Example of Data-Determined Statistical Distance: Biting-Fly Data

In this example the sample mean vector and sample variance-covariance matrix are used in place of the population mean vector and population variance-covariance matrix to calculate statistical distances. Table 6.15 on page 352 of the textbook contains data from a study that examines the differences between two morphologically-similar species of biting flies. Measurements were taken on $n = 70$ specimens, 35 from each of the two species. For the moment we will ignore the fact that there are two species and simply consider $p = 2$ variables measured on each of the 70 specimens using wing length (x_1) and wing width (x_2).

On the next page is the series of ***R*** code commands and output supporting the calculations given below. From the output we note that the sample mean vector is $\bar{\mathbf{x}} = \begin{pmatrix} 97.90 \\ 43.33 \end{pmatrix}$ and the sample variance-covariance matrix is $\mathbf{S} = \begin{pmatrix} 37.60 & 14.99 \\ 14.99 & 16.57 \end{pmatrix}$. The eigenvalues and eigenvectors of **S** are calculated using the ***R*** command `eigen(S)`. The eigenvalues of **S** are $\lambda_1 = 45.39$ and $\lambda_2 = 8.78$ with corresponding eigenvectors $\mathbf{e}_1 = \begin{pmatrix} 0.89 \\ 0.46 \end{pmatrix}$ and $\mathbf{e}_1 = \begin{pmatrix} -0.46 \\ 0.89 \end{pmatrix}$. Since $\mathbf{A} = \mathbf{S}^{-1} = \begin{pmatrix} 0.042 & -0.038 \\ -0.038 & 0.094 \end{pmatrix}$ (found using the ***R*** command `ginv(S)` – after first loading the MASS library), this means that the eigenvalues of **A** are $\lambda_1^{-1} = (45.39)^{-1} = 0.02$ and $\lambda_2^{-1} = (8.78)^{-1} = 0.11$, with the same eigenvectors as those of **S**. The axes of the ellipse centered at $\bar{\mathbf{x}}$ (that is the set of points all of which have statistical distance *c* from $\bar{\mathbf{x}}$) are

$$\bar{\mathbf{x}} \pm c\sqrt{\lambda_1}\mathbf{e}_1 = \begin{pmatrix} 97.90 \\ 43.33 \end{pmatrix} \pm c\sqrt{45.39}\begin{pmatrix} 0.89 \\ 0.46 \end{pmatrix} = \begin{pmatrix} 97.90 \pm 5.98c \\ 43.33 \pm 3.11c \end{pmatrix},$$

and

$$\bar{\mathbf{x}} \pm c\sqrt{\lambda_2}\mathbf{e}_2 = \begin{pmatrix} 97.90 \\ 43.33 \end{pmatrix} \pm c\sqrt{8.78}\begin{pmatrix} -0.46 \\ 0.88 \end{pmatrix} = \begin{pmatrix} 97.90 \pm 1.36c \\ 43.33 \pm 2.61c \end{pmatrix}.$$

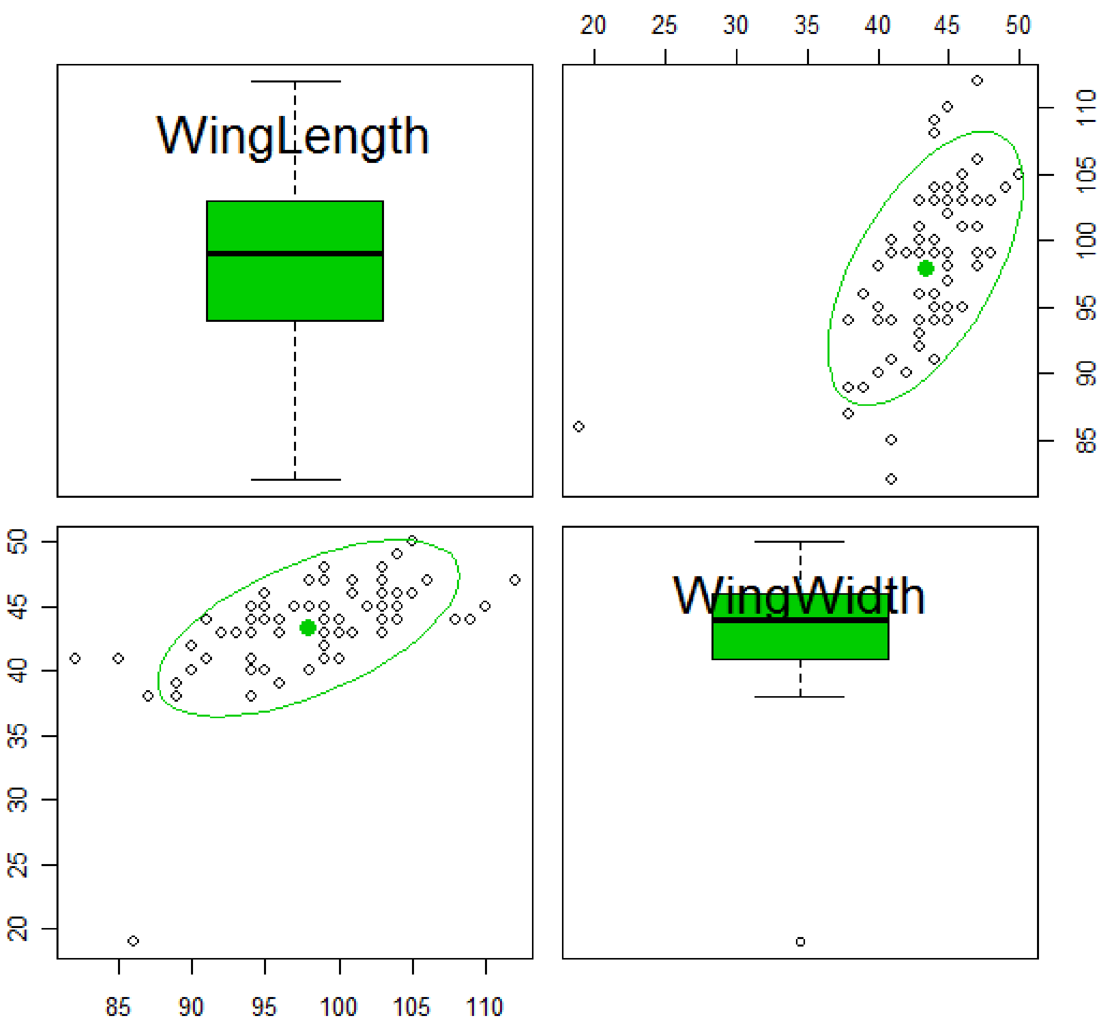
And finally, the formula for the squared statistical distance of a point $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ from $\bar{\mathbf{x}}$ is:

$$\begin{aligned} d^2(\mathbf{x}, \bar{\mathbf{x}}) &= (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \bar{\mathbf{x}}) = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \\ &= 0.042(x_1 - 97.90)^2 - 0.076(x_1 - 97.90)(x_2 - 43.33) + 0.094(x_2 - 43.33)^2 \end{aligned}$$

R code and Output for Biting Flies Example

```
> Flys <-
data.frame(WingLength=X[,1],WingWidth=X[,2],ThirdPalpLength=X[,3],ThirdPalpWidth=X[,4],FourthPalpLe
ngth=X[,5],LengthSeg12=X[,6],LengthSeg13=X[,7],Species=X[,8])
> mean(Flys)
     WingLength       WingWidth  ThirdPalpLength   ThirdPalpWidth
      97.900000       43.328571       37.342857        14.585714
FourthPalpLength      LengthSeg12      LengthSeg13          Species
      27.814286        9.614286        9.542857         0.500000
> #
> # construct the covariance matrix for the first two variables
> # (wing length and wing width)
> #
> S<-cov(Flys[,1:2])
> #
> # Find the e-values / e-vectors of S
> #
> eigen(S)
$values
[1] 45.394341  8.775846

$vectors
           [,1]       [,2]
[1,] -0.8871910  0.4614024
[2,] -0.4614024 -0.8871910

> #
> # need the MASS library to use the ginv command
> #
> library(MASS)
> #
> # calculate the inverse of S
> #
> A<-ginv(S)
> A
            [,1]        [,2]
[1,]  0.04159821 -0.03762762
[2,] -0.03762762  0.09438010
> #
> # Find the e-values / e-vectors of A (compare to those for S)
> #
> eigen(A)
$values
[1] 0.11394913 0.02202918

$vectors
           [,1]       [,2]
[1,] -0.4614024 -0.8871910
[2,]  0.8871910 -0.4614024

> spm(Flys[,1:2],diagonal=list(method="boxplot"),smooth=FALSE,regLine=FALSE,ellipse=
list(levels=c(0.75), robust=FALSE, fill=FALSE),main=c("Biting Flys: Wing Length & Wing Width"))
>
```

Biting Flys: Wing Length & Wing Width



(Courtesy of Dr. Roy St. Laurent)