Ma 322: Biostatistics
Solutions to Homework Assignment 10

Prof. Wickerhauser

Due Monday, April 19th, 2021

Read Chapter 16, "Working with Multivariate Data," pages 288–318 of our text.

NOTE: Machine-readable data for the problems below is in `http://www.math.wustl.edu/~victor/classes/ma322/hw10data.txt`. Cut and paste from that document into a text file, or into an `R` variable by use of the `scan()` function.

See `hw10R.txt` for the R commands used to compute these results.

1. The following 40 ordered pairs $x = (x_1, x_2)$ are samples from a bivariate normal population:

x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
2.864810	−0.087901	2.388924	−0.112396	2.404386	1.536228	0.980159	−1.113963
0.579622	2.072845	−1.170284	0.211460	−1.153178	0.435754	0.739514	2.413948
1.384192	4.185621	2.157917	3.993882	−2.040037	−0.076255	1.189135	−0.800904
3.015638	2.956750	1.360922	1.483508	−0.156409	0.444964	1.827972	0.590482
0.852800	0.633167	0.258943	1.706435	−0.467125	−0.712590	2.863697	−1.876853
1.744421	2.453734	1.788729	−1.266549	2.108316	−2.300278	1.364329	1.972333
4.754022	1.574119	2.610398	−0.411356	1.432215	1.049123	1.041985	0.760463
0.469449	1.740265	0.090927	2.289402	1.998294	3.047970	2.124222	0.543565
1.226427	1.741965	2.167013	1.948388	−0.963964	−1.826650	1.367142	1.569296
1.122402	−1.337069	1.074869	2.284006	−0.124088	0.895195	1.873769	1.341474

(a) Estimate the population mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and the variance matrix Σ in the bivariate normal density $\frac{1}{2\pi\sqrt{\det \Sigma}}\, e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$ of this population.

(b) Compute the eigenvalues of the estimated matrix Σ.

Solution: See `http://www.math.wustl.edu/~victor/classes/ma322/hw10R.txt` for ways to compute the answers automatically using R.

(a) Find the sample mean $\bar{X}$ by averaging the 40 vectors. This gives

$$\mu \approx \bar{X} = \frac{1}{40}\sum_{i=1}^{32}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}_i = \begin{pmatrix} 1.228812 \\ 0.898839 \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix}.$$

In fact, this data comes from a population with $\mu = \begin{pmatrix} 1.3 \\ 0.2 \end{pmatrix}$.

Find the approximation to Σ by averaging $(X-\bar{X})(X-\bar{X})^T$, giving up one degree of freedom for the estimation $\bar{X}$ of μ. This gives

$$\Sigma \approx \frac{1}{39}\sum_{i=1}^{40}\begin{pmatrix} (x_1-\bar{x}_1)^2 & (x_1-\bar{x}_1)(x_2-\bar{x}_2) \\ (x_2-\bar{x}_2)(x_1-\bar{x}_1) & (x_2-\bar{x}_2)^2 \end{pmatrix} = \begin{pmatrix} 1.1750965 & 0.310032 \\ 0.310032 & 2.457364 \end{pmatrix}.$$

(b) The eigenvalues of a 2×2 matrix $E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix}$ are given by the quadratic formula:

$$\lambda = \frac{1}{2}\left(E_{11}+E_{22} \pm \sqrt{(E_{11}-E_{22})^2 + 4E_{12}E_{21}}\right),$$

where $+$ gives λ_1, $-$ gives λ_2. Note that for a symmetric positive definite matrix, both eigenvalues must be positive and the formula results in $\lambda_1 \geq \lambda_2 > 0$.

Applying these formulas to the estimated matrix Σ gives $\lambda_1 = 2.574132$, $\lambda_2 = 1.634197$. □

2. The following data gives the hypothetical concentrations of three amino acids in centipede hæmolymph (mg/100ml) labeled by gender:

Male			Female		
Alanine	Aspartic Acid	Tyrosine	Alanine	Aspartic Acid	Tyrosine
7.0	17.0	19.7	7.3	17.4	22.5
7.3	17.2	20.3	7.7	19.8	24.9
8.0	19.3	22.6	8.2	20.2	26.1
8.1	19.8	23.7	8.3	22.6	27.5
7.9	18.4	22.0	6.4	23.4	28.1
6.4	15.1	18.1	7.1	21.3	25.8
6.6	15.9	18.7	6.4	22.1	26.9
8.0	18.2	21.5	8.6	18.8	25.5

(a) Perform three analyses of variance on the three amino acid concentrations individually to test whether their concentrations are the same in males and females.

(b) Using multivariate analysis of variance, analyze the three amino acid concentrations together to determine whether their concentrations are the same in males and females.

Solution: See `http://www.math.wustl.edu/~victor/classes/ma322/hw10R.txt` for one way to compute the answers automatically using R.

(a) Individual responses to sex are

Amino Acid	F statistic	$df1/df2$	p-value	H_0?
Alanine	0.0519	1/14	0.8232	Do not reject
Aspartic Acid	11.317	1/14	0.004632	Reject
Tyrosine	30.257	1/14	7.814×10^{-5}	Reject

(b) The single nonzero eigenvalue is $\lambda_1 = 10.468$, giving a Wilks' lambda value $W = 1/(1 + \lambda_1) = 0.087$. This corresponds to a variance ratio $F \approx 41.872$ on 3 and 12 degrees of freedom, giving an approximate p-value of 1.24×10^{-6}, so we soundly **reject the null hypothesis** that hæmolymph amino acid concentration is the same in both sexes of centipede.

Additional tests (Hotelling-Lawley, Pillai, and Roy) give identical results in this case. □

3. The following data is from a hypothetical experiment involving 10 male and 10 female birds. Half the birds of each sex were given a hormone treatment and half were not. Two measurements were then made on each bird: plasma calcium concentration (in mg/100 ml) and rate of evaporative water loss (in mg/min). Perform a two-factor bivariate Model I MANOVA on the data.

Hormone Treatment				No Hormone Treatment			
Female		Male		Female		Male	
Plasma Calcium	Water Loss	Plasma Calcium	Water Loss	Plasma Calcium	Water Loss	Plasma Calcium	Water Loss
16.5	76	14.5	80	39.1	71	32.0	65
18.4	71	11.0	72	26.2	70	23.8	69
12.7	64	10.8	77	21.3	63	28.8	67
14.0	66	14.3	69	35.8	59	25.0	56
12.8	69	10.0	74	40.2	60	29.3	52

Solution: See the file `http://www.math.wustl.edu/~victor/classes/ma322/hw10R.txt` for the outputs of R commands to perform this analysis.

Factor A test: Wilks $W = 0.182$, Roy $R = 4.496$, Hotelling-Lawley $H = 4.496$, and Pillai $P = 0.818$ all give $F \approx 33.717$, while $F_{\alpha(1),d,N-d-1} = F_{0.05(1),2,17} = 3.59$, so we **reject the null hypothesis** that hormone treatment has no effect.

Factor B test: Wilks $W = 0.830$, Roy $R = 0.206$, Hotelling-Lawley $H = 0.206$, and Pillai $P = 0.170$ all give $F = 1.541$, while $F_{\alpha(1),d,N-d-1} = F_{0.05(1),2,17} = 3.59$, so we **do not reject the null hypothesis** that sex has no effect.

Factor $A \times B$ test: Wilks $W = 0.853$, Roy $R = 0.173$, Hotelling-Lawley $H = 0.173$, and Pillai $P = 0.147$ all give $F = 1.295$, while $F_{\alpha(1),2g-2,2N-2g-2} = F_{0.05(1),6,30} = 2.42$, so we **do not reject the null hypothesis** that hormone treatment and sex have no interaction. □

4. For this problem, use the amino acid concentration data in Problem 2.

(a) Plot all pairs of amino acid concentrations on a 3×3 grid of graphs using the `R` command `pairs()`. Identify the plotted points by sex using "x" for males and "o" for females.

(b) Plot the 3-d scatterplot amino acid concentrations.

(Hint: `install.packages("scatterplot3d")`.)

(c) Find the principal components of the amino acid data and scree plot their importance. (Hint: `screeplot()` and `princomp()` are included in the standard `R` installation.)

(d) A centipede has the following amino acid concentrations in its hæmolymph:

Amino Acid	Concentration (mg/100ml)
Alanine	7.5
Aspartic Acid	18.1
Tyrosine	22.1

Use linear discriminant analysis to judge whether it is likelier to be male or female.

(Hint: `install.packages("MASS")` for function `lda()`.)

(e) Use cross-validation on linear discriminant analysis for the given data to estimate the probabilities of correctly classifying male and female centipedes from the concentrations of the three amino acids in their hæmolymph.

Solution: See the file `http://www.math.wustl.edu/~victor/classes/ma322/hw10R.txt` for the R commands used to perform this analysis.

(a) See Figure HW10,Ex.4(a) below.

(b) See Figure HW10,Ex.4(b) below.

(c) See Figure HW10,Ex.4(c) below.

(d) The output in `hw10R.txt` shows that it's a Male with 99.7% probability.

(e) For my 100 choices of training sets, the table showed 100% probability of correctly classifying Males, and 98.75% probability of correctly classifying Females. □

5. For this problem, again use the amino acid concentration data in Problem 2.

   A centipede has the following amino acid concentrations in its hæmolymph:

Amino Acid	Concentration (mg/100ml)
Alanine	7.55
Aspartic Acid	18.1
Tyrosine	23.3

   Use Mahalanobis distance to judge whether it is likelier to be male or female.

   Solution: See the file `http://www.math.wustl.edu/~victor/classes/ma322/hw10R.txt` for the R commands to perform this analysis and their outputs.

   Get Mahalanobis distances 64.47468 from the male cluster mean and 2.380527 from the female cluster mean. Conclude that this centipede is (much) likelier to be female. □

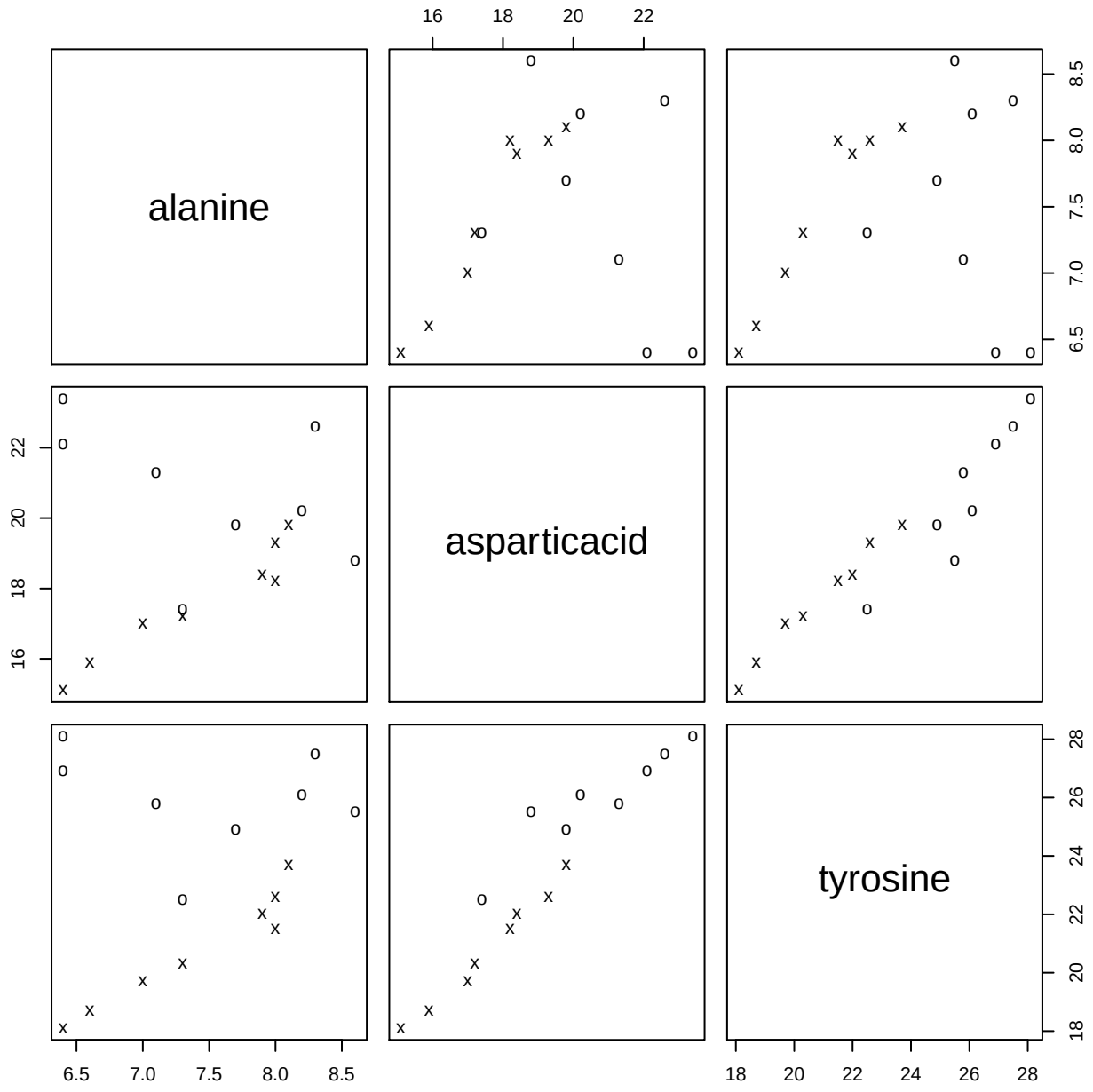

Figure 1: HW10,Ex.4(a): Pairs plot of a3 amino acid concentrations by gender (x=male, o=female).

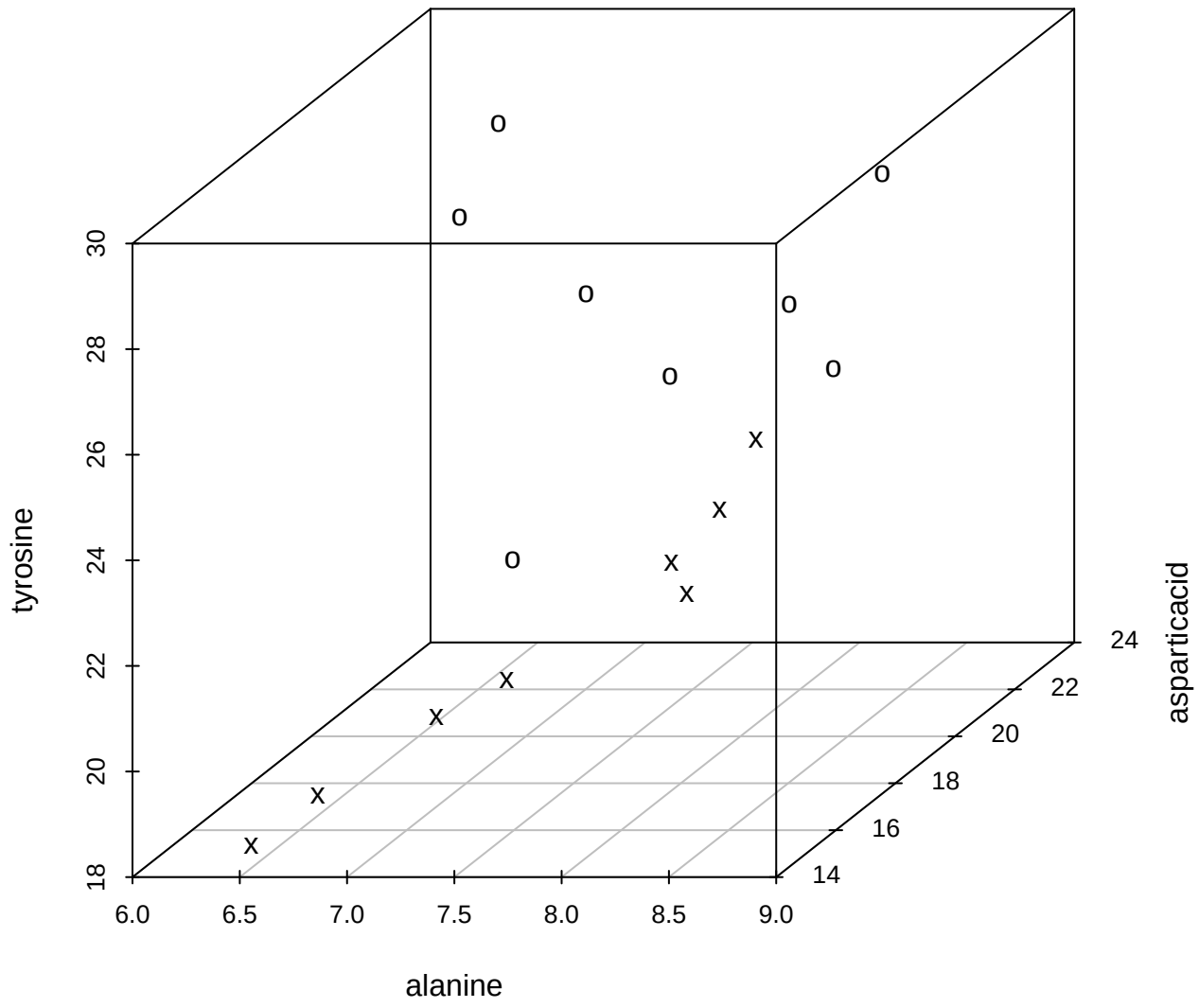

Figure 2: HW10,Ex.4(b): 3D scatterplot of amino acid concentrations by gender (x=male, o=female).

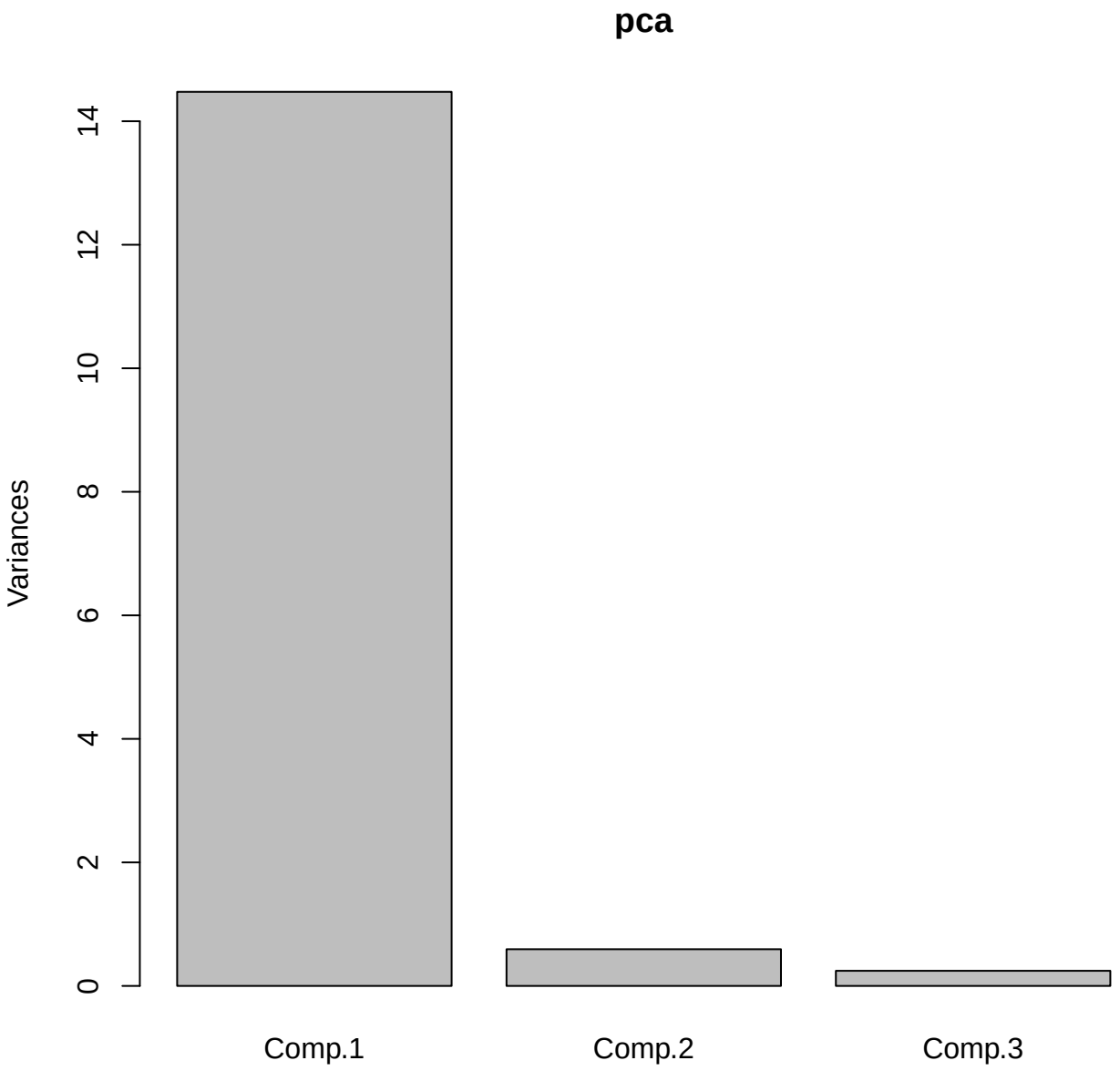

Figure 3: HW10,Ex.4(c): Screeplot of principal components of amino acid concentrations.