Absolute Measures of Fit in Latent Profile Analysis and Finite Mixture Models

Lecture 15

April 13, 2006

Clustering and Classification

Today’s Lecture

- Model fit assessment in Finite Mixture Models (as realized through LPA).
- Absolute versions of model fit for mixtures of continuous distributions.
- Why absolute measures of fit matter.

Finite Mixture Models

- Recall from last time that we stated that a finite mixture model expresses the distribution of **X** as a function of the sum of weighted distribution likelihoods:

$$f(\mathbf{X}) = \sum_{g=1}^{G} \eta_g f(\mathbf{X}|g)$$

- We are now ready to construct the LPA model likelihood.
- Here, we say that the conditional distribution of **X** given g is a sequence of independent normally distributed variables.

Latent Class Analysis as a FMM

Using some notation of Bartholomew and Knott, a latent profile model for the response vector of p variables ($i = 1, \ldots, p$) with K classes ($j = 1, \ldots, K$):

$$f(\mathbf{x}) = \sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(\frac{-(x_i - \mu_{ij})^2}{\sigma_{ij}^2}\right)$$

- η_j is the probability that any individual is a member of class j (must sum to one).
- x_i is the observed response to variable i.
- μ_{ij} is the mean for variable i for an individual from class j.
- σ_{ij}^2 is the variance for variable i for an individual from class j.

MVN

- The multivariate normal distribution function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

- The mean vector is $\boldsymbol{\mu}$.
- The covariance matrix is $\mathbf{\Sigma}$.
- Standard notation for multivariate normal distributions is $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.
- Visualizing the MVN is difficult for more than two dimensions, so I will demonstrate some plots with two variables - the bivariate normal distribution.

Expressing LPA with MVN

- Recall the LPA model has the strict assumption of local independence of variables given class.

- So, for each class j, we estimate a mean vector, $\boldsymbol{\mu}_j$, and a ***diagonal*** covariance matrix, $\boldsymbol{\Sigma}_j$:

$$\boldsymbol{\mu}_j = \begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{nj} \end{bmatrix} \qquad \boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma^2_{11} & 0 & \dots & 0 \\ 0 & \sigma^2_{21} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2_{ij} \end{bmatrix}$$

- We can then reexpress our LPA model by the MVN density (this will follow us throughout the remainder of the mixtures of MVN distributions).

Latent Class Analysis as a FMM

Using some notation of Bartholomew and Knott, a latent profile model for the response vector of p variables ($i = 1, \ldots, p$) with K classes ($j = 1, \ldots, K$):

$$f(\mathbf{x}) = \sum_{j=1}^{K} \eta_j \left[\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_j|^{1/2}} \exp\left(-(\mathbf{x} - \boldsymbol{\mu}_j)\mathbf{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)/2\right) \right]$$

- η_j is the probability that any individual is a member of class j (must sum to one).
- x_i is the observed response to variable i.
- $\boldsymbol{\mu}_j$ is the mean vector for class j.
- $\mathbf{\Sigma}_j$ is the ***diagonal*** covariance matrix for class j - implying conditional independence of variables.

LPA Example

- To illustrate the process of LPA, consider an example using Fisher's Iris data.
- The Mplus code is found on the next few slides.
- We will use the Plot command to look at our results.
- Also, this time we will try fitting multiple classes to see if our results change from time to time, and how the fit statistics look for each type of solution.
- Specifically, we will compare a two-class solution to a three-class solution (the correct one) and a 4-class solution.

```
title:
    2-Class Latent Profile Analysis
    of Fisher’s Iris Data;
data:
    file=iris.dat;
variable:
    names=x1-x4;
    classes=c(2);
analysis:
    type=mixture;
model:
%OVERALL%
%C#1%
x1-x4;
%C#2%
x1-x4;

OUTPUT:
    TECH1 TECH5 TECH8;
PLOT:
    TYPE=PLOT3;
    SERIES IS x1(1) x2(2) x3(3) x4(4);

SAVEDATA:
    FILE IS myfile2c.dat;
    SAVE = CPROBABILITIES;
```

```
title:
    3-Class Latent Profile Analysis
    of Fisher’s Iris Data;
data:
    file=iris.dat;
variable:
    names=x1-x4;
    classes=c(3);
analysis:
    type=mixture;
model:
%OVERALL%

%C#1%
x1-x4;
%C#2%
x1-x4;
%C#3%
x1-x4;

OUTPUT:
    TECH1 TECH5 TECH8;
PLOT:
    TYPE=PLOT3;
    SERIES IS x1(1) x2(2) x3(3) x4(4);

SAVEDATA:
    FILE IS myfile3c.dat;
    SAVE = CPROBABILITIES;
```

```
title:
    4-Class Latent Profile Analysis
    of Fisher's Iris Data;
data:
    file=iris.dat;
variable:
    names=x1-x4;
    classes=c(4);
analysis:
    type=mixture;
model:
%OVERALL%

%C#1%
x1-x4;
%C#2%
x1-x4;
%C#3%
x1-x4;
%C#4%
x1-x4;

OUTPUT:
    TECH1 TECH5 TECH8;
PLOT:
    TYPE=PLOT3;
    SERIES IS x1(1) x2(2) x3(3) x4(4);

SAVEDATA:
    FILE IS myfile4c.dat;
    SAVE = CPROBABILITIES;
```

Model Results

- The table below shows the results of our models in for each class solution:

Model	Parameters	Log L	AIC	BIC	Entropy
2-class	17	-386.185	806.371	857.551	1.000
3-class	26	-307.178	666.355	744.632	0.948
4-class	35	-264.848	599.695	705.067	0.948

- Based on AIC and BIC, we would choose the 4-class solution (and probably should try a 5-class model).

- Note that by adding multiple starting points, the 3-class and 4-class solutions started to demonstrate problems with:

  - Convergence in some iterations.

  - Multiple modes - something to think about!

- Any guesses as to why these problems didn't show up in the two-class solution?

More Fit Info Needed?

- The model fit section just discussed is often where many researchers stop in their evaluation of a LPA or FMM solution.
- But do we really know whether what we did resembles anything about the nature of our data?
- In other methods for analysis, for instance Structural Equation Modeling, we are very concerned that our model parameters resemble the observed characteristics of the data.
- For instance, the discrepancy between the observed covariance matrix and estimated covariance matrix is used in several goodness-of-fit indices.
- Well, similar measures can be constructed in FMM.

What Can We Do?

- We can begin look at our distributional assumptions and do some bivariate plots featuring confidence regions.
- We can also use the estimated parameters of our solution to "predict" what the moments of the variables should be like:
  - The mean for each item (we will not find much variation here regardless of model).
  - The covariance matrix for all pairs of items (this would be analogous to what we do in SEM).
  - Both of these can be done either by statistical properties of the models (hard sometimes) or by simulation (too easy sometimes).
- We can also look at the proportion of observations we would expect to classify correctly for each solution (although this is somewhat problematic).

Confidence Regions

- Just as with univariate statistics, we can construct "confidence intervals" for the mean vector for multivariate inference.
- These "intervals" are no longer for a single number, but for a set of numbers contained by the mean vector.
- The term Confidence Region is used to describe the multivariate confidence intervals.
- In general, a $100 \times (1-\alpha)\%$ confidence region for the mean vector of a p-dimensional normal distribution is the ellipsoid determined by all $\boldsymbol{\mu}$ such that:

$$n_j \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_j\right)' \boldsymbol{\Sigma}_j^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_j\right) = \frac{p(n_j - 1)}{(n_j - p)} F_{p, n_j - p}(\alpha)$$

Building CRs - Population

- To build confidence regions, recall our last lecture about the multivariate normal distribution...

  Specifically:

  $$(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \chi^2_p(\alpha)$$

  provides the confidence region containing $1-\alpha$ of the probability mass of the MVN distribution.

- We then calculated the axes of the ellipsoid by computing the eigenvalues and eigenvectors of the covariance matrix Σ:

  Specifically:

  $$(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

  has ellipsoids centered at $\boldsymbol{\mu}$, and has axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$.

Building CRs - Sample

- A similar function is used to develop the confidence region for the multivariate mean vector based on the sample mean $(\bar{\mathbf{x}})$ and covariance matrix $(\bar{\mathbf{S}})$.

- Note that because we are taking a sample rather than the population, the distribution of the squared statistical distance is no longer $\chi^2_p(\alpha)$ but rather $\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)$

- This means that the confidence region is centered at $(\bar{\mathbf{x}})$, and has axes $\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{(n-p)}F_{p,n-p}(\alpha)}\mathbf{e}_i$.

LPA CRs

- Let's assume, for simplicity (because n_j is only estimated), that the number of observations in each class is infinite.
- We could then draw ellipses for each class and see how they relate to our data.
- In this case, we have diagonal matrices for our covariance matrices within class.
- Our ellipses are somewhat difficult to draw in R.
- Difficulty shouldn't stop you from doing so, however.

LPA Simulation

- Given the difficulty in getting some plots to work, we can turn to simulation to achieve our fit evaluation objectives.
- It is here we will often see discrepancies between the model estimates and the data.
- Simulation is considerably easier than finding multivariate confidence ellipses, and can often be done much quicker.
- To demonstrate, look at the two text files with this week's lecture.

Final Thought

- Absolute measures of fit are available in FMM and will provide you with understandable and interpretable numbers and plots.
- The difficulty in evaluating model fit should not make you run from doing so (I suggest this to get you ready to write and review papers).
- Because of the complexity in the modeling aspects of FMM, people often forget about absolute measures of fit - which is a bad thing to do.
- When using absolute measures, you could be simply looking at two poorly-fitting models.

Next Time

- Estimation week begins on Tuesday - consider me one happy person.
- Our next class:
  - Marginal ML estimation of FMM.
  - Estimation topics in general.