An Introduction to the Multivariate Normal Distribution

EPSY 905: Fundamentals of Multivariate Modeling
Online Lecture #9
In This Lecture...

- Matrices in data
- The Multivariate Normal Distribution
DATA EXAMPLE AND R
A Guiding Example

- To demonstrate matrix algebra, we will make use of data.

- Imagine that I collected data SAT test scores for both the Math (SATM) and Verbal (SATV) sections of 1,000 students.

- The descriptive statistics of this data set are given below:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>SATV</th>
<th>SATM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>499.3</td>
<td>498.3</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>49.8</td>
<td>81.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation</th>
<th>SATV</th>
<th>SATM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATV</td>
<td>1.00</td>
<td>0.78</td>
</tr>
<tr>
<td>SATM</td>
<td>0.78</td>
<td>1.00</td>
</tr>
</tbody>
</table>
The Data...

In Excel:

In R:
MULTIVARIATE STATISTICS AND DISTRIBUTIONS
Multivariate Statistics

- Up to this point in this course, we have focused on the prediction (or modeling) of a single variable
  - Conditional distributions (aka, generalized linear models)

- Multivariate statistics is about exploring **joint distributions**
  - How variables relate to each other simultaneously

- Therefore, we must adapt our conditional distributions to have multiple variables, simultaneously (later, as multiple outcomes)

- We will now look at the joint distributions of two variables $f(x_1, x_2)$ or in matrix form: $f(X)$ (where $X$ is size N x 2; $f(X)$ gives a scalar/single number)
  - Beginning with two, then moving to anything more than two
  - We will begin by looking at **multivariate descriptive statistics**
    - Mean vectors and covariance matrices

- In this lecture, we only consider the **joint distribution** of sets of variables – but next time we will put this into a GLM-like setup
  - The **joint distribution** will then be conditional on other variables
Multiple Means: The Mean Vector

- We can use a vector to describe the set of means for our data

\[
\bar{x} = \frac{1}{N} X^T 1 = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_V \end{bmatrix}
\]

- Here 1 is a N x 1 vector of 1s
- The resulting mean vector is a V x 1 vector of means

- For our data: \( \bar{x} = \begin{bmatrix} 499.32 \\ 499.27 \end{bmatrix} = \begin{bmatrix} \bar{x}_{SATV} \\ \bar{x}_{SATM} \end{bmatrix} \)

- In R:

```r
# multivariate statistics ------------------------
N = (1/length(X[,1]))[1]
ONES = matrix(1,length(X[,1]),1)
XBAR = N*t(X)*%*%ONES
XBAR
```
Mean Vector: Graphically

- The mean vector is the center of the distribution of both variables
Covariance of a Pair of Variables

- The covariance is a measure of the relatedness
  - Expressed in the product of the units of the two variables:
    \[ s_{x_1x_2} = \frac{1}{N} \sum_{p=1}^{N} (x_{p1} - \bar{x}_1)(x_{p2} - \bar{x}_2) \]
  - The covariance between SATV and SATM was 3,132.22 (in SAT Verbal-Maths)
  - The denominator N is the ML version – unbiased is N-1

- Because the units of the covariance are difficult to understand, we more commonly describe association (correlation) between two variables with correlation
  - Covariance divided by the product of each variable’s standard deviation
Correlation of a Pair of Variables

- Correlation is covariance divided by the product of the standard deviation of each variable:

\[
r_{x_1x_2} = \frac{S_{x_1x_2}}{\sqrt{S_{x_1}^2} \sqrt{S_{x_2}^2}}
\]

- The correlation between SATM and SATV was 0.78

- Correlation is unitless – it only ranges between -1 and 1

  - If \( x_1 \) and \( x_2 \) both had variances of 1, the covariance between them would be a correlation
    - Covariance of standardized variables = correlation
Covariance and Correlation in Matrices

- The covariance matrix (for any number of variables $v$) is found by:

$$ S = \frac{1}{N} (X - 1\bar{x}^T) ^ T (X - 1\bar{x}^T) = \begin{bmatrix} S_{x_1}^2 & \cdots & S_{x_1x_v} \\ \vdots & \ddots & \vdots \\ S_{x_1x_v} & \cdots & S_{x_v}^2 \end{bmatrix} $$

- $S = \begin{bmatrix} 2,477.34 & 3,123.22 \\ 3,132.22 & 6,589.71 \end{bmatrix}$

- In R:

```r
> #calculating the mean vector:
> N = (1/length(X[,1]))[1]
> ONES = matrix(1,length(X[,1]),1)
> XBAR = N*t(X)%*%ONES
> XBAR
>       [,1]
> [1,] 499.32
> [2,] 498.27
> > #calculating the covariance matrix:
> > S = N*t(X-ONES)%*%(X-ONES)%*%t(XBAR))
> > S
>     [,1]       [,2]
> [1,] 2477.338 3132.224
> [2,] 3132.224 6589.707
```
From Covariance to Correlation

If we take the SDs (the square root of the diagonal of the covariance matrix) and put them into a diagonal matrix $D$, the correlation matrix is found by:

$$
R = D^{-1}SD^{-1} = \begin{bmatrix}
\frac{s_{x_1}^2}{\sqrt{s_{x_1}^2}} & \cdots & \frac{s_{x_1 x_p}}{\sqrt{s_{x_1} \cdot s_{x_p}}} \\
\frac{\sqrt{s_{x_1}^2}}{\sqrt{s_{x_1}^2}} & \vdots & \frac{\sqrt{s_{x_1 x_v}}}{\sqrt{s_{x_1} \cdot s_{x_v}}}
\end{bmatrix}
\begin{bmatrix}
\frac{s_{x_1}^2}{\sqrt{s_{x_1}^2}} & \cdots & \frac{s_{x_1 x_p}}{\sqrt{s_{x_1} \cdot s_{x_p}}} \\
\frac{\sqrt{s_{x_1}^2}}{\sqrt{s_{x_1}^2}} & \vdots & \frac{\sqrt{s_{x_1 x_v}}}{\sqrt{s_{x_1} \cdot s_{x_v}}}
\end{bmatrix}

= \begin{bmatrix}
1 & \cdots & r_{x_1 x_v} \\
\vdots & \ddots & \vdots \\
r_{x_1 x_v} & \cdots & 1
\end{bmatrix}
$$
Example Covariance Matrix

- For our data, the covariance matrix was:
  \[
  S = \begin{bmatrix}
  2477.34 & 3123.22 \\
  3132.22 & 6589.71 \\
  \end{bmatrix}
  \]

- The diagonal matrix \( D \) was:
  \[
  D = \begin{bmatrix}
  \sqrt{2477.34} & 0 \\
  0 & \sqrt{6589.71} \\
  \end{bmatrix} = \begin{bmatrix}
  49.77 & 0 \\
  0 & 81.18 \\
  \end{bmatrix}
  \]

- The correlation matrix \( R \) was:
  \[
  R = D^{-1}SD^{-1} = \begin{bmatrix}
  1 & 0 \\
  49.77 & 1 \\
  0 & 81.18 \\
  \end{bmatrix} \begin{bmatrix}
  2477.34 & 3123.22 \\
  3132.22 & 6589.71 \\
  \end{bmatrix} \begin{bmatrix}
  1 & 0 \\
  49.77 & 1 \\
  0 & 81.18 \\
  \end{bmatrix} = \begin{bmatrix}
  1.00 & 0.78 \\
  0.78 & 1.00 \\
  \end{bmatrix}
  \]
In R:

```r
> D = sqrt(diag(diag(S)))
> D
[,1]      [,2]
[1,] 49.77286 0.00000
[2,]  0.00000 81.17701
> Dinv = solve(D)
> Dinv
[,1]            [,2]
[1,] 0.02009127 0.00000000
[2,] 0.00000000 0.01231876
> R2 = Dinv%*%S%*%Dinv
> R2
[,1]            [,2]
[1,] 1.0000000 0.7752238
[2,] 0.7752238 1.0000000
> R
[,1]            [,2]
[1,] 1.0000000 0.7752238
[2,] 0.7752238 1.0000000
```
Generalized Variance

• The determinant of the covariance matrix is the **generalized variance**

  \[ \text{Generalized Sample Variance} = |S| \]

• It is a measure of spread across all variables
  - Reflecting how much overlap (covariance) in variables occurs in the sample
  - Amount of overlap reduces the generalized sample variance
  - Generalized variance from our SAT example: 6,514,104.5
  - Generalized variance if zero covariance/correlation: 16,324,929

```R
> gsv = det(S)
> gsv
[1] 6514104
```

• The generalized sample variance is:
  - Largest when variables are uncorrelated
  - Zero when variables form a linear dependency

• **In data:**
  - The generalized variance is seldom used descriptively, but shows up more frequently in maximum likelihood functions
Total Sample Variance

• The total sample variance is the sum of the variances of each variable in the sample
  
  ➢ The sum of the diagonal elements of the sample covariance matrix
  ➢ The trace of the sample covariance matrix

\[
Total \ Sample \ Variance = \sum_{i=1}^{V} s_{x_i}^2 = \text{tr } S
\]

• Total sample variance for our SAT example:

  > tsv = sum(diag(S))
  > tsv
  [1] 9067.045

• The total sample variance does not take into consideration the covariances among the variables
  
  ➢ Will not equal zero if linearly dependency exists

• **In data:**
  
  ➢ The total sample variance is commonly used as the denominator (target) when calculating variance accounted for measures
MULTIVARIATE DISTRIBUTIONS (VARIABLES ≥ 2)
Multivariate Normal Distribution

• The multivariate normal distribution is the generalization of the univariate normal distribution to multiple variables
  ➢ The bivariate normal distribution just shown is part of the MVN

• The MVN provides the relative likelihood of observing all $V$ variables for a subject $p$ simultaneously:
  
  \[ x_p = [x_{p1} \ x_{p2} \ \ldots \ \ x_{pV}] \]

• The multivariate normal density function is:

  \[
  f(x_p) = \frac{1}{\sqrt{V} \frac{1}{2} \sqrt{(2\pi)^2 |\Sigma|^2}} \exp \left[ - \frac{(x_p^T - \mu)^T \Sigma^{-1} (x_p^T - \mu)}{2} \right]
  \]
The Multivariate Normal Distribution

\[ f(x_p) = \frac{1}{V} \exp \left[ -\frac{(x_p - \mu)^T \Sigma^{-1} (x_p - \mu)}{2} \right] \]

- The mean vector is \( \mu = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \vdots \\ \mu_{x_V} \end{bmatrix} \)

- The covariance matrix is \( \Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_V} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 & \cdots & \sigma_{x_2 x_V} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{x_1 x_V} & \sigma_{x_2 x_V} & \cdots & \sigma_{x_V}^2 \end{bmatrix} \)

- The covariance matrix must be non-singular (invertible)
Comparing Univariate and Multivariate Normal Distributions

• The univariate normal distribution:

\[ f(x_p) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right] \]

• The univariate normal, rewritten with a little algebra:

\[ f(x_p) = \frac{1}{(2\pi)^{\frac{1}{2}}|\sigma^2|^{\frac{1}{2}}} \exp \left[ -\frac{(x - \mu)\sigma^{-\frac{1}{2}}(x - \mu)}{2} \right] \]

• The multivariate normal distribution

\[ f(x_p) = \frac{1}{V} \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \exp \left[ -\frac{(x_p^T - \mu)^T \Sigma^{-1} (x_p^T - \mu)}{2} \right] \]

➢ When \( V = 1 \) (one variable), the MVN is a univariate normal distribution.
The Exponent Term

- The term in the exponent \((without the) \left(-\frac{1}{2}\right)\) is called the **squared Mahalanobis Distance**

\[
d^2 (x_p) = (x_p^T - \mu)^T \Sigma^{-1} (x_p^T - \mu)
\]

- Sometimes called the statistical distance

- Describes how far an observation is from its mean vector, in standardized units

- Like a multivariate Z score (but, if data are MVN, is actually distributed as a \(\chi^2\) variable with DF = number of variables in X)

- Can be used to assess if data follow MVN
• Standard notation for the multivariate normal distribution of $v$ variables is $N_v(\mu, \Sigma)$
  - Our SAT example would use a bivariate normal: $N_2(\mu, \Sigma)$

• **In data:**
  - The multivariate normal distribution serves as the basis for most every statistical technique commonly used in the social and educational sciences
    - General linear models (ANOVA, regression, MANOVA)
    - General linear mixed models (HLM/multilevel models)
    - Factor and structural equation models (EFA, CFA, SEM, path models)
    - Multiple imputation for missing data

  - Simply put, the world of commonly used statistics revolves around the multivariate normal distribution
    - Understanding it is the key to understanding many statistical methods
Bivariate Normal Plot #1

\[ \mu = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma^2_{x_1} & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma^2_{x_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]

Density Surface (3D)

Density Surface (2D): Contour Plot
\[ \mu = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} \\ \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} \]

Density Surface (3D)

Density Surface (2D): Contour Plot
The multivariate normal distribution has some useful properties that show up in statistical methods.

- If $X$ is distributed multivariate normally:
  1. Linear combinations of $X$ are normally distributed.
  2. All subsets of $X$ are multivariate normally distributed.
  3. A zero covariance between a pair of variables of $X$ implies that the variables are independent.
  4. Conditional distributions of $X$ are multivariate normal.
To demonstrate how the MVN works, we will now investigate how the PDF provides the likelihood (height) for a given observation:

Here we will use the SAT data and assume the sample mean vector and covariance matrix are known to be the true:

\[
\mu = \begin{bmatrix} 499.32 \\ 498.27 \end{bmatrix};
S = \begin{bmatrix} 2,477.34 & 3,123.22 \\ 3,132.22 & 6,589.71 \end{bmatrix}
\]

We will compute the likelihood value for several observations (SEE EXAMPLE R SYNTAX FOR HOW THIS WORKS):

- \(x_{631} = [590 \ 730]; f(x) = 0.0000001393048\)
- \(x_{717} = [340 \ 300]; f(x) = 0.0000005901861\)
- \(\bar{x} = \overline{x} = [499.32 \ 498.27]; f(x) = 0.0000009924598\)

Note: this is the height for these observations, not the joint likelihood across all the data.

Next time we will use the R packaged named lavaan to find the parameters in \(\mu\) and \(\Sigma\) using maximum likelihood.
WRAPPING UP
Wrapping Up

- We are now ready to discuss multivariate models and the art/science of multivariate modeling

- Many of the concepts of univariate models carry over
  - Maximum likelihood
  - Model building via nested models

- All of the concepts involve multivariate distributions
Wrapping Up

• The last two classes set the stage to discuss multivariate statistical methods that use maximum likelihood

• Matrix algebra was necessary so as to concisely talk about our distributions (which will soon be models)

• The multivariate normal distribution will be necessary to understand as it is the most commonly used distribution for estimation of multivariate models

• Next week we will get back into data analysis – but for multivariate observations...using R’s lavaan package for path analysis
  ➢ Each term of the MVN will be mapped onto the lavaan() output