Notes on Multivariate Normal Distributions

Dr. Mohanty Stats 200

This document provides a few notes on multivariate normal (Gaussian) distributions to connect the univariate and bivariate cases with the more general, $P$ dimensional, case. In addition to detail on the vector notation, this document shows how to work with multivariate normals in R with actual data (i.e., using a dataset on popular movies).

Movie Data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><code>glimpse(movies)</code></td>
<td># glimpse() is in library(dplyr)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,961</td>
</tr>
<tr>
<td>Variables</td>
<td>11</td>
</tr>
<tr>
<td>$title$</td>
<td>Over the Hill to the Poorhouse, The Broad...</td>
</tr>
<tr>
<td>$genre$</td>
<td>Crime, Musical, Comedy, Comedy, Comedy, A...</td>
</tr>
<tr>
<td>$director$</td>
<td>Harry F. Millarde, Harry Beaumont, Lloyd ...</td>
</tr>
<tr>
<td>$year$</td>
<td>1920, 1929, 1933, 1935, 1936, 1937, 1939, ...</td>
</tr>
<tr>
<td>$duration$</td>
<td>110, 100, 89, 81, 87, 83, 102, 226, 88, 14...</td>
</tr>
<tr>
<td>$gross$</td>
<td>3.000000, 2.808000, 2.300000, 3.000000, 0....</td>
</tr>
<tr>
<td>$budget$</td>
<td>0.100000, 0.379000, 0.439000, 0.609000, 1....</td>
</tr>
<tr>
<td>$cast_facebook_likes$</td>
<td>4, 109, 995, 824, 352, 229, 2509, 1862, 11...</td>
</tr>
<tr>
<td>$votes$</td>
<td>5, 4546, 7921, 13269, 143086, 133348, 2918...</td>
</tr>
<tr>
<td>$reviews$</td>
<td>2, 107, 162, 164, 331, 349, 746, 863, 252,...</td>
</tr>
<tr>
<td>$rating$</td>
<td>4.8, 6.3, 7.7, 7.8, 8.6, 7.7, 8.1, 8.2, 7....</td>
</tr>
</tbody>
</table>

The PDF for the multivariate case is:

$$f(x) = \frac{1}{(2\pi)^{\frac{P}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)}$$

where:

- $P$ is the number of dimensions and also the length of the vectors $x$ and $\mu$. Note the number of dimensions is sometimes written as $N$, particularly when the focus is on theoretical properties, but once you have a data set, it is more conventional to say $N$ is the number of observations (above, $N = 2,961$) and $P$ is the dimensionality.

- $x = [x_1, x_2, ..., x_P]$ is a vector of observations on each dimension. In the movie dataset, if we were to focus in on only the numeric variables, $P = 8$ and $x_1 = [1920, 110, 3000000, 10000, ..., 4.8]$. 
• $\mu = [\mu_1, \mu_2, ..., \mu_P]$ is a vector of means on each dimension. With the movie data the population parameters are $\mu = [\mu_{\text{year}}, \mu_{\text{duration}}, \mu_{\text{gross}}, \mu_{\text{budget}}, ..., \mu_{\text{rating}}]$. The sample means can be found by using `colMeans()` on the appropriate columns of the data frame.

```r
colMeans(movies[, 4:11], digits = 1)
```

<table>
<thead>
<tr>
<th></th>
<th>year</th>
<th>duration</th>
<th>gross</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002.9</td>
<td>109.6</td>
<td>58.1</td>
<td></td>
</tr>
<tr>
<td>budget</td>
<td>cast_facebook_likes</td>
<td>votes</td>
<td></td>
</tr>
<tr>
<td>40.6</td>
<td>12393.8</td>
<td>109307.5</td>
<td></td>
</tr>
<tr>
<td>reviews</td>
<td>rating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>503.3</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• $\Sigma$ is the variance-covariance matrix (sometimes called the ‘variance matrix’ or the ‘covariance matrix’). $\Sigma$ contains variances on the diagonal and covariances off the diagonal.

$$
\Sigma = \begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1p} \\
\sigma_{21} & \sigma_2^2 & \sigma_{23} & \ldots & \sigma_{2p} \\
\sigma_{31} & \sigma_{32} & \sigma_3^2 & \ldots & \sigma_{3p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \ldots & \sigma_p^2
\end{bmatrix}
$$

$\Sigma$ is symmetric since, without loss of generality, $\sigma_{12} = \sigma_{21}$ (covariance is a measure of the strength and direction of the relationship between two variables, not whether one causes the other). Let’s look at the covariance matrix for just the first three numeric variables.

```r
cov(movies[, 4:6], digits = 1)
```

<table>
<thead>
<tr>
<th></th>
<th>year</th>
<th>duration</th>
<th>gross</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>98.9</td>
<td>-23.9</td>
<td>38.6</td>
</tr>
<tr>
<td>duration</td>
<td>-23.9</td>
<td>491.6</td>
<td>434.4</td>
</tr>
<tr>
<td>gross</td>
<td>38.6</td>
<td>434.4</td>
<td>5251.8</td>
</tr>
</tbody>
</table>

Just to pick a few examples...

```r
cov(movies$year, movies$gross)
[1] 38.59776

var(movies$gross)
[1] 5251.846
```

$\Sigma$ does not enter the equation directly. Instead, we have $|\Sigma|$, its determinant, and $\Sigma^{-1}$, its inverse. In the 2x2 case,

$$
|\Sigma| = \left| \begin{bmatrix}
\sigma_1^2 & \sigma_{12} \\
\sigma_{21} & \sigma_2^2
\end{bmatrix} \right| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2
$$

This quantity is necessarily positive (which is good since we must take its square root) and it achieves it maximum when their is zero covariance. For this reason, can be interpreted as a measure of how dependent the different dimensions are. Notice if the variance is one,
$0 < |\Sigma| \leq 1$ For example, the movie gross correlates strongly with budget (i.e., big budget films bring in more money) but not the year the movie was made. We see below that determinant approaches 1 for gross and year but is considerably lower for gross and budget.

```r
cor(movies$gross, movies$budget)
[1] 0.6408341
cor(movies$gross, movies$year)
[1] 0.05354855
```

Strictly speaking, division is not defined for matrix algebra but the inverse serves a similar purpose. By definition

$$\Sigma \Sigma^{-1} = I$$

where $I$ is the identity matrix:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

The inverse can be obtained in R with the command `solve`.

```r
# matrix multiplication is done with %*%
sigma.hat <- cov(movies[, 4:7])
sigma.hat %*% solve(sigma.hat) # obtains identity matrix up to numeric rounding
```

```r
year duration gross budget
year 1.0000000e+00 -2.012279e-16 -4.163336e-17 -2.775558e-17
duration -5.551115e-17 1.000000e+00 6.938894e-17 -5.551115e-17
gross 0.000000e+00 1.110223e-16 1.000000e+00 4.440892e-16
budget 0.000000e+00 0.000000e+00 0.000000e+00 1.000000e+00
```
The identity matrix is its own inverse ($II^{-1} = I$).

```r
diag(1, nrow = 4)  # construct diagonal matrix with 4 rows
[1,]   1   0   0   0
[2,]   0   1   0   0
[3,]   0   0   1   0
[4,]   0   0   0   1
```

That means that if all dimensions are independent, standard normals, $(x - \mu)'\Sigma^{-1}(x - \mu)$ simplifies to $(x - \mu)'(x - \mu)$. Of course, that’s a limiting case (at that point, the distribution is hardly worth calling multivariate).

More typically, $(x - \mu)'(x - \mu)$ is interpreted as (squared) Euclidean distance. $(x - \mu)'(x - \mu)$ yields the sum of squared distances between each observed point and the mean on that dimension:

$$
\begin{bmatrix}
(x_{i,1} - \mu_1) & (x_{i,2} - \mu_2)
\end{bmatrix}
\begin{bmatrix}
(x_{i,1} - \mu_1) \\
(x_{i,2} - \mu_2)
\end{bmatrix}
= (x_{i,1} - \mu_1)^2 + (x_{i,2} - \mu_2)^2
$$

where $i$ indexes movies $i = 1, 2, ..., 2961 = N$.

$\Sigma^{-1}$ is interpreted as a set of weights that accounts for dependence between the dimensions. A good exercise would be to use the definition of a matrix inverse and the population definition of correlation to obtain the definition of the bivariate normal distribution found in Rice. For more detail on weighted Euclidean distance, see Mahalanobis distance.

**Discussion**

Understanding multivariate Gaussians is essential for statistics for both statistical theory and data science applications. Multivariate Gaussians come up frequently in the regression context (see Rice 14.3) and are indispensable to high dimensional statistics and techniques like principal components analysis.