MATH 4397/6397 – Project 1

Data provided by:
Tony Frankino, Department of Biology, University of Houston

Note: Please include all your work, printouts of R code, etc, with your answer.

1 Introduction

The populations of various species of Drosophila (fruit fly) show wing shape variations. The wing shape is quantified by measuring the location of landmarks defined by the intersection of veins with each other or with the wing margin (see Figure 1). There are a total of 15 landmarks, each defined by two coordinates. Hence, a wing shape is described by a 30-dimensional vector. We want to investigate if the wing shape can differentiate between populations of the following three species of Drosophila:

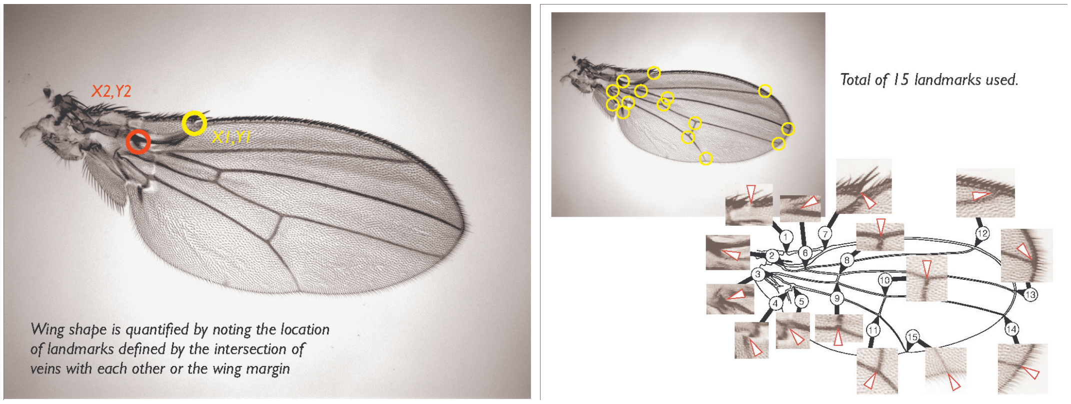
X2,Y2
X1,Y1
Wing shape is quantified by noting the location of landmarks defined by the intersection of veins with each other or the wing margin
Total of 15 landmarks used.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 1: Wing shape quantification

- D. Mauritiana
- D. Sechellia
- D. Simulans

The file `http://www.math.uh.edu/~bgb/biostats/Math4397/Project1/wing_xy.dat` contains the data for these three species.Read this file in R, into a variable called `wing_data` (use the function `read.table`). This creates a dataframe which contains the landmarks for 138 different flies, each belonging to one of the above three species. The entries from 1 to 42 are for D. Mauritiana, from 43 to 90 for D. Sechellia, and from 91 to 138 D. Simulans.

We begin by applying a data reduction technique. Our aim in the data reduction step is to obtain a scalar that quantifies the wing shape. This is done by taking the projection on a one-dimensional subspace. The choice of this subspace is described in the next section.

2 Data reduction

As noted above, the array `wing_data` contains the landmarks for 138 different flies. Hence we get a 138×30 matrix of observations (the matrix can be obatined from the dataframe via the R function `as.matrix`). Let us denote this matrix of observations by $\mathbf{Y}$. Perform data reduction in R as follows:

1. Find the mean of each column of $\mathbf{Y}$. The mean of column j is denoted by $\overline{Y}_j$. What does each of these means represent?

2. Next, for each column, subtract its mean from each of its entries. Call the new, "centered" matrix thus obtained $\mathbf{X}$. In other words, the i, j-th entry of the matrix $\mathbf{X}$ is given by:
$$\mathbf{X}_{i,j} = \mathbf{Y}_{i,j} - \overline{Y}_j .$$

3. Now calculate the matrix $\mathbf{C} = \frac{1}{137}\mathbf{X}^T\mathbf{X}$ (here $\mathbf{X}^T$ is the transpose of $\mathbf{X}$). What is the size of $\mathbf{C}$? What does the i, j-th entry of $\mathbf{C}$ represent statistically?

4. Diagonalize $\mathbf{C}$ using the function `eigen` in R.

5. Let α_{max} be the largest eigenvalue of $\mathbf{C}$ and $\mathbf{v}_{max}$ be the corresponding eigenvector. What do $\mathbf{v}_{max}$ and α_{max} represent? Explain in two or three sentences.

6. Let X_i denote the i row of $\mathbf{X}$ which corresponds to the i-th fly. Calculate z_i, the component of X_i in the direction of $\mathbf{v}_{max}$, given by the dot product
$$z_i = \langle X_i, \mathbf{v}_{max} \rangle .$$
This is the scalar that quantifies the wing shape for the i-th fly. Why do you think these z_i-values are good descriptors for the wing shape?

3 Hypothesis testing

Assume that the wing descriptors z_i are normally distributed with means $\mu_{mau}, \mu_{sim},$ and μ_{sec} for Mauritiana, Simulans, and Sechellia species respectively. Also assume that the standard deviations for the three populations, denoted by $\sigma_{mau}, \sigma_{sim},$ and σ_{sec}, are equal. Remember that z_1 through z_{42} correspond to Mauritiana, z_{43} through z_{90} correspond to Sechellia, and z_{91} through z_{138} correspond to Simulans.

1. Use an appropriate hypothesis test to determine if μ_{mau} is the same as μ_{sec}. Report the p-value.

2. Repeat this for μ_{mau} and μ_{sim}.

3. Repeat this for μ_{sec} and μ_{sim}.

Can you conclude that the expected value of the descriptor z_i differs between the three different species?