Mixture and Latent Class Analysis

Lecture 23

November 28, 2012

PSYC 943 (930) Foundations

of Multivariate Modeling

Lecture Outline

- Latent Class Analysis (LCA)
  - Underlying theory (general contexts)
  - Example analysis (and how to get estimates)
  - Interpretation of model parameters
  - Investigating model fit
- Extensions of the Technique:
  - Latent Profile Analysis (LPA)

Clusters Versus Classes

- When a researcher mentions they are going to be using cluster analysis, they are most likely referring to one of the following:
  - K-means clustering
  - Hierarchical clustering using distance methods
  - Discriminant analysis
  - Taxometrics
- Much less often, latent class analysis is included in the group
  - Although it too is useful for detecting clusters of observations
- For today's lecture, we will consider clusters to be synonymous with classes

LCA Versus Other Methods

- Although I am using the terms classes and clusters synonymously, the general approach of LCA differs from that of the other methods previously discussed
- LCA is a model-based method for clustering (or classification)
  - LCA fits a statistical model to the data in an attempt to determine classes
- The other methods listed on the previous slide do not explicitly state a statistical model
- By being model based, we are making very explicit assumptions about our data
  - Assumptions that can be tested

Latent Class Analysis

LCA Introduction

- Latent class models are commonly attributed to Lazarsfeld and Henry (1968)
- The final number of classes is not usually predetermined prior to analysis with LCA
  - The number of classes is determined through comparison of posterior fit statistics
  - The characteristics of each class is also determined following the analysis
  - Similar to K-means and hierarchical clustering techniques in this respect

Variable Types Used in LCA

- As it was originally conceived, LCA is an analysis that uses:
  - A set of binary-outcome variables - values coded as zero or one Examples include:
    - Test items - scored correct or incorrect
    - True/false questions
    - Gender
    - Anything else that has two possible outcomes

LCA Process

- For a specified number of classes, LCA attempts to:
  - For each class, estimate the probability that each variable is equal to one
  - Estimate the probability that each observation falls into each class
    - For each observation, the sum of these probabilities across classes equals one
    - This is different from K-means where an observation is a member of a class with certainty
  - Across all observations, estimate the probability that *any* observation falls into a class

LCA Estimation

- Estimation of LCA model parameters can be more complicated than other clustering methods:
  - In hierarchical clustering, a search process is used with new distance matrices being created for each step
  - K-means uses more of a brute-force approach - trying multiple random starting points then shifting cases between the different clusters until each is no longer shifted
  - Both methods relied on distance metrics to find clustering solutions
- LCA estimation uses distributional assumptions to find classes
- The distributional assumptions provide the measure of "distance" in LCA

LCA Distributional Assumptions

- Because (for today) we have discussed LCA with binary-outcome variables, the distributional assumptions of LCA must use a binary-outcome distribution
- Within each latent class, the variables are assumed to:
  - Be independent
  - Be distributed marginally as Bernoulli:
    - The Bernoulli distribution states:

$$f(x_i) = (\pi_i)^{x_i} (1 - \pi_i)^{(1-x_i)}$$

- The Bernoulli distribution is a simple distribution for a single event - like flipping a coin

Bernoulli Distribution Illustration

- To illustrate the Bernoulli distribution (and statistical likelihoods in general), consider the following example
- To illustrate the Bernoulli distribution, consider the result of the Michigan / Illinois football game as a binary-response item, X.
  - Let's say $X = 1$ if Illinois wins, and $X = 0$ if Michigan wins
  - My prediction is that Illinois has about an 87% chance of winning the game.
  - So, $\pi = 0.87$.
- Likewise, $P(X = 1) = 0.87$ and $P(X = 0) = 0.13$.

Bernoulli Distribution Illustration

- The likelihood function for X looks similar:
- If $X = 1$, the likelihood is:

$$f(x_i = 1) = (0.87)^1 (1 - 0.87)^{(1-1)} = 0.87$$

- If $X = 0$, the likelihood is:

$$f(x_i = 0) = (0.87)^1 (1 - 0.87)^{(1-0)} = 0.13$$

- This example shows you how the likelihood function of a statistical distribution gives you the likelihood of an event occurring
- In the case of discrete-outcome variables, the likelihood of an event is synonymous with the probability of the event occurring

Independent Bernoulli Variables

- To consider what independence of Bernoulli variables means, let's consider the another game this season: Michigan v. Ohio State
- Let's say we predict Michigan has a 97% chance of winning (or $\pi_2 = 0.97$).
- By assumption of independence of games, the probability of both Illinois and Michigan wining their games would be the product of the probability of winning each game separately:

$$P(X_1 = 1, X_2 = 1) = \pi_1 \pi_2 = 0.87 \times 0.97 = 0.84$$

- More generally, we can express the likelihood of any set of occurrences by:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J) = \prod_{j=1}^{J} \pi_j^{x_j} \left(1 - \pi_j\right)^{(1-x_j)}$$

Finite Mixture Models

- LCA models are special cases of more general models called Finite Mixture Models
- A finite mixture model expresses the distribution of a set of outcome variables, **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 LCA model likelihood
- Here, we say that the conditional distribution of **X** given g is a sequence of independent Bernoulli variables

Latent Class Analysis as a FMM

A latent class model for the response vector of J variables ($j = 1, \ldots, J$) with C classes ($c = 1, \ldots, C$):

$$f(\mathbf{x}_i) = \sum_{c=1}^{C} \eta_c \prod_{j=1}^{J} \pi_{jc}^{x_{ij}} \left(1 - \pi_{jc}\right)^{1 - x_{ij}}$$

- η_c is the probability that any individual is a member of class c (must sum to one)
- x_{ij} is the observed response of individual i to item j
- π_{jc} is the probability of a positive response to item j from an individual from class c

LCA Local Independence

- As shown in the LCA distributional form, LCA assumes all Bernoulli variables are independent given a class
  - This assumption is called Local Independence
  - It is also present in many other latent variable modeling techniques:
    - Item response theory
    - Factor analysis (with uncorrelated errors)
- What is implied is that any association between observed variables is accounted for only by the presence of the latent class
  - Essentially, this is saying that the latent class is the reason that variables are correlated

Estimation Process

- Successfully applying an LCA model to data involves the resolution to two key questions:

  1. How many classes are present?

  2. What does each class represent?

- The answer to the first question comes from fitting LCA models with differing numbers of classes, then choosing the model with the best fit (to be defined later)

- The answer to the second question comes from inspecting the LCA model parameters of the solution that was deemed to have fit best

LCA Estimation Software

- There are several programs that exist that can estimate LCA models
- The package to be used today will be Mplus (with the Mixture add-on)
  - The full version of Mplus is very useful for many statistical techniques.
- Other packages also exist:
  - Latent Gold
  - A user-developed procedure in SAS (proc lca)

LCA Example #1

LCA Example #1

- To illustrate the process of LCA, we will use the example presented in Bartholomew and Knott (p. 142)
- The data are from a four-item test analyzed with an LCA by Macready and Dayton (1977)
- The test data used by Macready and Dayton were items from a math test

Table 6.3: Macready and Dayton's (1977) data with posterior probabilities of belonging to the mastery state

Response pattern	Frequency	Expected frequency	$P\{\text{Master}\|\mathbf{x}\}$
1111	15	15.0	1.00
1110	7	6.2	1.00
1101	23	19.7	1.00
1100	7	8.9	0.91
1011	1	4.2	1.00
1010	3	1.9	0.90
1001	6	6.1	0.90
1000	13	12.9	0.18
0111	4	4.9	1.00
0110	2	2.1	0.97
0101	5	6.6	0.98
0100	6	5.6	0.47
0011	4	1.4	0.97
0010	1	1.3	0.44
0001	4	4.0	0.45
0000	41	41.0	0.02
	142	142	

- Ultimately, Macready and Dayton wanted to see if examinees could be placed into two groups:
  - Those who had mastered the material
  - Those who had not mastered the material

LCA Example #1

- Several considerations will keep us from assessing the number of classes in Macready and Dayton's data:
  - We only have four items
  - Macready and Dayton hypothesized two distinct classes: masters and non-masters
- For these reasons, we will only fit the two-class model and interpret the LCA model parameter estimates

Mplus Input

```
TITLE:      LCA of Macready and Dayton's data (1977).
            Two classes.
DATA:       FILE IS mddata.dat;
VARIABLE:   NAMES ARE u1-u4;
            CLASSES = c(2);
            CATEGORICAL = u1-u4;
ANALYSIS:   TYPE = MIXTURE;
            STARTS = 100 100;
OUTPUT:     TECH1 TECH10;
PLOT:       TYPE=PLOT3;
            SERIES IS u1(1) u2(2) u3(3) u4(4);
SAVEDATA:   FORMAT IS f10.5;
            FILE IS examinee_ests.dat;
            SAVE = CPROBABILITIES;
```

LCA Parameter Information Types

- Recall, we have three pieces of information we can gain from an LCA:
  - Sample information - proportion of people in each class (η_c)
  - Item information - probability of correct response for each item from examinees from each class (π_{jc})
  - Examinee information - posterior probability of class membership for each examinee in each class (α_{ic})

Estimates of η_c

From Mplus:

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THE ESTIMATED MODEL

    Latent
   Classes

       1          83.29149          0.58656
       2          58.70851          0.41344
```

η_c are proportions in far right column

Estimates of π_{jc}

From Mplus:

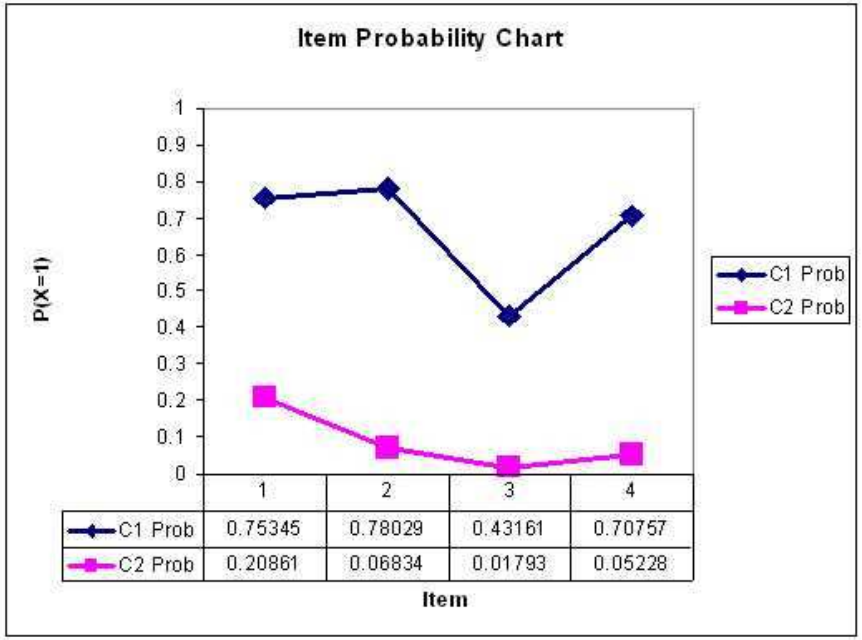
```
RESULTS IN PROBABILITY SCALE
Latent Class 1
 U1 Category 2          0.753    0.060
 U2 Category 2          0.780    0.069
 U3 Category 2          0.432    0.058
 U4 Category 2          0.708    0.063

Latent Class 2
 U1 Category 2          0.209    0.066
 U2 Category 2          0.068    0.056
 U3 Category 2          0.018    0.037
 U4 Category 2          0.052    0.057
```

π_{jc} are proportions in left column, followed by asymptotic standard errors

Interpreting Classes

- After the analysis is finished, we need to examine the item probabilities to gain information about the characteristics of the classes

- An easy way to do this is to look at a chart of the item response probabilities by class



- Here, we would say that Class 1 represents students who have mastered the material on the test

- We would say that Class 2 represents students who have not mastered the material on the test

Assessing Model Fit

Assessing Model Fit

- As with other statistical techniques, there is no one best way to assess the fit of an LCA model
- Techniques typically used can put into several general categories:
  - Model based hypothesis tests (absolute fit)
  - Information criteria
  - Measures based on distributional characteristics
  - Entropy

Model Based Measures

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

$$f(\mathbf{x}_i) = \sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} \left(1 - \pi_{ij}\right)^{1-x_i}$$

- Model based measures of fit revolve around the model function listed above
- With just the function above, we can compute the probability of *any* given response pattern
- Mplus gives this information using the TECH10 output option

Model Chi-squared Test

- The χ^2 test compares the sets of response patterns that were observed with the set of response patterns expected under the model
- To form the χ^2 test, one must first compute the probability of each response pattern using the latent class model equation displayed on the last slide
- The hypothesis tested is that the observed frequency is equal to the expected frequency
- If the test has a low p-value, the model is said to not fit
- To demonstrate the model χ^2 test, let's consider the results of the latent class model fit to the data from our running example (from Macready and Dayton, 1977)

Chi-squared Test Example

```
Class Probabilities:
Class   Probability
1       0.587
2       0.413

Item Parameters
class:          1
item    prob    SE(prob)
1       0.753   0.051
2       0.780   0.051
3       0.432   0.056
4       0.708   0.054

class:          2
item    prob    SE(prob)
1       0.209   0.060
2       0.068   0.048
3       0.018   0.029
4       0.052   0.044
```

Chi-squared Test Example

- To begin, compute the probability of observing the pattern $[1111]$...
- Then, to find the expected frequency, multiply that probability by the number of observations in the sample
- Repeat that process for all cells...
- The compute $\chi^2_p = \sum_r \frac{(O_r - E_r)^2}{E_r}$, where r represents each response pattern
- The degrees of freedom are equal to the number of response patterns minus model parameters minus one
- Then find the p-value, and decide if the model fits

Chi-squared from Mplus

RESPONSE PATTERN FREQUENCIES AND CHI-SQUARE CONTRIBUTIONS

Response	Frequency		Standard	Chi-square Contribution		
Pattern	Observed	Estimated	Residual	Pearson	Loglikelihood	Deleted
1	41.00	41.04	0.01	0.00	-0.08	
2	13.00	12.91	0.03	0.00	0.18	
3	6.00	5.62	0.16	0.03	0.79	
4	7.00	8.92	0.66	0.41	-3.39	
5	1.00	1.30	0.27	0.07	-0.53	
6	3.00	1.93	0.77	0.59	2.63	
7	2.00	2.08	0.05	0.00	-0.15	
8	7.00	6.19	0.33	0.10	1.71	
9	4.00	4.04	0.02	0.00	-0.07	
10	6.00	6.13	0.05	0.00	-0.26	
11	5.00	6.61	0.64	0.39	-2.79	
12	23.00	19.74	0.79	0.54	7.04	
13	4.00	1.42	2.18	4.70	8.29	
14	1.00	4.22	1.59	2.46	-2.88	
15	4.00	4.90	0.41	0.16	-1.62	
16	15.00	14.95	0.01	0.00	0.09	

Likelihood Ratio Chi-squared

- The likelihood ratio Chi-square is a variant of the Pearson Chi-squared test, but still uses the observed and expected frequencies for each cell

- The formula for this test is:

$$G = 2 \sum_r O_r \ln \left(\frac{O_r}{E_r} \right)$$

- The degrees of freedom are still the same as the Pearson Chi-squared test, however

Tests from Mplus

Chi-Square Test of Model Fit for the Binary and Ordered Categorical (Ordinal) Outcomes

Pearson Chi-Square	
Value	9.459
Degrees of Freedom	6
P-Value	0.1494
Likelihood Ratio Chi-Square	
Value	8.966
Degrees of Freedom	6
P-Value	0.1755

Chi-squared Problems

- The Chi-square test is reasonable for situations where the sample size is large, and the number of variables is small
  - If there are too many cells where the observed frequency is small (or zero), the test is not valid
- Note that the total number of response patterns in an LCA is 2^J, where J is the total number of variables
- For our example, we had four variables, so there were 16 possible response patterns
- If we had 20 variables, there would be a total of 1,048,576
  - Think about the number of observations you would have to have if you were to observe at least *one* person with each response pattern
  - Now think about if the items were highly associated (you would need even more people)

Model Comparison

- So, if model-based Chi-squared tests are valid only for a limited set of analyses, what else can be done?
- One thing is to look at comparative measures of model fit
- Such measures will allow the user to compare the fit of one solution (say two classes) to the fit of another (say three classes)
- Note that such measures are only valid as a means of relative model fit - what do these measures become if the model fits perfectly?

Log Likelihood

- Prior to discussing anything, let's look at the log-likelihood function, taken across all the observations in our data set

- The log likelihood serves as the basis for the AIC and BIC, and is what is maximized by the estimation algorithm

- The likelihood function is the model formulation across the joint distribution of the data (all observations):

$$L(\mathbf{x}_i) = \prod_{k=1}^{N} \left[\sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ki}} \left(1 - \pi_{ij}\right)^{1-x_{ki}} \right]$$

Log Likelihood

- The log likelihood function is the log of the model formulation across the joint distribution of the data (all observations):

$$LogL(\mathbf{x}_i) = \log\left(\prod_{k=1}^{N}\left[\sum_{j=1}^{K}\eta_j\prod_{i=1}^{p}\pi_{ij}^{x_{ki}}\left(1-\pi_{ij}\right)^{1-x_{ki}}\right]\right)$$

$$LogL(\mathbf{x}_i) = \sum_{k=1}^{N}\log\left(\sum_{j=1}^{K}\eta_j\prod_{i=1}^{p}\pi_{ij}^{x_{ki}}\left(1-\pi_{ij}\right)^{1-x_{ki}}\right)$$

- Here, the log function taken is typically base e - the natural log
- The log likelihood is a function of the observed responses for each person and the model parameters

Information Criteria

- The Akaike Information Criterion (AIC) is a measure of the goodness of fit of a model that considers the number of model parameters (q)

$$AIC = 2q - 2\log L$$

- Schwarz's Information Criterion (also called the Bayesian Information Criterion or the Schwarz-Bayesian Information Criterion) is a measure of the goodness of fit of a model that considers the number of parameters (q) and the number of observations (N):

$$BIC = q\log(N) - 2\log L$$

Fit from Mplus

TESTS OF MODEL FIT

Loglikelihood

H0 Value	-331.764

Information Criteria

Number of Free Parameters	9
Akaike (AIC)	681.527
Bayesian (BIC)	708.130
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	679.653
Entropy	0.754

Information Criteria

- When considering which model “fits” the data best, the model with the lowest AIC or BIC should be considered
- Although AIC and BIC are based on good statistical theory, neither is a gold standard for assessing which model should be chosen
- Furthermore, neither will tell you, overall, if your model estimates bear any decent resemblance to your data
- You could be choosing between two (equally) poor models - other measures are needed

Distributional Measures of Model Fit

- The model-based Chi-squared provided a measure of model fit, while narrow in the times it could be applied, that tried to map what the model said the data looked like to what the data actually looked like
- The same concept lies behind the ideas of distributional measures of model fit - use the parameters of the model to “predict” what the data should look like
- In this case, measures that are easy to attain are measures that look at:
  - Each variable marginally - the mean (or proportion)
  - The bivariate distribution of each pair of variables - contingency tables (for categorical variables), correlation matrices, or covariance matrices

Marginal Measures

- For each item, the model-predicted mean of the item (proportion of people responding with a value of one) is given by:

$$\hat{\bar{X}}_i = \hat{E}(X_j) = \sum_{x_j=0}^{M} \hat{P}(X_i = x_i)x_i = \sum_{j=1}^{J} \hat{\eta}_j \times \hat{\pi}_{ij}$$

- Across all items, you can then form an aggregate measure of model fit by comparing the observed mean of the item to that found under the model, such as the root mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\bar{X}}_i - \bar{X}_i)^2}{I}}$$

- Often, there is not much difference between observed and predicted mean (depending on the model, the fit will always be perfect)

Marginal Measures from Mplus

From Mplus (using TECH10):

```
UNIVARIATE MODEL FIT INFORMATION

                                  Estimated Probabilities
Variable                    H1              H0    Standard Residual
U1
  Category 1             0.472           0.472           0.000
  Category 2             0.528           0.528           0.000
U2
  Category 1             0.514           0.514           0.000
  Category 2             0.486           0.486           0.000
U3
  Category 1             0.739           0.739           0.000
  Category 2             0.261           0.261           0.000
U4
  Category 1             0.563           0.563           0.000
  Category 2             0.437           0.437           0.000
```

Bivariate Measures

- For each pair of items (say a and b, the model-predicted probability of both being one is given in the same way:

$$\hat{P}(X_a = 1, X_b = 1) = \sum_{j=1}^{J} \hat{\eta}_j \times \hat{\pi}_{aj} \times \hat{\pi}_{bj}$$

- Given the marginal means, you can now form a 2 x 2 table of the probability of finding a given pair of responses to variable a and b:

	a		
b	0	1	
0			$1 - \hat{P}(X_b = 1)$
1		$\hat{P}(X_a = 1, X_b = 1)$	$\hat{P}(X_b = 1)$
	$1 - \hat{P}(X_a = 1)$	$\hat{P}(X_a = 1)$	1

Bivariate Measures

- Given the model-predicted contingency table (on the last slide) for every pair of items, you can then form a measure of association for the items
- There are multiple ways to summarize association in a contingency table
- Depending on your preference, you could use:
  - Pearson correlation
  - Tetrachoric correlation
  - Cohen's kappa.
- After that, you could then summarize the discrepancy between what your model predicts and what you have observed in the data
  - Such as the RMSE, MAD, or BIAS.

Bivariate Measures from Mplus

From Mplus (using TECH10):

BIVARIATE MODEL FIT INFORMATION

		Estimated Probabilities		
Variable U1	Variable U2	H1	H0	Standard Residual
Category 1	Category 1	0.352	0.337	0.391
Category 1	Category 2	0.120	0.135	-0.540
Category 2	Category 1	0.162	0.177	-0.483
Category 2	Category 2	0.366	0.351	0.387

Entropy

- The entropy of a model is defined to be a measure of classification uncertainty
- To define the entropy of a model, we must first look at the posterior probability of class membership, let's call this $\hat{\alpha}_{ic}$ (notation borrowed from Dias and Vermunt, date unknown - online document)
- Here, $\hat{\alpha}_{ic}$ is the estimated probability that observation i is a member of class c
- The entropy of a model is defined as:

$$EN(\boldsymbol{\alpha}) = -\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}\log\alpha_{ij}$$

Relative Entropy

- The entropy equation on the last slide is bounded from $[0, \infty)$, with higher values indicated a larger amount of uncertainty in classification
- Mplus reports the *relative* entropy of a model, which is a rescaled version of entropy:

$$E = 1 - \frac{EN(\boldsymbol{\alpha})}{N \log J}$$

- The relative entropy is defined on $[0, 1]$, with values near one indicating high certainty in classification and values near zero indicating low certainty

Fit from Mplus

TESTS OF MODEL FIT

Loglikelihood

H0 Value	-331.764

Information Criteria

Number of Free Parameters	9
Akaike (AIC)	681.527
Bayesian (BIC)	708.130
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	679.653
Entropy	0.754

Latent Class Analysis: Wrap Up

LCA Limitations

- LCA has limitations which can make its general application difficult:
  - Classes not known prior to analysis
  - Class characteristics not know until after analysis
- Both of these problems are related to LCA being an exploratory procedure for understanding data
- Diagnostic Classification Models can be thought of as one type of a confirmatory LCA
  - By placing constraints on the class item probabilities and specifying what our classes mean prior to analysis

LCA Summary

- Latent class analysis is a model-based technique for finding clusters in binary (categorical) data
- Each of the variables is assumed to:
  - Have a Bernoulli distribution
  - Be independent given class
- Additional reading: Lazarsfeld and Henry (1968). Latent structure analysis

Latent Profile Analysis

LPA Introduction

- Latent profile models are commonly attributed to Lazarsfeld and Henry (1968)
- As it was originally conceived, LPA is an analysis that uses:
  - A set of continuous (metrical) variables - values allowed to range anywhere on the real number line
- The number of classes (an integer ranging from two through...) must be specified prior to analysis

LPA Process

- For a specified number of classes, LPA attempts to:
  - For each class, estimate the mean and variance for each variable
  - Estimate the probability that each observation falls into each class
    - For each observation, the sum of these probabilities across classes equals one
  - Across all observations, estimate the probability that *any* observation falls into a class

LPA Distributional Assumptions

- Because LPA works with continuous variables, the distributional assumptions of LPA must use a continuous distribution

- Within each latent class, the variables are assumed to:
  - Be independent
  - (Marginally) be distributed normal (or Gaussian):
    - For a single variable, the normal distribution function is:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(\frac{-(x_i - \mu_x)^2}{\sigma_x^2}\right)$$

Joint Distribution

- Because, conditional on class, we have normally distributed variables in LPA, we could also phrase the likelihood as coming from a multivariate normal distribution (MVN):

- The next set of slides describes the MVN

- What you must keep in mind is that our variables are set to be independent, conditional on class, so the within class covariance matrix will be diagonal

Multivariate Normal Distribution

- The generalization of the well-known normal distribution to multiple variables is called the multivariate normal distribution (MVN)
- Many multivariate techniques rely on this distribution in some manner
- Although real data may never come from a true MVN, the MVN provides a robust approximation, and has many nice mathematical properties
- Furthermore, because of the central limit theorem, many multivariate statistics converge to the MVN distribution as the sample size increases

MVN

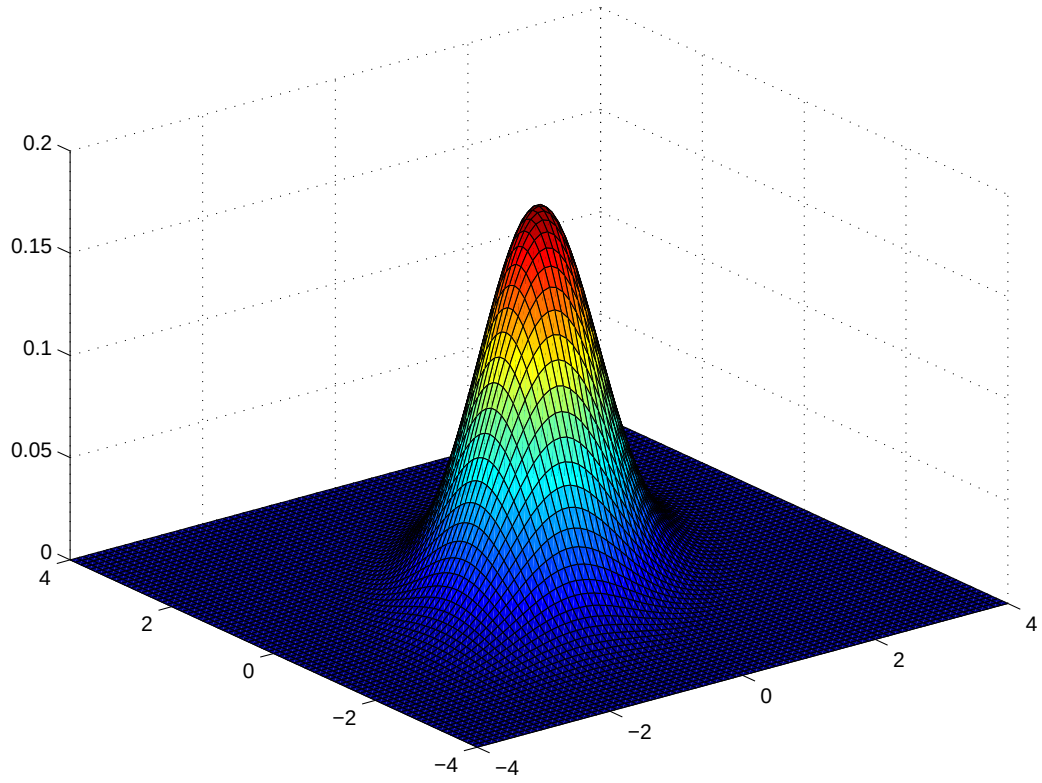
- The multivariate normal distribution function is:

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

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

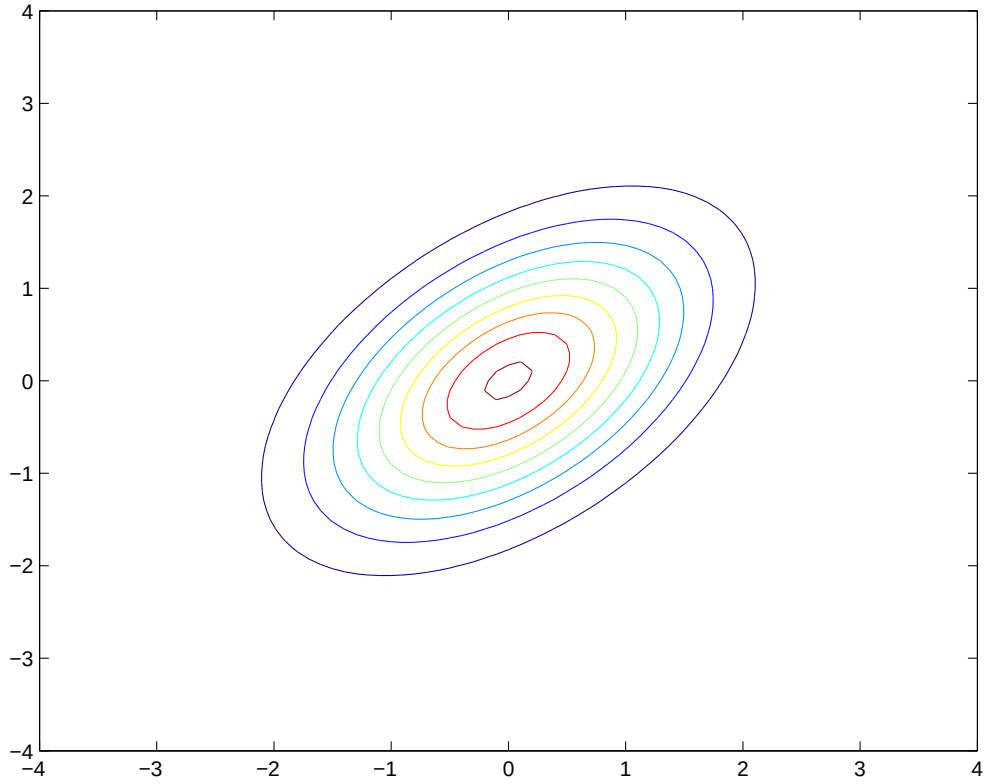
Bivariate Normal Plot #1

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



Bivariate Normal Plot #1a

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



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 Profile Analysis as a FMM

A latent profile model for the response vector of J variables ($j = 1, \ldots, J$) with C classes ($c = 1, \ldots, C$):

$$f(\mathbf{x}_i) = \sum_{c=1}^{C} \eta_c \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_{jc}^2}} \exp\left(\frac{-(x_i - \mu_{jc})^2}{\sigma_{jc}^2}\right)$$

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

LPA Example

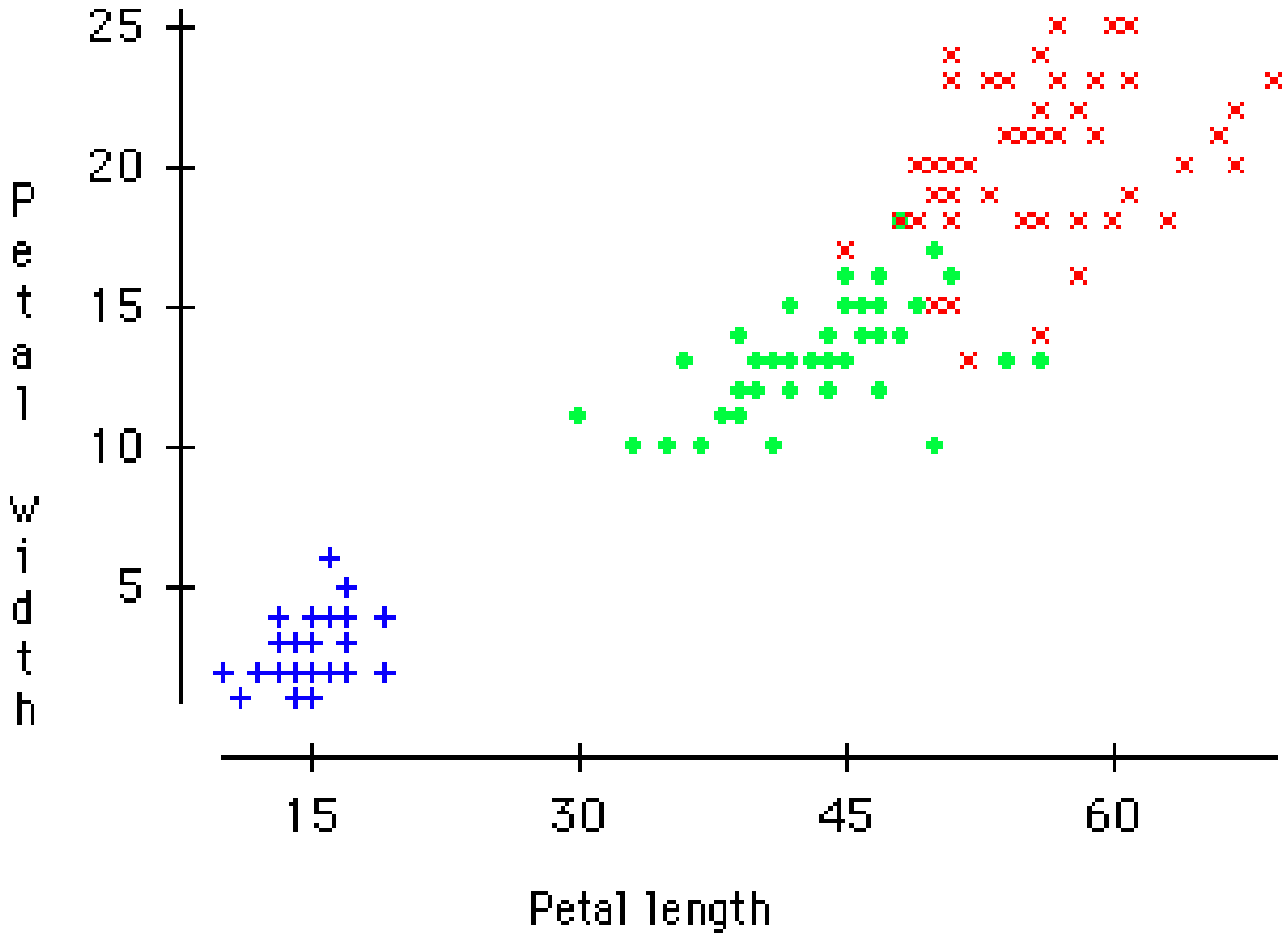
- To illustrate the process of LPA, consider an example using Fisher's Iris data
  - From CMU's DASL:

    The Iris dataset was introduced by R. A. Fisher as an example for discriminant analysis. The data report four characteristics (sepal width, sepal length, pedal width and pedal length) of three species of Iris flower.

- 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

LPA Example



```
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:
OUTPUT:
    TECH1 TECH5 TECH10;
PLOT:
    TYPE=PLOT3;
    SERIES IS x1(1) x2(2) x3(3) x4(4);

SAVEDATA:
    FILE IS myfile2c.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	13	-488.915	1003.830	1042.968	0.991
3-class	18	-361.426	758.851	813.042	0.957
4-class	23	-310.117	666.234	735.479	0.945

- 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 else to worry about!

Model Results

- The use of information criteria in this example highlights some of the problems with using such methods
- The data came from three distinct flowers
- The analysis suggested having more than three groups extracted
- Such problems are prevalent with many FMM techniques
- This highlights the need for use of validation techniques for any result obtained using these methods
- We will reject the 4-class solution and examine the 3-class solution because of our prior knowledge about the flowers

Model Results

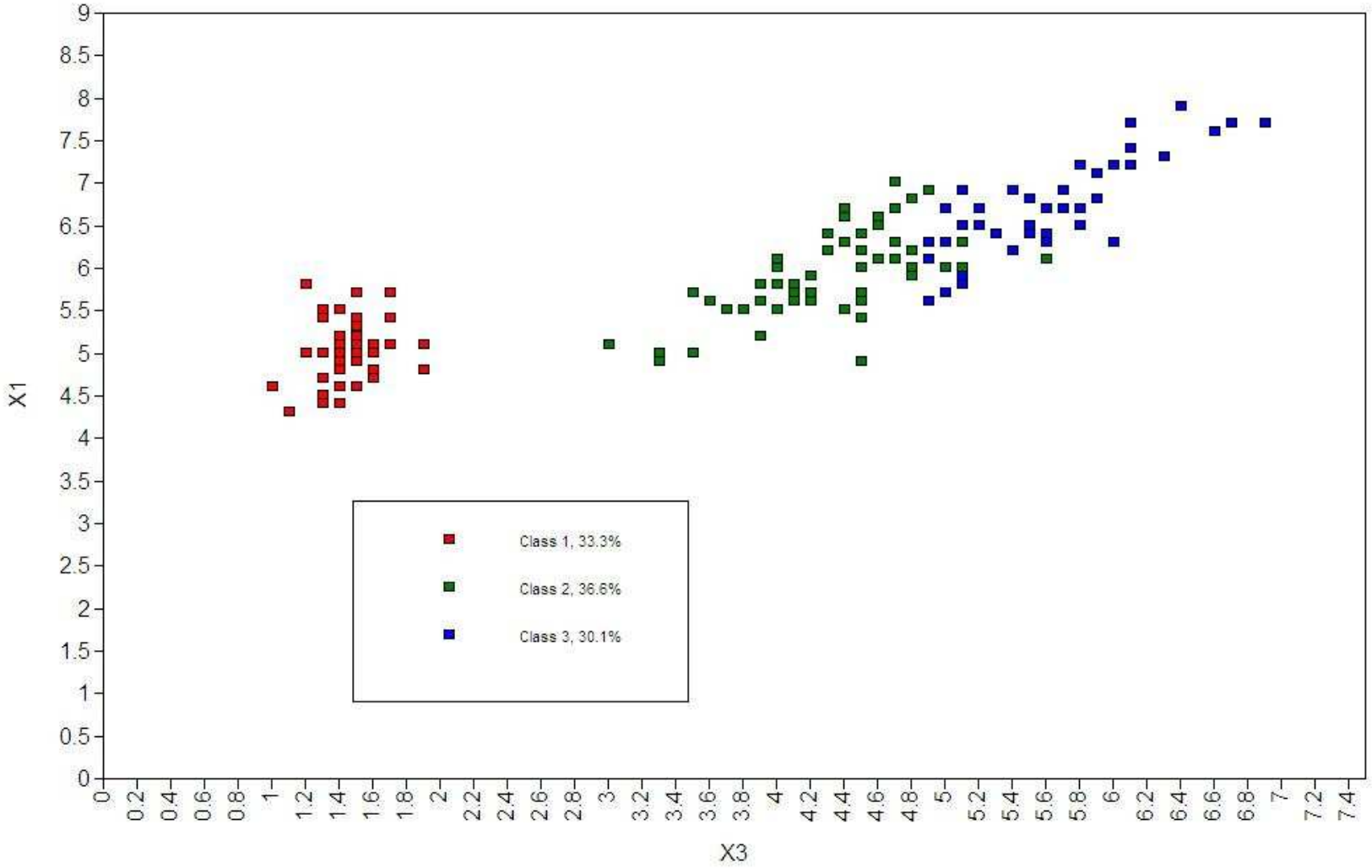
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED
THE ESTIMATED MODEL

Latent Classes		
1	50.00000	0.33333
2	54.88812	0.36592
3	45.11188	0.30075

Model Results

Latent Class 1			
Means			
X1	5.006	0.049	101.442
X2	3.428	0.053	64.595
X3	1.462	0.024	60.132
X4	0.246	0.015	16.673
Latent Class 2			
Means			
X1	5.920	0.079	75.391
X2	2.748	0.051	54.285
X3	4.327	0.122	35.533
X4	1.352	0.071	18.936
Latent Class 3			
Means			
X1	6.678	0.150	44.464
X2	3.023	0.054	55.527
X3	5.611	0.155	36.164
X4	2.070	0.062	33.501
Variances			
X1	0.236	0.030	7.957
X2	0.107	0.014	7.643
X3	0.187	0.027	6.862
X4	0.038	0.007	5.083

Model Results



LPA Summary

- Latent profile analysis is a model-based technique for finding clusters in continuous data
- Each of the variables is assumed to be:
  - Have a normal distribution
  - Be independent given class
- It is the continuous-distribution analog to LCA

Concluding Remarks

Concluding Remarks

- Many extensions of the models presented today can be found in recent empirical research articles
- Other methods include:
  - Growth mixture models
    - Methods for detecting groups that have differing growth trajectories
  - Diagnostic classification models
    - Methods for confirmatory analysis with classes
    - Used to diagnose psychological disorders and knowledge states of students
  - General Finite Mixture Models
    - Quite literally, any statistical distribution can be made into a mixture model

Concluding Remarks

- Today was a whirlwind tour of model-based clustering methods
- The methods described today are useful tools for the detecting of clusters within data
- Many of the ways to detect clusters can lead to problematic conclusions
  - Especially if information criteria are used to assess model fit
  - This speaks to the need for validation
- Proponents of the method say such techniques gives researchers the potential for very powerful analysis conclusions
- I say these techniques suffer from the same issues that other exploratory techniques suffer from - they are prone to problems and can lead to conclusions which are spurious at best and dangerous at worst