Latent Profile Analysis

Lecture 14

April 4, 2006

Clustering and Classification

Today's Lecture

- Latent Profile Analysis (LPA).
- LPA as a specific case of a Finite Mixture Model.
- How to do LPA.

LPA Introduction

- Latent profile models are commonly attributed to Lazarsfeld and Henry (1968).

- Like K-means and hierarchical clustering techniques, the final number of latent classes is not usually predetermined prior to analysis with latent class models.
  - The number of classes is determined through comparison of posterior fit statistics.
  - The characteristics of each class is also determined following the analysis.

Variable Types Used in LPA

- 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. Examples include:
- 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 statistical likelihood of each variable.
  - 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.

LPA Estimation

- Estimation in LPA is more complicated than in previous methods discussed in this course.
  - In agglomerative hierarchical clustering, a search process was used with new distance matrices being created for each step.
  - K-means used more of a brute-force approach - trying multiple starting points.
  - Both methods relied on distance metrics to find clustering solutions.
- LPA estimation uses distributional assumptions to find classes.
- The distributional assumptions provide the measure of "distance" in LPA.

Overview

Latent Profile Analysis
➤ LPA Input
➤ LPA Process
➤ LPA Estimation
➤ Assumptions

MVN

LPA as a FMM

LPA Example #1

Wrapping Up

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.

Univariate Normal Distribution

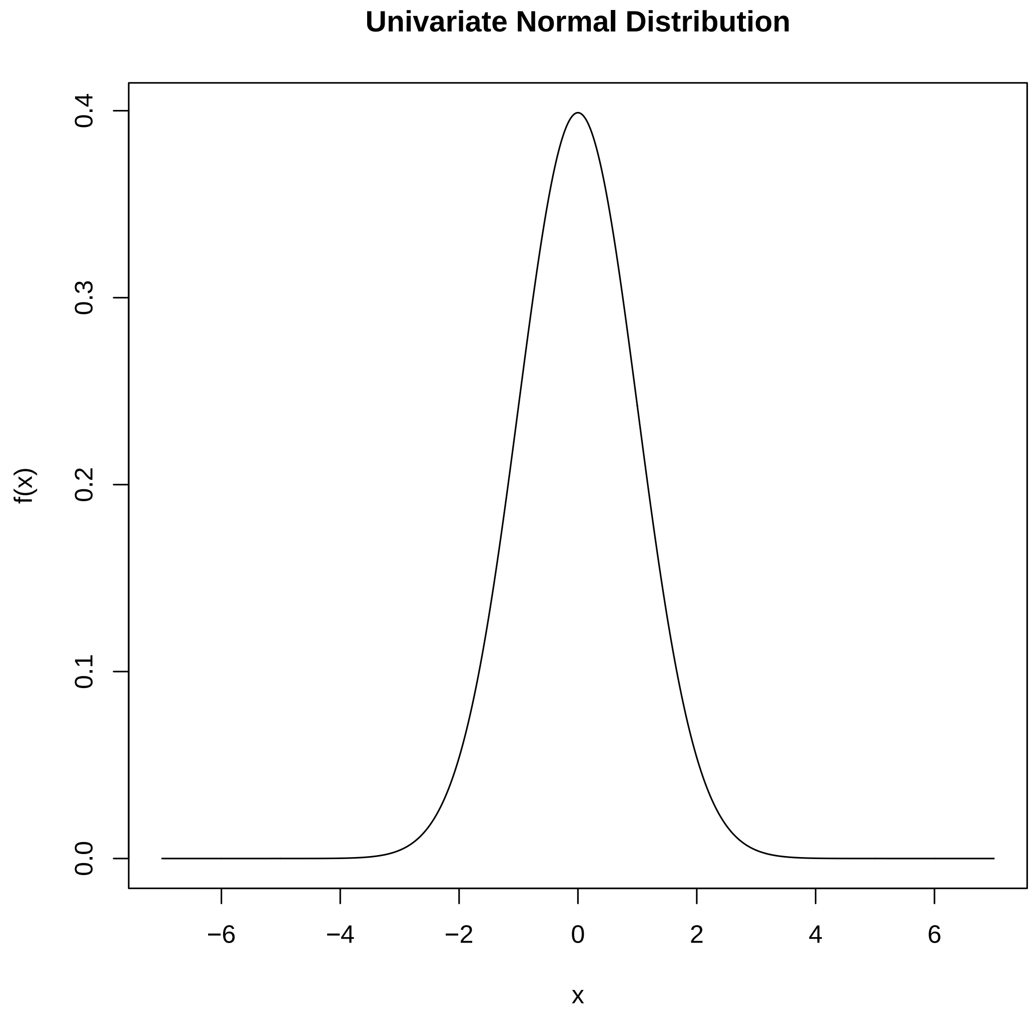
- The univariate normal distribution function is:

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

- The mean is μ_x.
- The variance is σ_x^2.
- The standard deviation is σ_x.
- Standard notation for normal distributions is $N(\mu_x, \sigma_x^2)$, which will be extended for the MVN distribution.

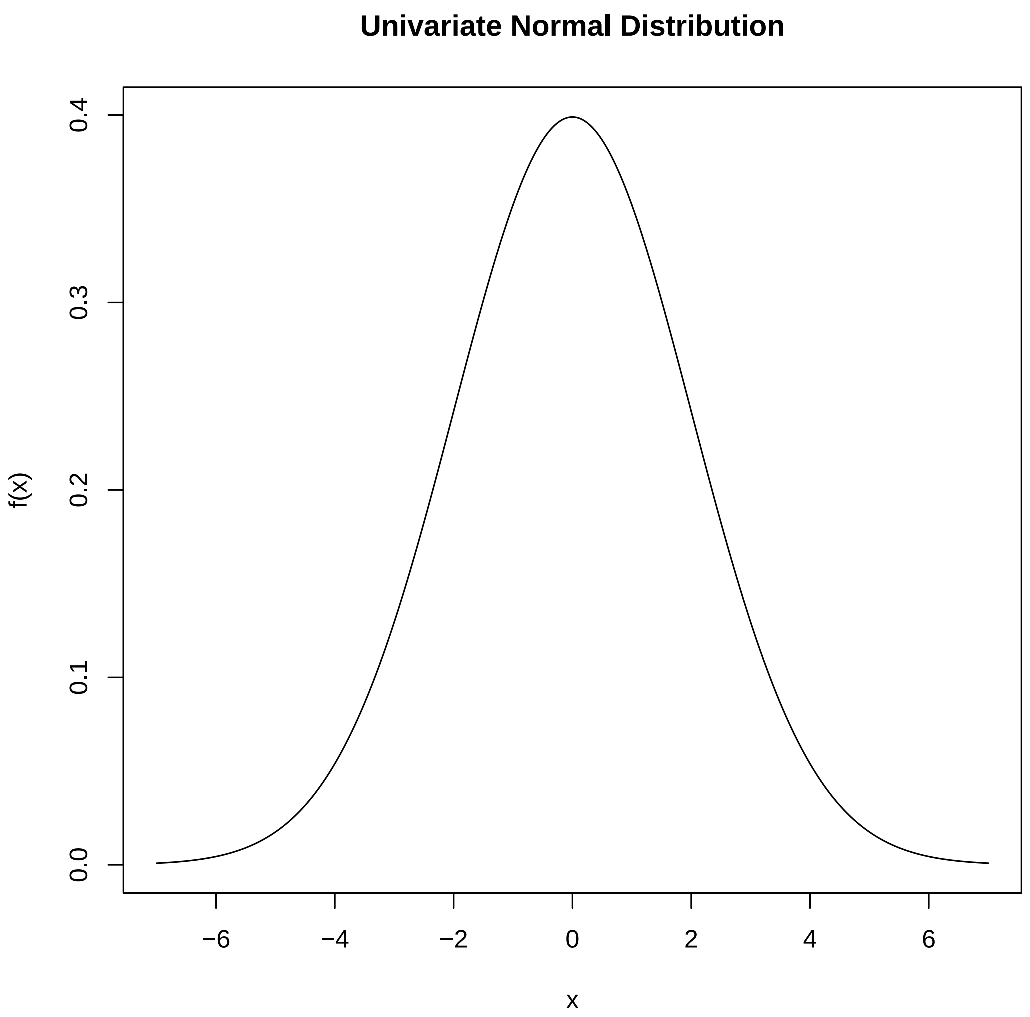
Univariate Normal Distribution

$$N(0,1)$$



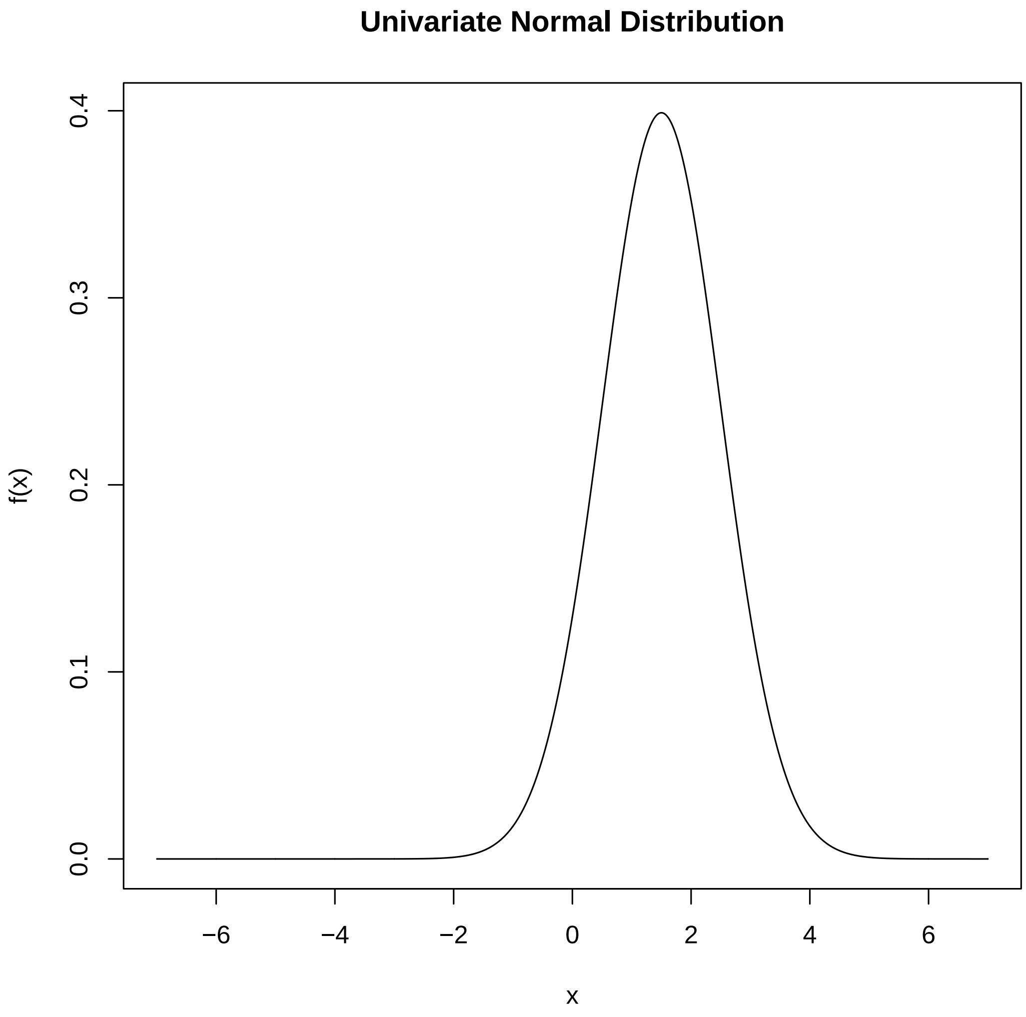
Univariate Normal Distribution

$$N(0,2)$$



Univariate Normal Distribution

$$N(3,1)$$



UVN - Notes

- Recall that the area under the curve for the univariate normal distribution is a function of the variance/standard deviation.

- In particular:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954$$

- Also note the term in the exponent:

$$\left(\frac{(x-\mu)}{\sigma} \right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

- This is the square of the distance from x to μ in standard deviation units, and will be generalized for the MVN.

MVN

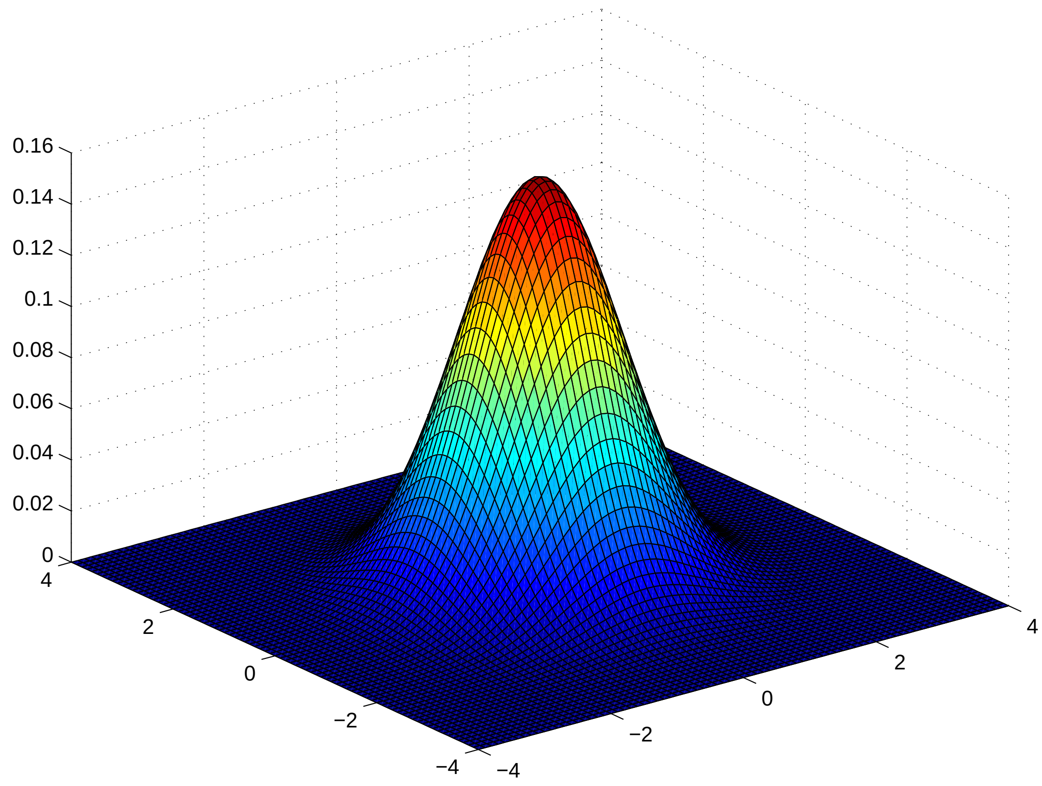
- The multivariate normal distribution function is:

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

- 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.

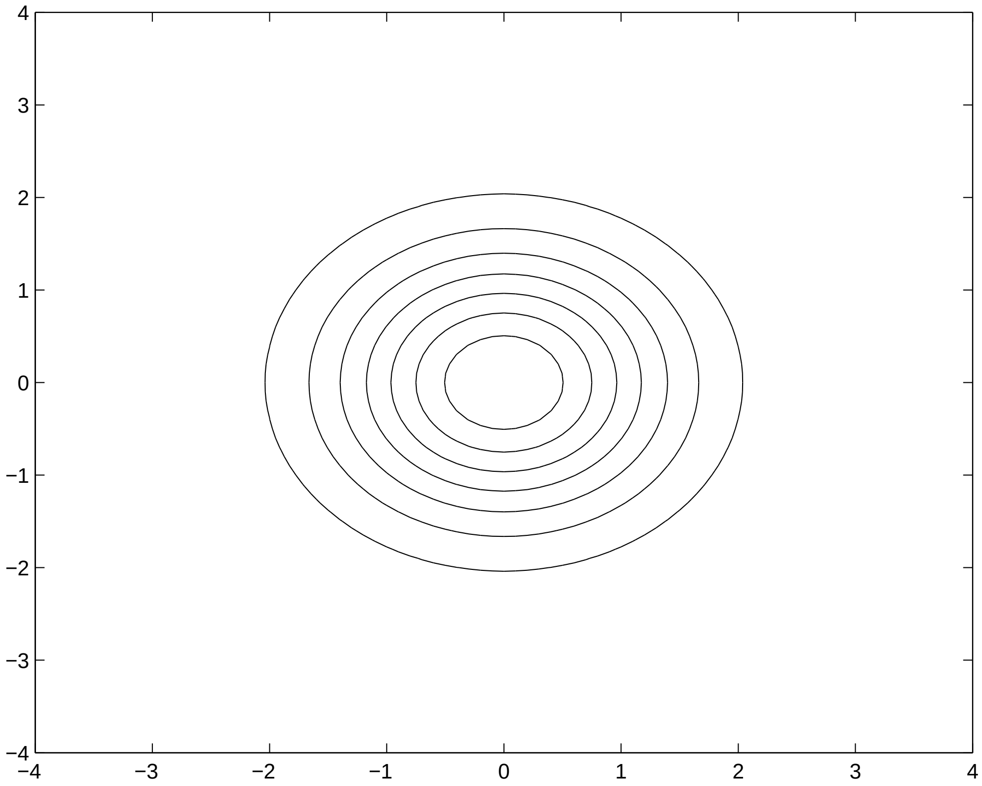
Bivariate Normal Plot #1

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



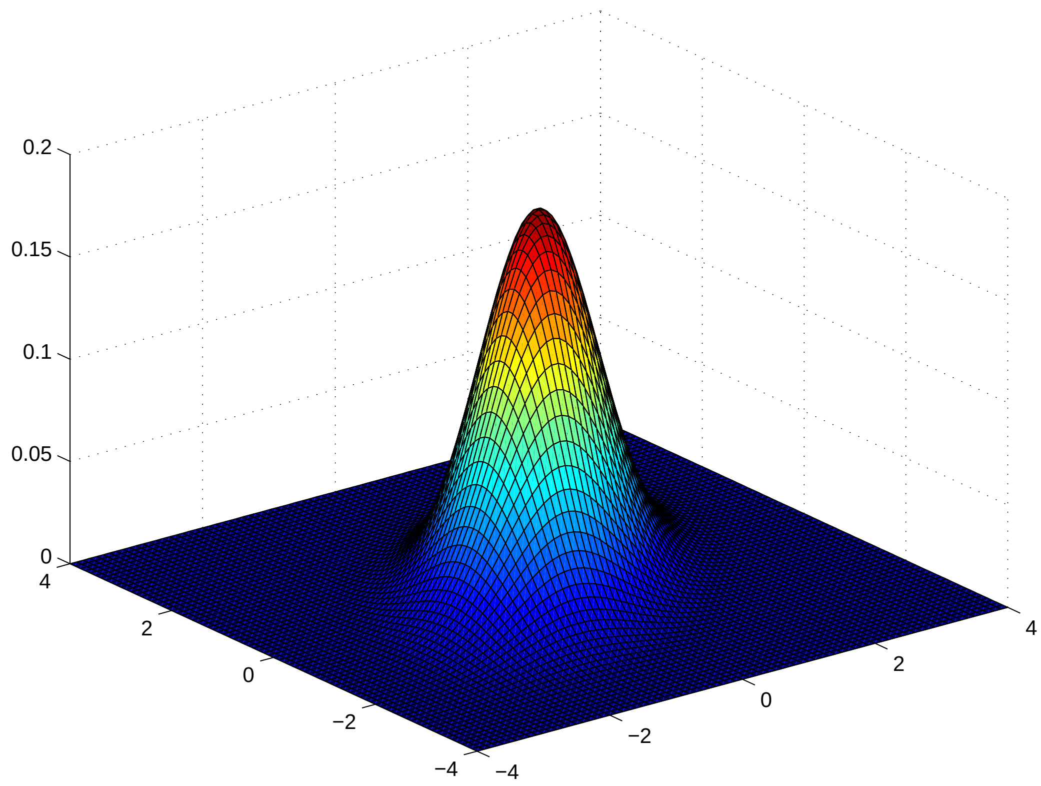
Bivariate Normal Plot #1a

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



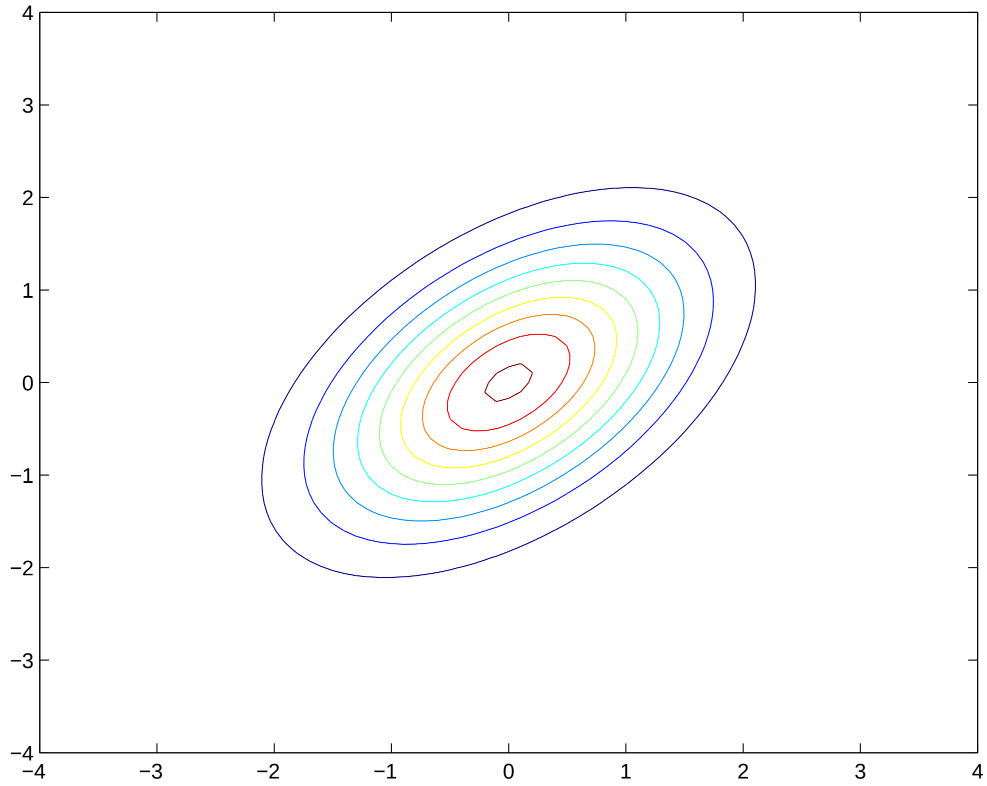
Bivariate Normal Plot #2

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



Bivariate Normal Plot #2

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



MVN Contours

- The lines of the contour plots denote places of equal probability mass for the MVN distribution.
- These contours can be constructed from the eigenvalues and eigenvectors of the covariance matrix.
  - The direction of the ellipse axes are in the direction of the eigenvalues.
  - The length of the ellipse axes are proportional to the constant times the eigenvector.
- Specifically:

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

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

MVN Contours, Continued

- Contours are useful because they provide confidence regions for data points from the MVN distribution.
- The multivariate analog of a confidence interval is given by an ellipsoid, where c is from the Chi-Squared distribution with p degrees of freedom.
- 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.

MVN Contour Example

- Imagine we had a bivariate normal distribution with:

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

- The covariance matrix has eigenvalues and eigenvectors:

$$\boldsymbol{\lambda} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix}, \mathbf{E} = \begin{bmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{bmatrix}$$

- We want to find a contour where 95% of the probability will fall, corresponding to $\chi^2_2(0.05) = 5.99$

MVN Contour Example

- This contour will be centered at $\boldsymbol{\mu}$.
- Axis 1:

$$\boldsymbol{\mu} \pm \sqrt{5.99 \times 1.5}\begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} 2.12 \\ 2.12 \end{bmatrix}, \begin{bmatrix} -2.12 \\ -2.12 \end{bmatrix}$$

- Axis 2:

$$\boldsymbol{\mu} \pm \sqrt{5.99 \times 0.5}\begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix} = \begin{bmatrix} -1.22 \\ 1.22 \end{bmatrix}, \begin{bmatrix} 1.22 \\ -1.22 \end{bmatrix}$$

MVN Properties

- The MVN distribution has some convenient properties.
- If **X** has a multivariate normal distribution, then:

1. Linear combinations of **X** are normally distributed.
2. All subsets of the components of **X** have a MVN distribution.
3. Zero covariance implies that the corresponding components are independently distributed.
4. The conditional distributions of the components are MVN.

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 ($j = 1, \ldots, K$):

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

- 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.

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.

```
title:
    Latent Profile Analysis of Fisher's Iri
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 myfile.dat;
     SAVE = CPROBABILITIES;
```

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 means by class.

Final Thought

- LPA is a wonderful technique to use to find classes with very specific types of data.
- We have only scratched the surface of LPA techniques.
- We will discuss estimation and other models in the weeks to come.

Next Time

- No class next two meetings (4-6 and 4-11).
- Our next class:
  - More LPA examples/facets.
  - A bit about estimation of such models.