37902 Foundations of Advanced Quantitative Marketing

Session V

Topics

1. Random coefficients model with continuous heterogeneity distribution
2. Simplifying the covariance matrix

1 Continuous heterogeneity distribution

The discrete or LCM model discussed above has many appealing features - easy to interpret as it has the segmentation connotation, it is a flexible distribution, etc. But there are several issues:

1) Number of parameters explodes as number of segments goes up.

2) Likelihood is ill beheaved especially when number of segments goes up.

3) It does not capture "tail" behavior very well. Now households with very small or very large parameter values are not very well represented in the estimates.

For these reasons researchers may prefer specifying a parametric form for the distribution. This simplifies the problem in one sense: we only need to estimate the parameter of the distribution. Further it keeps a lid on the total number of parameters. However, estimation could be a lot more complicated.

The typical assumption for the heterogeneity distribution is the multivariate normal. What does that mean? Recall from session 3 notes that the parametric vector for household i was represented as

$$\Theta_i = \{\alpha_{ij},\ \ j = 1, ..., J-1,\ \beta_i\}$$

The discrete distribution approach said that Θ_i can only take S values, $\Theta_1, \Theta_2, ..., \Theta_S$. Instead here we will assume that Θ_i comes from a multivariate normal distribution, i.e., one that has $(J-1+M)$ "variates", where M is the dimensionality of the β_i vector. What does this mean? It says that if I could estimate a Θ_i for each household and draw the histogram of these $\Theta_i's$, it would look like a giant normal distribution in $(J-1+M)$ dimensions!

Problem is we cannot estimate a separate Θ_i (why?). So we will end up estimating only the mean and covariance matrix of the multivariate normal (MVN) distribution, Θ and $\sum$.

So what is new here is estimating the $\sum$. Let us consider the yogurt example again. In this case Θ_i is 5-dimensional: $\{\alpha_1^i, \alpha_2^i, \alpha_3^i, \beta_f^i, \beta_p^i\}$. So the mean vector Θ will be $\{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p\}$. This represents the average Θ_i across all households. The covariance matrix $\sum$ looks like:

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{1f} & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{2f} & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{3f} & \sigma_{3p} \\ \sigma_{f1} & \sigma_{f2} & \sigma_{f3} & \sigma_{ff} & \sigma_{fp} \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \sigma_{pf} & \sigma_{pp} \end{bmatrix}$$

So it appears that there are 5×5 parameters to be estimated. However, since the matrix is symmetric $\sigma_{12} = \sigma_{21}$, etc. so we have only $\frac{5\times6}{2} = 15$ parameters to estimate. In general with M-variates we have $\frac{M(M+1)}{2}$ parameters. Let us call them $\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{1f}, \sigma_{1p}, \sigma_{22}, \sigma_{23}, \sigma_{2f}, \sigma_{2p}, \sigma_{33}, \sigma_{3f}, \sigma_{fp}, \sigma_{pp}$. Next we need to consider the main property of a covariance matrix, i.e., that it is positive-definite. How can we impose this condition in the estimation of the above parameters? Fortunately, $\sum$ is like a "square" so it has a square root. This is known as the Cholesky decomposition of $\sum$. IOW,

$$\sum = \Gamma'\Gamma \text{ where } \Gamma \text{ is an upper-triangular matrix.}$$

$$\Gamma = \begin{pmatrix} a & b & c & d & e \\ 0 & f & g & h & i \\ 0 & 0 & j & k & l \\ 0 & 0 & 0 & m & n \\ 0 & 0 & 0 & 0 & p \end{pmatrix}$$

Where $a \sim p$ can be any number, positive or negative. So

$$\Gamma'\Gamma = \begin{pmatrix} a^2 & ab & ac & ad & ae \\ ab & b^2+f^2 & bc+fg & bd+hf & be+fi \\ ac & bc+gf & c^2+g^2+j^2 & cd+gh+jk & ce+gi+jl \\ ad & bd+hf & cd+gh+jk & d^2+h^2+k^2+m^2 & de+hi+kl+mn \\ ae & be+fi & ce+gi+jl & de+hi+kl+mn & e^2+i^2+l^2+n^2+p^2 \end{pmatrix}$$

It is easy to verify that $\Gamma'\Gamma$ is indeed positive definite. Further since Γ has the same number of parameters as $\sum$, we preserve full flexibility of the coveriance matrix parameters in the estimation.

Before we return to the estimation of the RC model, we look at how we would make draws from a MVN distribution that has mean Θ and covariance $\sum$.

- Step 1: Make D draws for each variate of the MVN from a standard normal distribution N(0, 1).
- Step 2: Arrange the draws as a $D \times M$ matrix where M is the dimensionality of $\sum$ (5 in the yogurt example). This means we have $D \times M$ draws from N(0, 1). Call this matrix Q.
- Step 3: Post-multiply Q by Γ, i.e., multiply a $D \times M$ matrix by an $M \times M$ matrix to get another $D \times M$ matrix. The $D \times M$ matrix that results has an MVN distribution$\sim MVN(0, \sum)$
- Step 4: Add the vector Θ', the $1 \times M$ vector of mean parameters to the $D \times M$ matrix of draws, i.e., $(\Theta' + Q\Gamma)$. This new matrix has draws from an MVN distribution with mean Θ and covariance matrix $\sum$. Note: I am assuming that the mean vector Θ is an $M \times 1$ vector, hence the need to transpose before adding to $Q\Gamma$.

Now that we know how to make draws from MVN distribution, let us return to estimating the RC model. Recall what we are trying to do. We want to allow each household to have a vector Θ_i such that the distribution of Θ_i across households follows $MVN(\Theta, \sum)$. Now, each Θ_i is a draw from $MVN(\Theta, \sum)$. Problem is we don't know each household's specific draw. What does this sound like? Just like the segments - a priori we did not know which latent class a household belongs to. Similarly, if we can make D draws from $MVN(\Theta, \sum)$, we don't know which specific draw is most likely the one for a given household. But we can take the expectation of a household's likelihood function across these draws just as we computed the expected likelihood for a household across S segments. So the steps in the estimation are:

- Step 1: Pick starting values for Θ and Γ. These would be:
  - $\{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p, a, b, c, d, e, f, g, h, i, j, k, l, m, n, p\}$
- Step 2: Make D draws each for M univariate standard normal variables. Construct $Q = D \times M$ matrix.
- Step 3: Post-multiply Q by Γ (recall Γ has all the parameters $\{a, b, c, d, ..., p\}$.). Call this Q_G.

- Step 4: Add $\{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p\}$ to Q_G. This gives us draws from $MVN(\Theta, \sum)$. Call this matrix Q_{GT}. This matrix also has dimensionality $D \times M$.
- Step 5: Recognize that each column of Q_{GT}, $d = 1, ..., D$ is just like the parameters for each of the S segments in the latent class model - there we had S sets of starting parameters. Now we have D sets of starting parameters.
- Step 6: Compute the likelihood for each household i, for each draw d, i.e. $L_{i|d}$ (similar to $L_{i|s}$ we computed previously).
- Step 7: Compute the average likelihood $L_i = \frac{1}{D}\sum_{d=1}^{D} L_{i|d}$ (note that we don't have "segment sizes" as the draws themselves reflect the probabilities as they come from a normal distribution).
- Step 8: Take the log of $L_i = \frac{1}{D}\sum_{d=1}^{D} L_{i|d}$ from the step above and add the likelihoods across the N households. This gives the sample log-likelihood = LL.
- Step 9: Choose parameters Θ, Γ to maximize the log-likelihood LL.

The LL is written as :

$$\begin{aligned} LL &= \ln\left[\prod_{i=1}^{N}\left\{\int_{\Theta_i}\left(\prod_{t=1}^{T_i}\left(\prod_{j=1}^{J} P_{ijt|\Theta_i}^{\delta_{ijt}}\right)\right) f(\Theta_i)d\Theta_i\right\}\right] \\ &= \sum_{i=1}^{N}\ln\left\{\int_{\Theta_i}\left(\prod_{t=1}^{T_i}\left(\prod_{j=1}^{J} P_{ijt|\Theta_i}^{\delta_{ijt}}\right)\right) f(\Theta_i)d\Theta_i\right\} \\ &= \sum_{i=1}^{N}\ln\left\{\frac{1}{D}\sum_{d=1}^{D}\left[\prod_{t=1}^{T_i}\left(\prod_{j=1}^{J} P_{ijt|\Theta_d}^{\delta_{ijt}}\right)\right]\right\} \end{aligned}$$

2 Simplifying the covariance matrix

Recall from the RC model that we are trying to estimate the following

$$U_{ijt} = \alpha_{ij} + X_{jt}\beta_i + \epsilon_{ijt}$$

i: household, j: brand, t: purchase occasion

$\Theta_i = \{\alpha_{ij}, \beta_i, j = 1, ..., J-1\}$ is the vector of household $i's$ parameters.

We assume $\Theta_i \sim MVN(\Theta, \sum)$, where unknown parameters are the mean vector $\Theta = \{\alpha_j, \beta, j = 1, ..., J-1\}$, and Γ, where $\sum = \Gamma'\Gamma$ and Γ is the upper triangular matrix.

Note that the number of unknown parameters is $\Gamma = \frac{R(R+1)}{2}$, where R is the dimensionality of Θ (in the yogurt case $\Theta = \{\alpha_1, \alpha_2, \alpha_3, \beta_f, \beta_p\}$, so $R = 5$. So Γ has $\frac{5*6}{2} = 15$ parameters).

Now suppose that we had 20 brands. Their own Γ would have $\frac{22*23}{2} = 253$ parameters! This could be a colossal pain to estimate. So what do we do? There are 2 common approaches.

a) **The characteristics approach**

Suppose the 20 brands we wanted to estimate were the following

Yoplait Strawberry Lowfat	Dannon Strawberry Lowfat
Yoplait Raspberry Lowfat	Dannon Raspberry Lowfat
Yoplait Strawberry Nonfat	Dannon Strawberry Nonfat
Yoplait Raspberry Nonfat	Dannon Raspberry Nonfat
Yoplait Plain Lowfat	Dannon Plain Lowfat
Yoplait Plain Nonfat	Dannon Plain Nonfat
Weight Wathers Strawberry Lowfat	Weight Watchers Plain Lowfat
Weight Watchers Raspberry Lowfat	Weight Watchers Plain Nonfat
Weight Watchers Strawberry Nonfat	Hiland Plain Lowfat
Weight Watchers Raspberry Nonfat	Hiland Plain Nonfat

Consumers now choose between these 20 brands. The characteristics approach essentially decomposes the preference for a given product, say yoplait strawberry low-fact into its three component attributes:

1) Brand with levels {Yoplait, Dannon, Hiland, Weight Watchers}

2) Flavor with levels {Strawberry, Raspberry, Plain}

3) Fat content {Low-fat, Nonfat}

So

$$\alpha_{i,Y,S,L} = \alpha_{i,Y} + \alpha_{i,S} + \alpha_{i,LF}$$

Now we can assume that each of $\alpha_{i,Y}$, $\alpha_{i,S,}$ and $\alpha_{i,LF}$ (and the corresponding $\alpha_{i,D}$, $\alpha_{i,H}$, $\alpha_{i,W}$, $\alpha_{i,R}$, $\alpha_{i,P}$, $\alpha_{i,NF}$) has a univariate normal distribution with means $\alpha_Y, \alpha_S,$ and α_{LF} and corresponding standard deviations σ_Y, σ, and σ_{LF}. For identification we set $\alpha_W = \alpha_P = \alpha_{NF} = 0$ and $\sigma_W = \sigma_P = \sigma_{NF} = 0$. In effect we will have 6 mean parameters and 6 variance parameters to estimate. With this setup we have dramatically lowered the number of parameters to be estimated.

- **Determining the coefficients**

Let's consider two brands {Y, S, LF} and {D, R, LF}. Their mean preferences are $\{\alpha_Y + \alpha_S + \alpha_{LF}\}$ and $\{\alpha_D + \alpha_R + \alpha_{LF}\}$. The variances of these preferences are $(\sigma_Y^2 + \sigma_S^2 + \sigma_{LF}^2)$ and $(\sigma_D^2 + \sigma_R^2 + \sigma_{LF}^2)$. The covariance in these preferences is σ_{LF}^2 (since LF is the only common characteristic between them).

So generalizing and using the terminology we used in class:

$$\alpha_{ij} = \sum_{c=1}^{C} \sum_{l_c=1}^{L_C} \alpha_{ic_{l_c}} I_{jc_{l_c}}$$

where $I_{jc_{l_c}} = \begin{cases} 1 & if\ brand\ j\ has\ attribute\ level\ l_c\ in\ attribute\ c \\ 0 & otherwise \end{cases}$

Benefit: Simple, easy to understand, useful prediction tool for line extensions.

Limitation: Assume additive separability - if Yoplait's Greek yogurt tastes different from Dannon's Greek yogurt, this won't be reflected in the model. Also all correlations in preferences are constrained to be positive. (Why?)

b) The factor-analytic approach

Instead of estimating the 20×20 covariance matrix of preferences, we claim that the covariance can be adequately represented by a small number of factors F, where $F << \#$ brands in the analysis. Now we project the covariance matrix down to a lower dimensional space F. So

$$\Sigma = A\omega\omega' A'$$

Where $A = R \times F$ matrix of locations; $\omega = F \times 1$ vector of attribute weights.

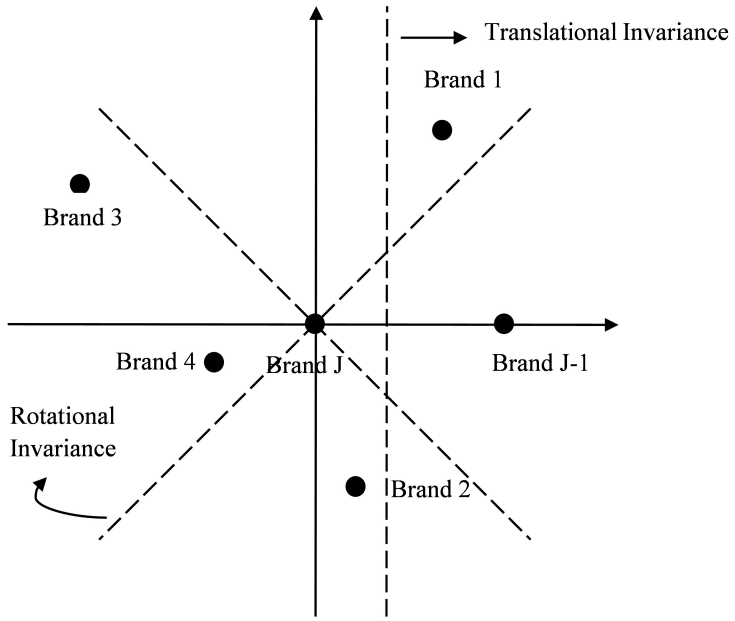
Assumption: All consumers in the market agree on the locations but disagree on how important each factor is to them. So ω is specific to i (household), and A is common across household.

For a 2-Dimensional map, $F = 2$, the unknown parameters would be $2 \times R$, i.e., we need to estimate $2R$ parameters.

ω_i for household i represents an $F \times 1$ dimensional vector of univariate standard normal draws. So instead of drawing R[1] univariate normals and then post multiplying by the Chokesky factor to get an MVN draw $\sim MVN(0, \Sigma)$, we now make F univariate normal draws and pre-multiply them with the matrix A to get the draws from $\sim MVN(0, \Sigma)$. As long as $F << R$ we will end up saving a ton of parameters to estimate. In the estimation we usually start with $F = 1$, $F = 2$, etc. and proceed till BIC increases.

- **Identification**

One of the nice things about the factor structure is that the location matrix A can be displayed in an $F-$dimensional plot. So if $F = 2$, we can create a map such as this one:



Note that we have placed Brand J at the origin. We do this because if you recall, in the RC model, although there were 4 brands, only 3 of them figured in the covariance matrix. This is the identical restriction. We call this restriction one that ensures translational invariance of the map. The idea is that since relative locations do not change if I moved the map back and forth, I need something to keep it from going back and forth.

Also notice from the picture that Brand $(J-1)$ has been constrained to lie along the X-axis. This ensures rotational invariance. What does this mean? The idea is that if I rotated the map, it would not affect the

[1] I have used D previously to denote number of draws in the RC model.

relative locations so I don't want the A parameters to constantly change in my estimation in a way that keeps relative locations fixed. To fix this issue I place $J-1$ on X-axis.

There is a 3^{rd} constraint as well, i.e., scale invariance. The idea is that if I multiply (i.e., scale) A by 10 and divide (i.e., scale) ω by $\frac{1}{10}$, my covariance matrix will be unaffected. To avoid this from happening in the estimation, we fix the variance of ω at 1. This is the reason why when drawing ω we made standard normal draws, i.e., draws that have unit variance.

Okay, to actually do the estimation. Let us say you

1) Make $D = 10 \times 2$ univariate standard normal draws. Call this matrix ω.
2) Then pick starting values for the A matrix of dimension $(J-1) \times 2 - 1$ (assume for now that Σ only applies to the preferences and not to the price and feature coefficients).
3) Multiply A by ω' to get a $(J-1) \times 10$ vector of draws. These now correspond to 10 draws from MVN distribution with distribution $MVN(0, \Sigma = A\omega'\omega A')$.
4) Pick starting values for the mean vector of preference and price and feature sensitivity $= \Theta$.
5) Add the first $(J-1)$ elements of Θ to $A\omega'$. This will give draws from an MVN with mean Θ_{J-1} and covariance matrix $A\omega'\omega A'$.
6) Now proceed as we did for the RC model with D draws.
7) Increase the number of factors till BIC stops you.
8) For identification, $(J-1) \times F - (F-1) \leq \frac{J(J-1)}{2}$.