Chapter 2
Multivariate Distributions

This chapter describes the multivariate location and dispersion (MLD) model, random vectors, the population mean, the population covariance matrix, and the classical MLD estimators: the sample mean and the sample covariance matrix. Some important results on Mahalanobis distances and the volume of a hyperellipsoid are given. Often methods of multivariate analysis work best when the variables $x_1, ..., x_p$ are linearly related. Section 2.4 discusses power transformations to remove gross linearities from the variables.

2.1 Introduction

Definition 2.1. An important *multivariate location and dispersion model* is a joint distribution with joint probability density function (pdf)

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z})d\boldsymbol{z}$ for suitable sets A.

Notation: Usually a vector $\boldsymbol{x}$ will be column vector, and a row vector $\boldsymbol{x}^T$ will be the transpose of the vector $\boldsymbol{x}$. However,

$$\int_A f(\boldsymbol{z})d\boldsymbol{z} = \int_A f(z_1, ..., z_p)dz_1 \cdots dz_p.$$

The notation $f(z_1, ..., z_p)$ will be used to write out the components z_i of a joint pdf $f(\boldsymbol{z})$ although in the formula for the pdf, e.g., $f(\boldsymbol{z}) = c\exp(\boldsymbol{z}^T\boldsymbol{z})$, $\boldsymbol{z}$ is a column vector.

D. J. Olive, *Robust Multivariate Analysis*,
https://doi.org/10.1007/978-3-319-68253-2_2

Definition 2.2. A $p \times 1$ *random vector* $\boldsymbol{x} = (x_1, ..., x_p)^T = (X_1, ..., X_p)^T$ where $X_1, ..., X_p$ are p random variables. A *case* or *observation* consists of the p random variables measured for one person or thing. For multivariate location and dispersion, the ith case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are n cases, and context will be used to determine whether $\boldsymbol{x}$ is the random vector or the observed value of the random vector. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{x}_i$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Notation: *In the theoretical sections of this text,* $\boldsymbol{x}_i$ *will sometimes be a random vector and sometimes the observed data.* Some texts, for example Johnson and Wichern (1988, pp. 7, 53), use $\boldsymbol{X}$ to denote the $n \times p$ data matrix and an $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, *R* will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence x corresponds to $\boldsymbol{W}$, x[,1] is the first column of x, and $x[4,]$ is the 4th row of x.

2.2 The Sample Mean and Sample Covariance Matrix

The population location vector $\boldsymbol{\mu}$ need not be the population mean, but often the population mean is denoted by $\boldsymbol{\mu}$. For elliptically contoured distributions, such as the multivariate normal distribution, $\boldsymbol{\mu}$ is usually the point of symmetry for the population distribution. See Chapter 3.

Definition 2.3. If the second moments exist, the *population mean* of a random $p \times 1$ vector $\boldsymbol{x} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(X_1), ..., E(X_p))^T,$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T] =$$

$$E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})[E(\boldsymbol{x})]^T = (\sigma_{ij}) = (\sigma_{i,j}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}.$$

That is, the ij entry of $\text{Cov}(\boldsymbol{x})$ is $\text{Cov}(X_i, X_j) = \sigma_{ij} = E([X_i - E(X_i)][X_j - E(X_j)])$. The $p \times p$ population correlation matrix $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$. That is, the ij entry of $\text{Cor}(\boldsymbol{x})$ is $\text{Cor}(X_i, X_j) =$

$$\frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Let the $p \times p$ population standard deviation matrix

$$\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}}).$$

Then

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta}\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\Delta}, \tag{2.1}$$

and

$$\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}. \tag{2.2}$$

Let the population standardized random variables

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$$

for $i = 1, ..., p$. Then $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{z})$ is the covariance matrix of $\boldsymbol{z} = (Z_1, ..., Z_p)^T$.

Definition 2.4. Let random vectors $\boldsymbol{x}$ be $p \times 1$ and $\boldsymbol{y}$ be $q \times 1$. The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is the $p \times q$ matrix

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T] =$$

$$E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{y}^T] = E(\boldsymbol{x}\boldsymbol{y}^T) - E(\boldsymbol{x})[E(\boldsymbol{y})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}$$

assuming the expected values exist. Note that the $q \times p$ matrix $\text{Cov}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}^T$, and $\text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{x}, \boldsymbol{x})$.

A $p \times 1$ random vector $\boldsymbol{x}$ has an *elliptically contoured distribution,* if $\boldsymbol{x}$ has pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{2.3}$$

and we say $\boldsymbol{x}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. See Chapter 3. If second moments exist for this distribution, then

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\boldsymbol{x}) = \mathrm{c_x}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$$

for some constant $c_x > 0$ where the ij entry is $\mathrm{Cov}(X_i, X_j) = \sigma_{i,j}$.

Definition 2.5. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 2.6. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\mathbf{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\mathbf{1}\mathbf{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

Definition 2.7. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

Let the standardized random variables

$$Z_j = \frac{x_j - \overline{x}_j}{\sqrt{S_{jj}}}$$

for $j = 1, ..., p$. Then the sample correlation matrix $\boldsymbol{R}$ is the sample covariance matrix of the $\boldsymbol{z}_i = (Z_{i1}, ..., Z_{ip})^T$ where $i = 1, ..., n$.

Often it is useful to standardize variables with a robust location estimator and a robust scale estimator. The R function `scale` is useful. The R code below shows how to standardize using

$$Z_j = \frac{x_j - \text{MED}(x_j)}{\text{MAD}(x_j)}$$

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the jth variable: $x_{1j}, ..., x_{nj}$. See Definitions 1.3 and 1.5. Some of these results are illustrated with the following R code.

```
x <- buxx[,1:3]; cov(x)
                len       nasal     bigonal
len     118299.9257 -191.084603 -104.718925
nasal     -191.0846   18.793905   -1.967121
bigonal   -104.7189   -1.967121   36.796311

cor(x)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000
z <- scale(x)
cov(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
cov(z)    #in the length variable
                len      nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
```

```
bigonal   -6.981262  -0.218569  4.088479

cor(z)
                len       nasal     bigonal
len       1.00000000 -0.12815187 -0.05019157
nasal    -0.12815187  1.00000000 -0.07480324
bigonal  -0.05019157 -0.07480324  1.00000000

apply(z,2,median)
len   nasal bigonal
0       0       0
#scaled data has coord. median = (0,0,0)^T
apply(z,2,mad)/1.4826
len   nasal bigonal
1       1       1 #scaled data has unit MAD
```

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of thumb 2.1. Multivariate procedures start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) gave this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. See the paragraph below Theorem 3.7.

The population and sample correlation are measures of the strength of a **linear relationship** between two random variables, satisfying $-1 \leq \rho_{ij} \leq 1$ and $-1 \leq r_{ij} \leq 1$. Let the $p \times p$ sample standard deviation matrix

$$\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}}).$$

Then

$$\boldsymbol{S} = \boldsymbol{DRD}, \tag{2.4}$$

and

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}. \tag{2.5}$$

The inverse covariance matrix or inverse correlation matrix can be used to find the partial correlation $r_{ij,\boldsymbol{x}(ij)}$ between x_i and x_j where $\boldsymbol{x}(ij)$ is the vector of predictors with x_i and x_j deleted where $i \neq j$. This partial correlation is the correlation of x_i and x_j after eliminating the linear effects

of $\boldsymbol{x}(ij)$ from both variables: regress x_i and x_j on $\boldsymbol{x}(ij)$ and get the two sets of residuals, and then find the correlation of the two sets of residuals. If $p \geq 3$ and $\boldsymbol{S}^{-1} = (S^{ij})$, then

$$r_{ij,\boldsymbol{x}(ij)} = \frac{-S^{ij}}{(S^{ii}S^{jj})^{1/2}} = \frac{-r^{ij}}{(r^{ii}r^{jj})^{1/2}}.$$

Srivastava and Khatri(1979, p. 53) proved this result. The second equality holds since $\boldsymbol{R}^{-1} = \boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D} = (r^{ij}) = (S^{ij}\sqrt{S_{ii}}\sqrt{S_{jj}})$.

Some R code illustrating this result is shown below. The function `lsfit` is used to regress x_1 on x_3 and then regress x_2 on x_3. Note that $\boldsymbol{x}(i = 1, j = 2) = x_3$ once x_1 and x_2 have been deleted since $p = 3$.

```
x <- buxx[,1:3]; z<-solve(cor(x))
z #inverse correlation matrix

                len      nasal    bigonal
len     1.02042523 0.13535798 0.06134196
nasal   0.13535798 1.02358206 0.08336109
bigonal 0.06134196 0.08336109 1.00931453

out1 <- lsfit(x[,3],x[,1])$resid
out2 <- lsfit(x[,3],x[,2])$resid
cor(out1,out2)
[1] -0.1324439

-z[1,2]/sqrt(z[1,1]*z[2,2])
[1] -0.1324439

zz <- solve(var(x)) #inverse covariance matrix
-zz[1,2]/sqrt(zz[1,1]*zz[2,2])
[1] -0.1324439
```

2.3 Mahalanobis Distances

Definition 2.8. Let $\boldsymbol{A}$ be a positive definite symmetric matrix. Then the *Mahalanobis distance* of $\boldsymbol{x}$ from the vector $\boldsymbol{\mu}$ is

$$D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}.$$

Typically $\boldsymbol{A}$ is a dispersion matrix. The *population squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{2.6}$$

Estimators of multivariate location and dispersion $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are of interest. The *sample squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}). \tag{2.7}$$

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{2.8}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\,\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e}^T\boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are orthogonal: $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is positive definite iff all of its eigenvalues are positive, and positive semidefinite iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 2.1. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50–51), let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 2.2. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Points $\boldsymbol{x}$ with the same distance $D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}^{-1})$ lie on a hyperellipsoid. Let matrix $\boldsymbol{A}$ have determinant $\det(\boldsymbol{A}) = |\boldsymbol{A}|$. Recall that

$$|\boldsymbol{A}^{-1}| = \frac{1}{|\boldsymbol{A}|} = |\boldsymbol{A}|^{-1}.$$

See Johnson and Wichern (1988, pp. 49–50, 102–103) for the following theorem.

Theorem 2.3. Let $h > 0$ be a constant, and let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Then $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\} =$

$$\{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{A}^{-1}) \leq h^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{A}^{-1}) \leq h\}$$

defines a hyperellipsoid centered at $\boldsymbol{\mu}$ with volume

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{A}|^{-1/2} h^p.$$

Let $\boldsymbol{\mu} = \boldsymbol{0}$. Then the axes of the hyperellipsoid are given by the eigenvectors $\boldsymbol{e}_i$ of $\boldsymbol{A}$ with half length in the direction of $\boldsymbol{e}_i$ equal to $h/\sqrt{\lambda_i}$ for $i = 1, ..., p$.

In the following theorem, the shape of the hyperellipsoid is determined by the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$: $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Note $\boldsymbol{\Sigma}^{-1}$ has the same eigenvectors as $\boldsymbol{\Sigma}$ but eigenvalues equal to $1/\lambda_i$ since $\boldsymbol{\Sigma e} = \lambda \boldsymbol{e}$ iff $\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma e} = \boldsymbol{e} = \boldsymbol{\Sigma}^{-1} \lambda \boldsymbol{e}$. Then divide both sides by $\lambda > 0$ since $\boldsymbol{\Sigma} > 0$ and is symmetric. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at squared distance $\boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors of $\boldsymbol{\Sigma}$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. Taking $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$ or $\boldsymbol{A} = \boldsymbol{S}^{-1}$ in Theorem 2.3 gives the volume results for the following two theorems.

Theorem 2.4. Let $\boldsymbol{\Sigma}$ be a positive definite symmetric matrix, e.g., a dispersion matrix. Let $U = D_{\boldsymbol{x}}^2 = D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The hyperellipsoid

$$\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\},$$

where $h^2 = u_{1-\alpha}$ and $P(U \leq u_{1-\alpha}) = 1 - \alpha$, is the highest density region covering $1 - \alpha$ of the mass for an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution (see Definitions 3.2 and 3.3) if g is continuous and decreasing. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at a squared distance $\boldsymbol{w}^T \boldsymbol{S}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\boldsymbol{e}_i$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{\Sigma}|^{1/2} h^p.$$

Theorem 2.5. Let the symmetric sample covariance matrix $\boldsymbol{S}$ be positive definite with eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p > 0$. The hyperellipsoid

$$\{\boldsymbol{x} | D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$$

is centered at $\overline{\boldsymbol{x}}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{S}|^{1/2} h^p.$$

Let $\boldsymbol{w} = \boldsymbol{x} - \overline{\boldsymbol{x}}$. Then points at a squared distance $\boldsymbol{w}^T \boldsymbol{S}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$.

From Theorem 2.5, the volume of the hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\}$ is proportional to $|\boldsymbol{S}|^{1/2}$ so the squared volume is proportional to $|\boldsymbol{S}|$. Large $|\boldsymbol{S}|$ corresponds to large volume while small $|\boldsymbol{S}|$ corresponds to small volume.

Definition 2.9. The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

Following Johnson and Wichern (1988, pp. 103–106), a generalized variance of zero is indicative of extreme degeneracy, and $|\boldsymbol{S}| = 0$ implies that at least one variable X_i is not needed given the other $p - 1$ variables are in the multivariate model. Two necessary conditions for $|\boldsymbol{S}| \neq 0$ are $n > p$ and that $\boldsymbol{S}$ has full rank p. If $\boldsymbol{1}$ is an $n \times 1$ vector of ones, then

$$(n-1)\boldsymbol{S} = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T),$$

and $\boldsymbol{S}$ is of full rank p iff $\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T$ is of full rank p.

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 2.1 was made with the *Arc* software of Cook and Weisberg (1999a). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 2.1a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

Note that the ellipsoids are narrow with high positive correlation. In Figure 2.1b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

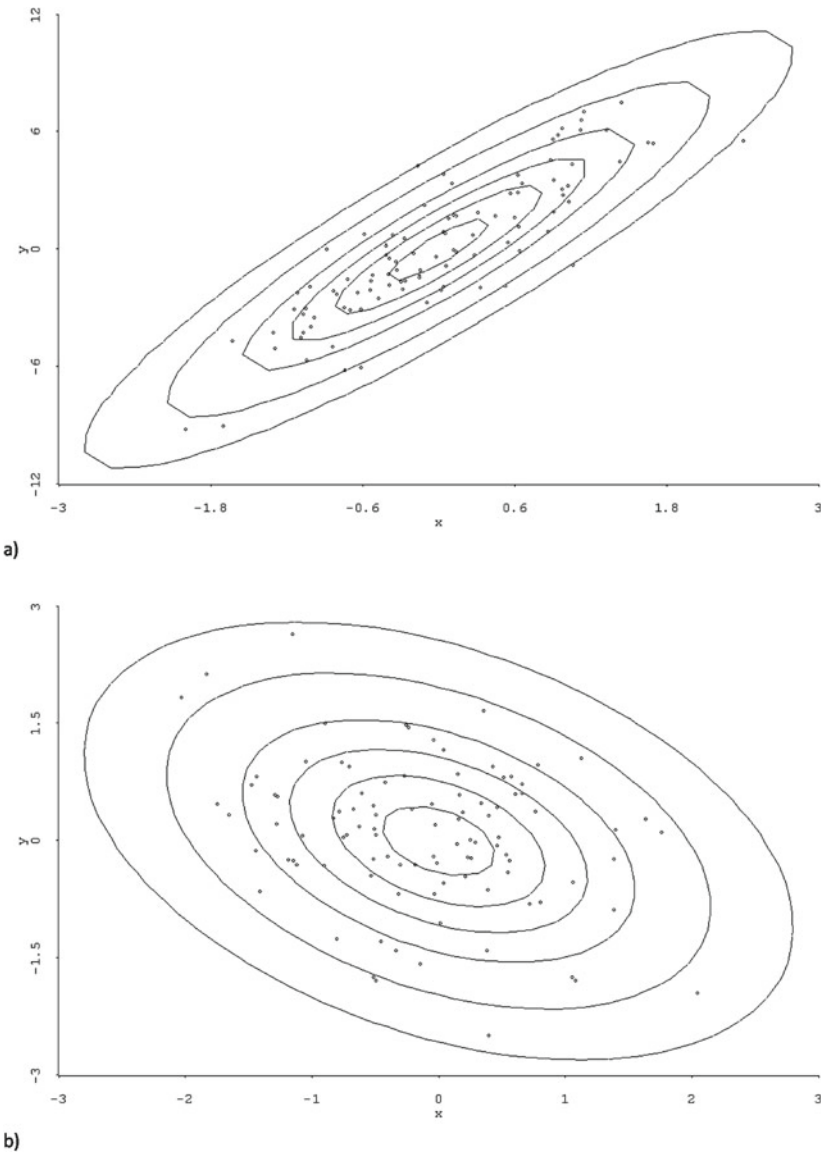

Fig. 2.1 Highest Density Regions for 2 MVN Distributions

2.4 Predictor Transformations

In regression, there is a response variable $w_1 = Y$ of interest, and predictor variables $w_2, ..., w_p$ are used to predict Y. In multivariate analysis, all p random variables $x_1, ..., x_p$ are of interest.

Predictor transformations are used to remove gross nonlinearities in the predictors w_i or the random variables x_i, and this technique is often very useful. Power transformations are particularly effective, and the techniques of this section are often useful for general regression problems, not just for multivariate analysis. A power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. The *modified power transformation* also has $x = t_0(w) = \log(w)$, but for $\lambda \neq 0$,

$$x = t_\lambda(w) = \frac{w^\lambda - 1}{\lambda}.$$

For both the power and modified power transformations, often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{2.9}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder," e.g., from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, e.g., if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as ± 2 and ± 3 can always be added.

Definition 2.10. A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal bivariate relationships between the random variables.

Often nine or ten variables can be placed in a scatterplot matrix. The names of the variables appear on the diagonal of the scatterplot matrix. The software *Arc* gives two numbers, the minimum and maximum of the variable, along with the name of the variable. The software *R* labels the values of each variable in two places; see Example 2.2 below. Let one of the variables be W. All of the marginal plots above and below W have W on the horizontal axis. All of the marginal plots to the left and the right of W have W on the vertical axis.

If n is large and the p random variables come from an elliptically contoured distribution, then the subplots in the scatterplot matrix should be linear. Nonlinearities suggest that the data does not come from an elliptically contoured distribution. There are several rules of thumb that are useful for visually selecting a power transformation to remove nonlinearities from the random variables.

Rule of thumb 2.2. a) If strong nonlinearities are apparent in the scatterplot matrix of the random variables $x_1, ..., x_p$, it is often useful to remove the nonlinearities by transforming the random variables using power transformations.

b) Use theory if available.

c) Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Assume that all values of X_1 and X_2 are positive. Then the following six rules are often used.

d) The **log rule** states that a positive predictor that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $X > 0$ and $\max(X)/\min(X) > 10$ suggests using $\log(X)$.

e) The **range rule** states that a positive predictor that has the ratio between the largest and smallest values less than two should not be transformed. So $X > 0$ and $\max(X)/\min(X) < 2$ suggests keeping X.

f) The *bulging rule* states that changes to the power of X_2 and the power of X_1 can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power of X_2. If the curve is hollow down (the bulge points up), increase the power of X_2. If the curve bulges toward large values of X_1, increase the power of X_1. If the curve bulges toward small values of X_1, decrease the power of X_1. See Tukey (1977, pp. 173–176).

g) The **ladder rule** appears in Cook and Weisberg (1999a, p. 86).
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

h) If it is known that $X_2 \approx X_1^\lambda$ and the ranges of X_1 and X_2 are such that this relationship is one to one, then

$$X_1^\lambda \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation X_1^λ or $X_2^{1/\lambda}$ will linearize the plot. Note that $\log(X_2) \approx \lambda \log(X_1)$, so taking logs of both variables will also linearize the plot. This relationship frequently occurs if there is a volume present. For example, let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere.

i) The *cube root rule* says that if X is a volume measurement, then the cube root transformation $X^{1/3}$ may be useful.

Theory, if available, should be used to select a transformation. Frequently, more than one transformation will work. For example, if W = weight and X_1

$=$ volume $= (X_2)(X_3)(X_4)$, then W versus $X_1^{1/3}$ and $\log(W)$ versus $\log(X_1) = \log(X_2)+\log(X_3)+\log(X_4)$ may both work. Also if W is linearly related with X_2, X_3, X_4 and these three variables all have length units mm, say, then the units of X_1 are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

Suppose that all values of the variable w to be transformed are positive. The log rule says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. This rule often works wonders on the data, and the log transformation is the most used (modified) power transformation. If the variable w can take on the value of 0, use $\log(w + c)$ where c is a small constant like 1, 1/2, or 3/8.

To use the ladder rule, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. Consider the ladder of powers

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger. For example, if both variables are **right skewed**, then there will be many more cases in the lower left of the plot than in the upper right. Hence small values of both variables need spreading.

Consider the ladder of powers. Often no transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

Example 2.1. Examine Figure 2.2. Let $X_1 = w$ and $X_2 = x$. Since w is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot looks roughly like the northwest corner of a square, then small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 2.2a, small values of w need spreading. Notice that the plotted points bulge up toward small values of the horizontal variable. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 2.2b, large values of x need spreading. Notice that the plotted points bulge up toward large values of the horizontal variable. If the plot looks roughly like the southwest corner of a square, as in Figure 2.2c, then small values of both variables need spreading. Notice that the plotted points bulge down toward small values of the horizontal variable. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 2.2d, small val-

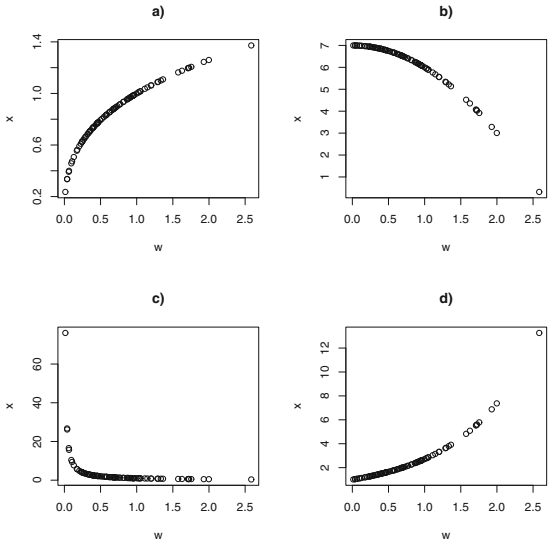

Fig. 2.2 Plots to Illustrate the Bulging and Ladder Rules

ues of x need spreading. Notice that the plotted points bulge down toward large values of the horizontal variable.

Example 2.2. Mussel Data. Cook and Weisberg (1999a, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The response is *muscle mass* M in grams, and the predictors are a constant, the *length* L, *height* H, and the *width* W of the shell in mm, and the *shell mass* S. Figure 2.3 shows the scatterplot matrix of the predictors L, H, W, and S. Examine the variable *length*. Length is on the vertical axis on the three top plots, and the right of the scatterplot matrix labels this axis from 150 to 300. Length is on the horizontal axis on the three leftmost marginal plots, and this axis is labeled from 150 to 300 on the bottom of the scatterplot matrix. The marginal plot in the bottom left corner has length on the horizontal and shell on the vertical axis. The marginal plot that is second from the top and second from the right has height on the horizontal and width on the vertical axis. If the data is stored in x, the plot can be made with the following command in *R*.

```
pairs(x,labels=c("length",'"width","height","shell"))
```

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinear relationship. The minimum value of shell is 10 while the max is 350. Since $350/10 = 35 > 10$, the log rule suggests that $\log S$ may be useful. If $\log S$ replaces S in the scatterplot matrix, then there may be some nonlinearity present in the plot of $\log S$ versus W with small values of W needing spreading. Hence the ladder rule suggests reducing λ from 1, and we tried $\log(W)$.

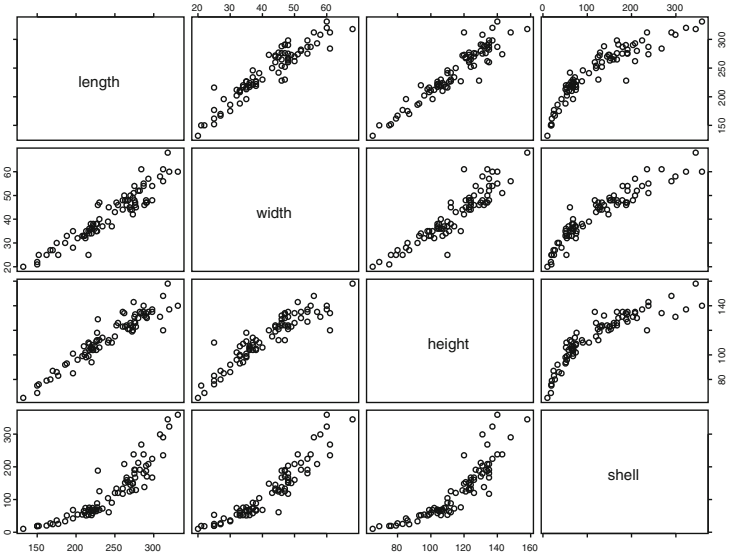

Fig. 2.3 Scatterplot Matrix for Original Mussel Data Predictors

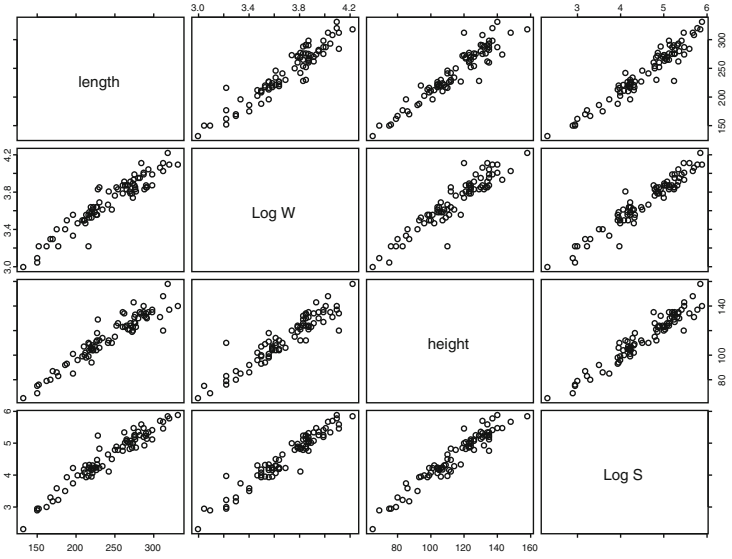

Fig. 2.4 Scatterplot Matrix for Transformed Mussel Data Predictors

Figure 2.4 shows that taking the log transformations of W and S results in a scatterplot matrix that is much more linear than the scatterplot matrix of Figure 2.3. Notice that the plot of W versus L and the plot of $\log(W)$ versus L both appear linear. This plot can be made with the following commands.

```
z <- x; z[,2] <- log(z[,2]); z[,4] <- log(z[,4])
pairs(z,labels=c("length","Log W","height","Log S"))
```

The plot of *shell* versus *height* in Figure 2.3 is nonlinear, and small values of *shell* need spreading since if the plotted points were projected on the horizontal axis, there would be too many points at values of *shell* near 0. Similarly, large values of *height* need spreading.

2.5 Summary

The following three quantities are important.

1) $E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(x_1), ..., E(x_p))^T$.

2) The $p \times p$ *population covariance matrix* $\text{Cov}(\boldsymbol{x}) = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = (\sigma_{ij}) = \boldsymbol{\Sigma_x}$.

3) The $p \times p$ *population correlation matrix* $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho_x} = (\rho_{ij})$.

4) The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma_{x,y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$.

5) Let the $p \times p$ matrix $\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}})$. Then $\boldsymbol{\Sigma_x} = \boldsymbol{\Delta \rho_x \Delta}$, and $\boldsymbol{\rho_x} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma_x} \boldsymbol{\Delta}^{-1}$.

6) The $n \times p$ data matrix

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}.$$

7) The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n} \boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $p \times 1$ vector of ones.

8) The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

9) $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\ \overline{\boldsymbol{x}}^T = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T(\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T) = \boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}$. Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

10) The **sample correlation matrix** $\boldsymbol{R} = (r_{ij})$.

11) Let the $p \times p$ sample standard deviation matrix $\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}})$. Then $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$, and $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}$.

12) The spectral decomposition of the symmetric matrix $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$.

13) Let $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ be a positive definite $p \times p$ symmetric matrix. Let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

14) The population squared Mahalanobis distance $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$.

15) The sample squared Mahalanobis distance $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})$.

16) The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

17) The hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \le h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} |\boldsymbol{S}|^{1/2} h^p.$$

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$. If $\overline{\boldsymbol{x}} = \boldsymbol{0}$, the axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Here $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \ne j$ while $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$.

18) A **scatterplot** of x versus y is used to visualize the conditional distribution of $y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the bivariate relationships of the p random variables.

19) There are several guidelines for **choosing power transformations**. First, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. The **ladder rule:** consider the **ladder of powers**

$$-1, -0.5, -1/3, 0, 1/3, 0.5, \quad \text{and} \quad 1.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

20) Suppose that all values of the variable w to be transformed are positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$.

21) If p random variables come from an elliptically contoured distribution, then the subplots in the scatterplot matrix should be linear.

22) For multivariate procedures with p variables, we want $n \geq 10p$. This rule of thumb will be used for the sample covariance matrix $\boldsymbol{S}$, the sample correlation matrix $\boldsymbol{R}$, and procedures that use these matrices such as principal component analysis, factor analysis, canonical correlation analysis, Hotelling's T^2, discriminant analysis for each group, and one way MANOVA for each group.

2.6 Complements

Section 2.3 will be useful for principal component analysis and for prediction regions. Fan (2017) gave a useful one-number summary of the correlation matrix that acts like a squared correlation.

2.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

2.1. Assuming all relevant expectations exist, show $\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$.

2.2. Suppose $Z_i = \dfrac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$. Show $\mathrm{Cov}(Z_i, Z_j) = \mathrm{Cor}(X_i, X_j)$.

2.3. Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{x})$. Show that $c\boldsymbol{x}$ is also an eigenvector of $\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

2.4. i) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{x})$. Show that $c\boldsymbol{x}$ is also an eigenvector of $\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

ii) Let $\boldsymbol{\Sigma}$ be a $p \times p$ matrix with the eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$. Find the eigenvalue eigenvector pairs of $\boldsymbol{A} = c\boldsymbol{\Sigma}$ where $c \neq 0$ is a real number.

2.5. Suppose $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{e})$. Then $\boldsymbol{A}\boldsymbol{e} = \lambda\boldsymbol{e}$ so $\boldsymbol{A}^2\boldsymbol{e} = \boldsymbol{A}\boldsymbol{A}\boldsymbol{e} = \boldsymbol{A}\lambda\boldsymbol{e}$. Find an eigenvalue eigenvector pair for $\boldsymbol{A}^2$.

2.6. Suppose $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalue eigenvector pair $(\lambda, \boldsymbol{e})$. Then $\boldsymbol{A}\boldsymbol{e} = \lambda\boldsymbol{e}$ so $\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{e} = \boldsymbol{A}^{-1}\lambda\boldsymbol{e}$. Find an eigenvalue eigenvector pair for $\boldsymbol{A}^{-1}$.

Problems using ARC

2.7*. This problem makes plots similar to Figure 2.1. Data sets of $n = 100$ cases from two multivariate normal $N_2(\mathbf{0}, \boldsymbol{\Sigma}_i)$ distributions are generated and plotted in a scatterplot along with the 10%, 30%, 50%, 70%, 90%, and 98% highest density regions where

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Activate *Arc* (Cook and Weisberg 1999a). Generally this will be done by finding the icon for *Arc* or the executable file for *Arc*. Using the mouse, move the pointer (cursor) to the icon and press the leftmost mouse button twice, rapidly. This procedure is known as *double clicking* on the icon. A window should appear with a "greater than" > prompt. The menu *File* should be in the upper left corner of the window. Move the pointer to *File* and hold the leftmost mouse button down. Then the menu will appear. Drag the pointer down to the menu command *load*. Then click on *data* and then click on *demo-bn.lsp*. (You may need to use the *slider bar* in the middle of the screen to see the file *demo-bn.lsp*: click on the arrow pointing to the right until the file appears.) In the future, these menu commands will be denoted by "File > Load > Data > demo-bn.lsp." These are the commands needed to activate the file *demo-bn.lsp*.

a) In the *Arc* dialog window, enter the numbers
0 0 1 4 0.9 and 100. Then click on *OK*.

The graph can be printed with the menu commands "File>Print," but it will generally save paper by placing the plots in the *Word* editor.

Activate *Word* (often by double clicking on the *Word* icon). Click on the screen and type "Problem 2.7a." In *Arc,* use the menu commands "Edit>Copy." In *Word,* click on the *Paste* icon near the upper left corner of *Word* and hold down the leftmost mouse button. This will cause a menu to appear. Drag the pointer down to *Paste*. The plot should appear on the screen. (Older versions of *Word,* use the menu commands "Edit>Paste.") **In the future**, "paste the output into *Word*" will refer to these mouse commands.

b) Either click on *new graph* on the current plot in *Arc* or reload *demo-bn.lsp*. In the *Arc* dialog window, enter the numbers
0 0 1 1 −0.4 and 100. Then place the plot in *Word*.

After editing your *Word* document, get a printout by clicking on the upper left *icon*, select "Print" then select "Print." (Older versions of *Word* use the menu commands "File>Print.")

To save your output on your flash drive G, click on the icon in the upper left corner of *Word*. Then drag the pointer to "Save as." A window will appear, click on the *Word Document* icon. A "Save as" screen appears. Click on the right "check" on the top bar, and then click on "Removable Disk (G:)." Change the file name to HW2d7.docx, and then click on "Save."

To exit from *Word* and *Arc*, click on the "X" in the upper right corner of the screen. In *Word*, a screen will appear and ask whether you want to save changes made in your document. Click on *No*. In *Arc,* click on *OK*.

2.8*. In *Arc* enter the menu commands "File>Load>Data" and open the file *mussels.lsp*. Use the commands "Graph&Fit>Scatterplot Matrix of." In the dialog window, select H, L, S, W, and M (so select M last). Click on "OK" and include the scatterplot matrix in *Word*. The response M is the edible part of the mussel while the 4 predictors are shell measurements. Are any of the marginal predictor relationships nonlinear? Is $E(M|H)$ linear or nonlinear?

2.9*. Activate the McDonald and Schwing (1973) *pollution.lsp* data set with the menu commands "File > Load > Removable Disk (G:) > pollution.lsp." Scroll up the screen to read the data description. Often simply using the log rule on the predictors with $\max(x)/\min(x) > 10$ works wonders.

a) Make a scatterplot matrix of the first nine predictor variables and *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select DENS, EDUC, HC, HOUS, HUMID, JANT, JULT, NONW, NOX, and MORT. Then click on *OK*.

A scatterplot matrix with slider bars will appear. Move the slider bars for NOX, NONW, and HC to 0, providing the log transformation. In *Arc*, the diagonals have the min and max of each variable, and these were the three predictor variables satisfying the log rule. Open *Word*.

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the scatterplot matrix into the *Word* document. Print the graph.

b) Make a scatterplot matrix of the last six predictor variables. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select OVR65, POOR, POPN, PREC, SO, WWDRK, and MORT. Then click on *OK*. Move the slider bar of SO to 0 and copy the plot into *Word*. Print the plot as described in a).

R Problem

Note: For the following problem, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/mrsashw.txt) into *R*.

2.10. Use the following *R* commands to make 100 multivariate normal (MVN) $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC lognormal cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS)*.

Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box- or ball-shaped plot is linear.)