Foundations of Machine Learning*, MIT Press, Cambridge, MA, 2012.

[14] Nielsen. Nielsen estimation 111.4 million television homes in the U.S. for the 2015-16 television season. New York, NY, 2015.

[15] PwC. Global entertainment and media outlook, 2015-2019. New York, NY, 2015.

[16] Rode, A. Literature review: non-unitary models of the household (theory and evidence). Working paper, U. California, Santa Barbara, CA, 2011.

Appendix A. Details of the LDA Model

Table A1. PCA eigenvalues

COMPO-NENTS	EIGEN-VALUES	DIFF.	PROP. VAR.	CUM. VAR.
1	4.250	2.146	0.425	0.425
2	2.104	0.964	0.210	0.635
3	1.140	0.108	0.114	0.749
4	1.032	0.335	0.103	0.853
5	0.697	0.334	0.070	0.922
6	0.363	0.168	0.036	0.959
7	0.195	0.079	0.020	0.978
8	0.116	0.029	0.012	0.990
9	0.087	0.071	0.009	0.998
10	0.016	.	0.002	1.000

Appendix B. Bayesian Estimation Model

Table B1. Covariance and correlation matrix, $\overline{\beta_j}$

	Ed	*En*	*If*	*It*	*K*	*L*	*M*	*N*	*S*
Ed	15.21 (1.16)	0.83	0.41	0.83	0.72	0.87	0.68	0.80	0.82
En	14.27 (1.15)	19.58 (1.26)	0.34	0.80	0.69	0.79	0.74	0.79	0.81
If	3.99 (1.02)	3.81 (1.11)	6.25 (0.73)	0.50	0.26	0.29	0.28	0.34	0.38
It	9.85 (1.06)	10.72 (1.13)	3.81 (0.95)	9.24 (1.05)	0.77	0.72	0.63	0.76	0.77
K	13.45 (1.24)	14.58 (1.38)	3.09 (1.04)	11.14 (1.18)	22.89 (1.84)	0.66	0.62	0.65	0.70
L	16.25 (1.25)	16.64 (1.26)	3.40 (1.15)	10.38 (1.09)	15.13 (1.28)	22.73 (1.66)	0.71	0.76	0.77
M	13.70 (1.25)	16.83 (1.21)	3.67 (1.09)	9.79 (1.17)	15.28 (1.42)	17.38 (1.23)	26.54 (2.41)	0.74	0.77
N	15.61 (1.66)	17.40 (1.79)	4.26 (1.30)	11.56 (1.47)	15.58 (1.81)	18.04 (1.82)	19.09 (1.92)	25.08 (2.75)	0.78
S	13.25 (1.24)	14.90 (1.24)	3.98 (1.21)	9.68 (1.23)	13.82 (1.46)	15.15 (1.31)	16.42 (1.29)	16.18 (1.94)	17.16 (1.38)

Notes. The *Chinese* genre is fixed as before, with similar abbreviations.

Table B2. Household heterogeneity model results

GENRE	*D1*	*D2*	*D3*	*D4*	*D5*	*D6*	*D6*
Chinese	Fixed						
Education	.96 (.36)	1.34 (.41)	.98 (.55)	.56 (.61)	.60 (.58)	.64 (.59)	.70 (.60)
Entertainment	1.20 (.42)	1.26 (.51)	1.30 (.73)	.09 (.73)	.58 (.71)	.33 (.74)	.27 (.73)
Infotainment	.28 (.34)	.84 (.46)	1.38 (.51)	1.65 (.80)	1.60 (.79)	.79 (.69)	1.50 (.75)
International	.59 (.30)	1.20 (.35)	.90 (.48)	.31 (.50)	.48 (.46)	.16 (.50)	.58 (.48)
Kids	1.19 (.44)	1.80 (.47)	.84 (.69)	.75 (.74)	.85 (.74)	.74 (.73)	.83 (.72)
Lifestyle	1.33 (.44)	2.30 (.50)	1.52 (.70)	.72 (.78)	.74 (.79)	1.36 (.75)	1.34 (.77)
Movie	1.52 (.54)	1.81 (.59)	2.33 (.80)	-.47 (.98)	-.63 (.91)	.38 (1.00)	.16 (.99)
News	1.97 (.49)	2.89 (.55)	2.07 (.72)	.57 (.74)	1.09 (.69)	1.11 (.75)	.92 (.74)
Sports	1.16 (.40)	2.21 (.47)	1.51 (.65)	.18 (.68)	.34 (.67)	.48 (.69)	.86 (.68)

Notes. The *Chinese* genre is fixed. *D1* is a dummy for *Nationality*; *D2* and *D3* are for dwelling types, and *D4* to *D6* are regions of residence.