Market Segmentation Based on Spending

STAT927 HW4

Haowei Liu

April 2021

1 Introduction

In making marketing-related decisions, we often want to identify different types of customers based on demographic characteristics and their past behaviors. In subscription business model, we may want to conduct campaigns that target customers who are mostly likely to convert. In other instances, we may need to differentiate the more loyal customers from the others, and offer specialized discount programs. The first step towards any actions would be identifying customers segments: in this analysis, I use a concise mall customer data to demonstrate how to group customers together based on their spending behaviors using Multivariate-Gaussian Mixture model. I believe this analysis paradigm can be easily extended to more complex situations where we may have more information about the consumers, or that there are more types of consumers.

2 Data Exploration

The data I will be using is the **Mall Customer Segmentation Data** from Kaggle.[1] In this dataset, we have access to 200 customers' basic demographic information as collected through membership program, in addition to spending score that quantifies each customer's spending behavior; customers with higher spending scores are those we'd like to tag as high-spenders. I recognize this dataset is rather a simplified version of consumer behavior data in reality, since it is difficult to summarize spending behaviors with a single number, and we might also have richer but messier information regarding customer characteristics. However, I believe market segmentation problems can be formulated in similar manner, where we aim to group customers based on some quantitative metrics, controlling for confounders. Therefore, here I use a simple dataset to demonstrate a potential solution a rather complex problem in hope that similar analysis can be carried on other forms of data as well.

2.1 Summary Statistics

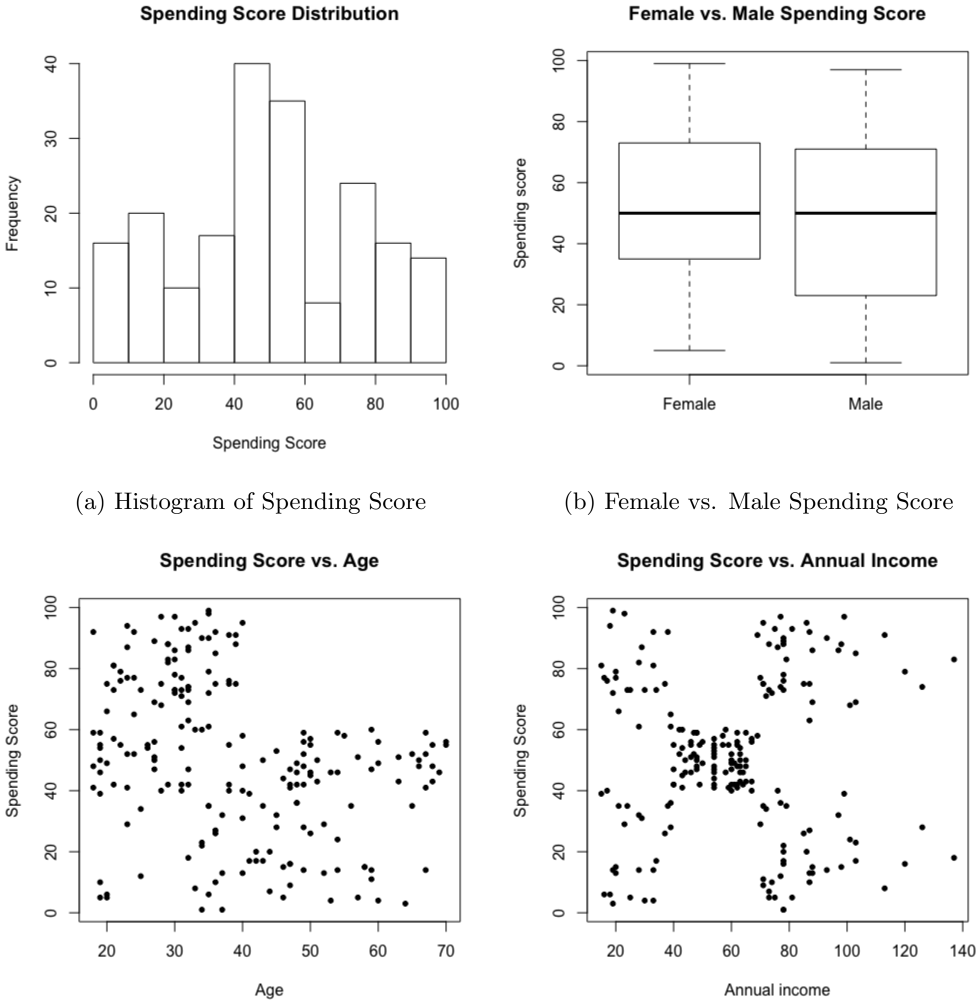
	Mean	**Median**	**Min & Max**
Male	0.44	n.a	0 & 1
Age	38.85	36	18 & 70
Annual Income	60.56	61.5	15 & 137

Table 1: Summary Statistics of Regression Variables

To understand how spending score interacts with customer characteristics, I created the following figures. First of all, we see that spending score is centered around 50, following a fairly symmetrical distribution. And from **Figure 1b**, we see no discernible difference between male and female customers. In addition, there's no clear correlation between spending and age/income; especially in **Figure 1d**, data points seem

[1] Mall Customer Segmentation Data, Vijay Choudhary, https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python.

to develop in two directions. It is my hope that a mixture model would be able to model these diverging trends by segmentation.



(a) Histogram of Spending Score

(b) Female vs. Male Spending Score

(c) Spending Score vs. Age

(d) Spending Score vs. Income

Figure 1: Summary Plots of Regression Variables

3 Model Specification

The motivation behind using mixture model is that I believe consumers can have innately different spending habit. In other words, two people with the same age, gender, and income could have different spending preferences. In this case, I will use a two-part mixture model to identify high-spender vs. low-spender.[2] Next, given consumer type, I can then use two separate linear regression to capture the relationship between spending score and personal characteristics. The coefficients of these regressors would follow two independent multivariate normal distribution, allowing for individual variance. The spending score can be expressed in the following equations.

$$Y_{ij} = X_{ij}\tilde{\beta}_j + e_{ij} \tag{1}$$

[2] Since the spending score depends on a few variables, high-spenders do no always have higher spending score than low-spenders. I use the term high-spenders to indicate people who tend to have higher spending score in general.

The outcome variable **Y** in this equation is spending score. **X** refers to independent variables, a vector consisting of gender, age, and annual income, and $\tilde{\beta}$ is the corresponding coefficient vector. e is an error term that allows for individual variance, and is of distribution $e_{ij} \sim N(0, \sigma^2)$, assuming the variance parameter is shared across group. The subscripts j refers to the type of spender each customer is, so we would have two sets of coefficients for high-spender and low-spenders.

In order to build the full probability model, I'll first specify a few additional notations and prior assumptions. I use I_i to denote customer segment, the latent variable: $I_i = 1$ means individual i is a high-spender, and $I_i = 0$ means i is a low-spender.

$$p(I_i = 1) = \alpha \tag{2}$$

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \tag{3}$$

$$p(\alpha) = Beta(1, 1) \tag{4}$$

Given the parameters specified above, the full posterior probability distribution can be written as follows:

$$\begin{aligned} p(\alpha, \tilde{\beta}, \sigma^2, I|Y) &\propto p(Y, I|\alpha, \tilde{\beta}, \sigma^2)p(\tilde{\beta}, \sigma^2)p(\alpha) \\ &\propto \frac{1}{\sigma^2}\prod_i^N \alpha^{I_i}(1-\alpha)^{1-I_i}[(\sigma^2)^{-\frac{1}{2}}exp\{\frac{-1}{2\sigma^2}(Y - X\tilde{\beta}_1)'(Y - X\tilde{\beta}_1)\}]^{I_i} \\ &\quad * \ [(\sigma^2)^{-\frac{1}{2}}exp\{\frac{-1}{2\sigma^2}(Y - X\tilde{\beta}_0)'(Y - X\tilde{\beta}_0)\}]^{1-I_i} \end{aligned} \tag{5}$$

Given this full posterior probability distribution, I can compute the posterior distribution of each parameter, and compute posterior interval using Gibbs sampler. The steps to obtain samples of relevant parameters are shown below.

1. $\mathbf{I_i}$ (latent variable of spending type):

$$p(I_i = 1|\alpha, \tilde{\beta}_1, \sigma^2, Y) \propto \frac{1}{\sigma^2}\ \alpha\ (\sigma^2)^{-\frac{1}{2}}exp\{\frac{-1}{2\sigma^2}(Y - X\tilde{\beta}_1)'(Y - X\tilde{\beta}_1)\} \tag{6}$$

$$p(I_i = 0|\alpha, \tilde{\beta}_0, \sigma^2, Y) \propto \frac{1}{\sigma^2}\ (1-\alpha)\ (\sigma^2)^{-\frac{1}{2}}exp\{\frac{-1}{2\sigma^2}(Y - X\tilde{\beta}_0)'(Y - X\tilde{\beta}_0)\} \tag{7}$$

We can draw samples of $I_i = 1$ or $I_i = 0$ given probability expressed in Equation 6 and 7.

2. α (probability of customer being high spender):

$$\begin{aligned} p(\alpha|Y, I, \tilde{\beta}, \sigma^2) &\propto \prod_i^N \alpha^{I_i}\ (1-\alpha)^{1-I_i} \\ &\propto \alpha^{\sum I_i}(1-\alpha)^{N-\sum I_i} \\ &\sim Beta(\sum I_i + 1, N - \sum I_i + 1) \end{aligned} \tag{8}$$

3. β_1 (regression coefficients of high-spenders):
Since we are only concerned with the distribution of high-spenders, we can treat this as a regression problem on customers with assigned $I_i = 1$. Thus the distribution can be computed as a multivariate normal distribution on the subset of high-spenders. Let $V_{\beta_1} = (X_1'X_1)^{-1}$, and $\hat{\beta_1} = (X_1'X_1)^{-1}X_1'Y_1$, where X_1 and Y_1 indicates customers with $I_i = 1$:

$$p(\tilde{\beta_1}|\alpha, I = 1, \sigma^2, Y) \sim MVN(\hat{\beta_1}, \sigma^2 V_{\beta_1}) \tag{9}$$

4. β_0 (regression coefficients of low-spenders):
Similar to the calculation of β_1, we can compute the probability distribution of β_2 using the subset of data with assigned $I_i = 0$. Let $V_{\beta_0} = (X_0'X_0)^{-1}$, and $\hat{\beta_0} = (X_0'X_0)^{-1}X_0'Y_0$, where X_0 and Y_0 indicates customers with $I_i = 0$:

$$p(\tilde{\beta_0}|\alpha, I = 0, \sigma^2, Y) \sim MVN(\hat{\beta_0}, \sigma^2 V_{\beta_0}) \tag{10}$$

5. σ^2 (variance term for high-spenders and low-spenders):
   As mentioned earlier, two regressions are assumed to have the same variance term, and the posterior probabiltiy can be expressed as an inverse Gamma distribution. To compute sum of squared errors, we can simply divide the data given value of **I**, and compute the predicted value using corresponding β:

$$p(\sigma^2|\tilde{\beta}, X, Y, I) \sim InvGamma(\frac{1}{2}(N-p),\ \frac{1}{2}\sum_{i:I=1}^{N}(Y_i - X\tilde{\beta_1})^2 + \frac{1}{2}\sum_{i:I=0}^{N}(Y_i - X\tilde{\beta_0})^2) \tag{11}$$

These five-step sampling is then repeated to obtain adequate samples for the purpose of removing burnout samples and autocorrelation. The exact process is described in the next section.

4 Result and Evaluation

To obtain independent samples from Gibbs sampling, I ran two chains separately with different starting values. I then checked for convergence between these two chains by visualizing the changes in parameters; upon examination, I found that the data converge fairly quickly, and threw away the first 100 samples as burnout. Then I checked for auto-correlation, and found that by keeping every 20 samples, I end up with data with little auto-correlation. By combining these two chains, I end up with 4000 number of independent samples for all the parameters described above.[3] The 95% posterior confidence interval for these parameters are shown in **Table 2**.

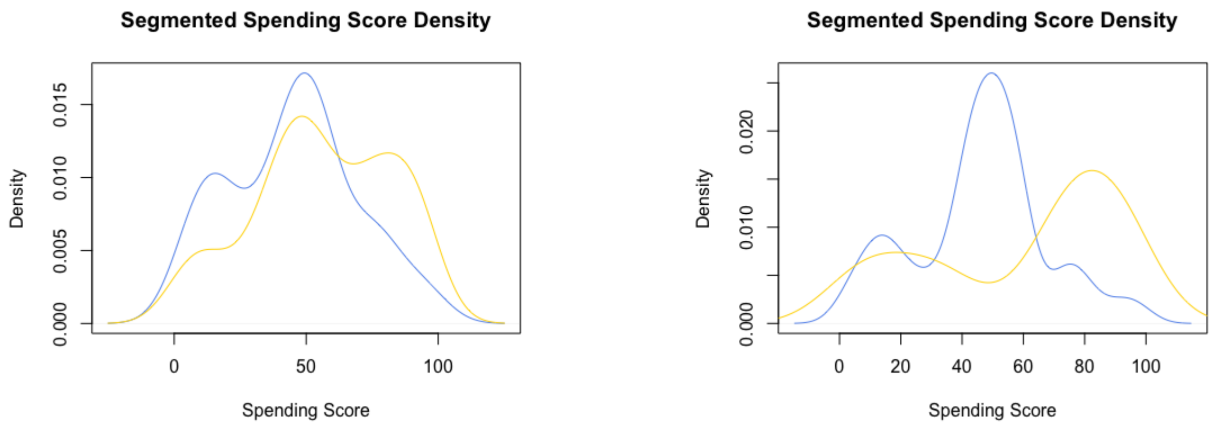
	High-spenders	**Low-spenders**
Alpha	[0.418, 0.582]	
Sigma_squared	[161.72, 262.61]	
Intercept	[9.22, 45.07]	[87.81, 107.91]
Male	[-4.04, 9.23]	[-12.90, -0.26]
Age	[-0.71, -0.09]	[-0.36, 0.08]
Annual Income	[0.58, 0.86]	[-0.87, -0.62]

Table 2: 95% Posterior Interval for Model Parameters

To interpret the result of this model, we can first take a look at the value of α, i.e. what percentage of customers are considered high-spenders. It seems that the segmentation is quite balanced between groups. As for intercepts, the two groups seem to differ quite a lot in the positive range; this can be partly explained by the difference signs in income coefficients, as the 95% interval is strictly positive for high-spenders, and negative for low-spenders. The gender indicator is not significant for high-spenders but slightly negative for low-spenders, which might be consistent with commonly perceived spending habit difference between genders. Finally we see that age coefficient is negative for high-spenders, along with positive income effect, I suspect there could be a correlation between age and income that could be captured if we were to include an interaction term between age and annual income in future analysis.

In order to understand how this mixture model is capturing relationship between personal characteristics and spending outcome, I visualize the breakdown between spending distributions in the following density plot. It should be noted here that I use sampled group assignment, variable I_i, to determine which group a customer belongs to; in other words, each customer is tagged as high-spender or low-spender depending on the expected value of I_i, calculated as $\frac{\sum_i^M I_i}{M}$, where M is the number of total samples. I choose 0.5 as the determining threshold. The blue line in **Figure 2a** is the density of low-spenders' spending score, and the gold line is that of high-spenders'. It's quite straightforward from the graph that the distribution of high-spenders' spending score is skewed more towards higher end. It does appear that even without enforcing any segmentation requirement, our mixture model has identified two groups with somewhat different spending scores. And

[3] Since the assignment of groups is agnostic of the significance in group 1 and 0, before combining chains, I made sure the model parameters are consistent within group.



(a) High vs. Low Spender Density Plot: cutoff = 0.5

(b) High vs. Low Spender Density Plot: cutoff = 0.8

Figure 2: Spending Density Plot after Segmentation

it can also be shown from **Figure 2b** that as we increase the threshold of high-spender assignment, the difference between distributions becomes more pronounced. However, a caveat in interpreting this result is that since we dealing with multiple covariates, spending score is ultimately affected by each and every independent variables.

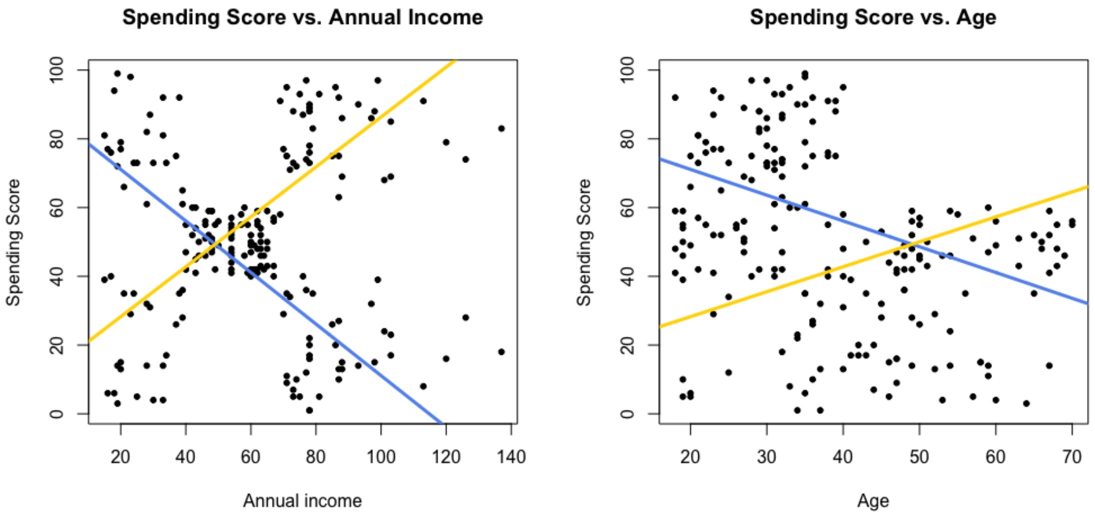
From **Table 2**, we see that the income coefficients are quite different for high and low spenders, one being in the positive range and the other being negative. This would mean that there are two types of consumers: one would increase spending as their income increases, and the other would do the opposite, holding all other things constant. We can plot these relationships by computing the spending score for an average aged male for varying level of income in **Figure 3a**. Again, blue indicates low-spenders and gold line represents high-spenders. It can be shown that a two-part mixture models successfully captures the two diverging trend in spending habit. For high-spenders, income has a positive effect on their spending behaviors. Thus in application, a promotional membership program may incentivize these customers to spend more, while the same may not necessarily be true for low-spenders. Similarly, for spending score and age, I computed the spending score for a male with average income of different ages and plotted in **Figure 3b**. We observe similar patterns: high-spenders have positive trend and low-spenders have a negative trend. When we compare these two trendlines with true data, we see that by allowing for two groups, the points can be traced better with two directions.

Additionally we can quantify the predictive ability of our model by calculating the absolute difference between true spending score, and predicting spending score given spender group assignment and average value of sampled parameters.[4] We can also assess whether a mixture model performs better than a single linear regression model by comparing the resulting absolute error. **Table 3** shows that the mean error from mixture model is much smaller than that of a single linear regression. This indicates that using a mixture model is better at depicting the underlying patterns than assuming all customers have the same spending preference.

	Mixture Model			**Linear Regression**
	High-spenders	Low-spenders	All customers	
Mean Abs. Error	9.57	9.07	9.31	20.37

Table 3: Mean Absolute Error of Model Prediction

[4] Similar to methods discussed earlier, I take the average of sampled group assignment, and use 0.5 as a separating threshold between groups.



(a) Fitted lines of spending scores over annual income

(b) Fitted lines of spending scores over age

Figure 3: Fitted relationship between spending scores and customer characteristics after segmentation

5 Discussion

I have thus shown how to perform customer segmentation using a multivariate-Gaussian model, and given this specific dataset, mixture model is better at capturing trends in customer behaviors. With this segmentation information, we can potentially implement more targeted promotion and advertisement to existing customers. This model can also provide some insights on expected value of a new customer, as we can generate a predictive spending score interval given mixture probability and group-specific parameters; however, the application of this model to new customers is somewhat limited in my opinion. As shown above, the mixture parameter α centers around 0.5, which means we would be essentially making a random guess on whether a customer is a high-spender or low-spender.

On the other hand, the current model is based on an aggregate spending score that only provides a static view on consumer behavior; this doesn't allow us to track changes in consumer spending over time. If we were to have consumer data on a more granular level, a time-series data for example, it's possible to build a sequential model that monitors spending and other behavior over time, and make predictions on future behaviors, which can then be used for marketing insights. In sum, I believe there are many interesting extensions to the current model in achieving market segmentation goal given data of bigger scope.