

Chapter 2

Multivariate Distributions

2.1 Introduction

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

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

for a $p \times 1$ random vector \mathbf{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 $P(\mathbf{x} \in A) = \int_A f(\mathbf{z})d\mathbf{z}$ for suitable sets A .

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

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

The notation $f(z_1, \dots, z_p)$ will be used to write out the components z_i of a joint pdf $f(\mathbf{z})$ although in the formula for the pdf, eg $f(\mathbf{z}) = c \exp(\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \mathbf{z})$, \mathbf{z} is a column vector.

Definition 2.2. A $p \times 1$ *random vector* $\mathbf{x} = (x_1, \dots, x_p)^T = (X_1, \dots, X_p)^T$ where X_1, \dots, 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 i th case is $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^T$. There are n cases, and context will be used to determine whether \mathbf{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 $\mathbf{x}_1, \dots, \mathbf{x}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of \mathbf{x}_i is $f(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data \mathbf{x}_i has been observed and stored in an $n \times p$ matrix

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

where the i th row of \mathbf{W} is the i th case \mathbf{x}_i^T and the j th column \mathbf{v}_j of \mathbf{W} corresponds to n measurements of the j th random variable X_j for $j = 1, \dots, p$. Hence the n rows of the data matrix \mathbf{W} correspond to the n cases, while the p columns correspond to measurements on the p random variables X_1, \dots, 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, \mathbf{x}_i will sometimes be a random vector and sometimes the observed data.* Johnson and Wichern (1988, p. 7, 53) uses \mathbf{X} to denote the $n \times p$ data matrix and a $n \times 1$ random vector, relying on the context to indicate whether \mathbf{X} is a random vector or data matrix. Software tends to use different notation. For example, *R/Splus* will use commands such as

`var(x)`

to compute the sample covariance matrix of the data. Hence x corresponds to \mathbf{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

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

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

and the $p \times p$ population covariance matrix

$$\begin{aligned}\text{Cov}(\mathbf{x}) &= E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T] = E[(\mathbf{x} - E(\mathbf{x}))\mathbf{x}^T] = \\ &E(\mathbf{x}\mathbf{x}^T) - E(\mathbf{x})[E(\mathbf{x})]^T = ((\sigma_{i,j})) = \mathbf{\Sigma}_x.\end{aligned}$$

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

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

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

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

Then

$$\mathbf{\Sigma}_x = \mathbf{\Delta}\boldsymbol{\rho}\mathbf{\Delta}, \quad (2.1)$$

and

$$\boldsymbol{\rho} = \mathbf{\Delta}^{-1}\mathbf{\Sigma}_x\mathbf{\Delta}^{-1}. \quad (2.2)$$

Let the population standardized random variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Definition 2.7. The **sample correlation matrix**

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

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

Let the standardized random variables

$$Z_i = \frac{x_i - \bar{x}_i}{\sqrt{S_{ii}}}$$

for $i = 1, \dots, p$. Then \mathbf{R} is the sample covariance matrix of $\mathbf{z} = (Z_1, \dots, Z_p)^T$.

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

$$\mathbf{D} = \text{diag}(\sqrt{S_{11}}, \dots, \sqrt{S_{pp}}).$$

Then

$$\mathbf{S} = \mathbf{D} \mathbf{R} \mathbf{D}, \tag{2.4}$$

and

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

2.3 Distances

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

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

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

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

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

$$D_{\mathbf{x}}^2(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}). \quad (2.7)$$

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

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \quad (2.8)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In the following theorem, the shape of the hyperellipsoid is determined by the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$: $(\lambda_1, \mathbf{e}_1), \dots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$. Note $\boldsymbol{\Sigma}^{-1}$ has the same eigenvectors as $\boldsymbol{\Sigma}$ but eigenvalues equal to $1/\lambda_i$ since $\boldsymbol{\Sigma}\mathbf{e} = \lambda\mathbf{e}$ iff $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{e} = \mathbf{e} = \boldsymbol{\Sigma}^{-1}\lambda\mathbf{e}$. Then divide both sides by $\lambda > 0$ since $\boldsymbol{\Sigma} > 0$ and is symmetric. Let $\mathbf{w} = \mathbf{x} - \boldsymbol{\mu}$. Then points at squared distance $\mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w} = h^2$ 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 \mathbf{e}_i is $h\sqrt{\lambda_i}$. Taking $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ or $\mathbf{A} = \mathbf{S}^{-1}$ 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, eg a dispersion matrix. Let $U = D_{\mathbf{x}}^2 = D_{\mathbf{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The hyperellipsoid $\{\mathbf{x} | D_{\mathbf{x}}^2 \leq h^2\} = \{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq h^2\}$, 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 Definition 3.3) if g is continuous and decreasing. Let $\mathbf{w} = \mathbf{x} - \boldsymbol{\mu}$. Then points at squared distance $\mathbf{w}^T \mathbf{S}^{-1} \mathbf{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes

are given by the eigenvectors \mathbf{e}_i where the half length in the direction of \mathbf{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 \mathbf{S} be positive definite with eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p > 0$. The hyperellipsoid

$$\{\mathbf{x} | D_{\mathbf{x}}^2(\bar{\mathbf{x}}, \mathbf{S}) \leq h^2\} = \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \leq h^2\}$$

is centered at $\bar{\mathbf{x}}$. The volume of the hyperellipsoid is

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

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

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

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

Following Johnson and Wichern (1988, p. 103-106), a generalized variance of zero is indicative of extreme degeneracy, and $|\mathbf{S}| = 0$ 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 $|\mathbf{S}| \neq 0$ are $n > p$ and that \mathbf{S} has full rank p . If $\mathbf{1}$ is an $n \times 1$ vector of ones, then

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

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

If \mathbf{X} and \mathbf{W} 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 $\{\mathbf{x} : (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 2.1 was made with the *Arc* software of Cook and Weisberg (1999). The 10%,

30%, 50%, 70%, 90% and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \mathbf{0}$. 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.

2.4 Predictor Transformations

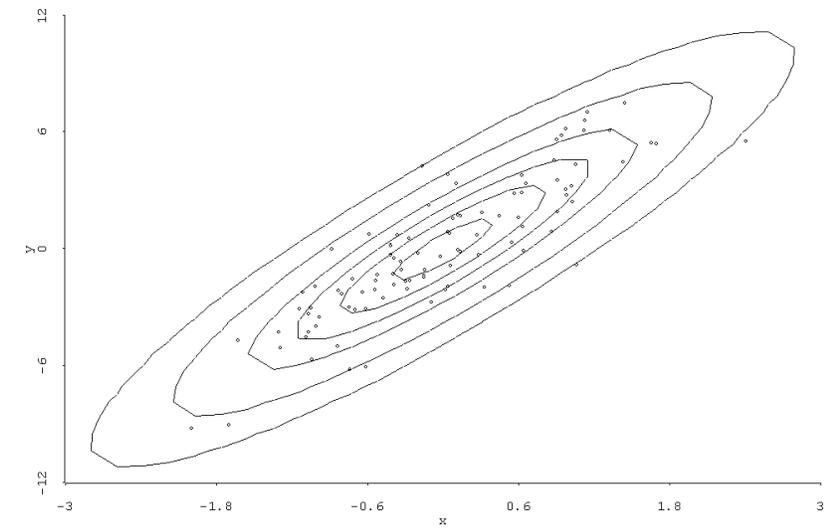
Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful. Power transformations are particularly effective, and the techniques of this section are often useful for general regression problems, not just for multivariate analysis. A power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

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

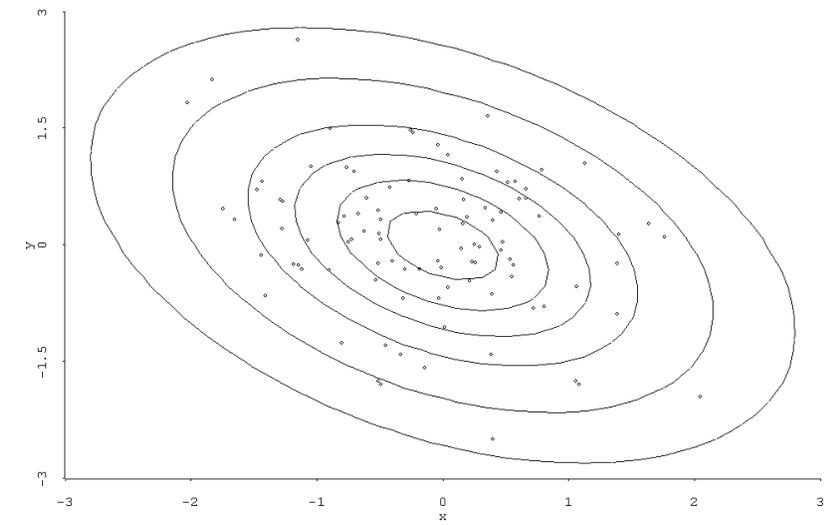
is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes “down the ladder,” eg from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, eg if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then 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 predictors.

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,



a)



b)

Figure 2.1: Highest Density Regions for 2 MVN Distributions
20

along with the name of the variable. The software *R/Splus* 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 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 predictors.

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

Rule of thumb 2.1. a) If strong nonlinearities are apparent in the scatterplot matrix of the predictors w_2, \dots, w_p , it is often useful to remove the nonlinearities by transforming the predictors 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 towards large values of X_1 increase the power of X_1 . If the curve bulges towards small values of X_1 decrease the power of X_1 . See Tukey (1977, p. 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

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

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 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 $W = \text{weight}$ and $X_1 = \text{volume} = (X_2)(X_3)(X_4)$, 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.

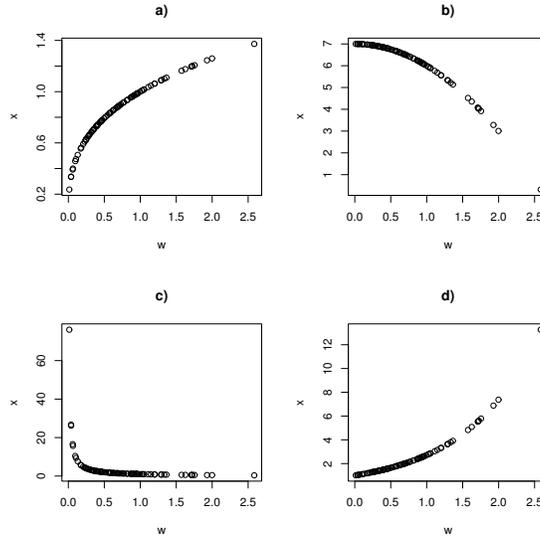


Figure 2.2: Plots to Illustrate the Bulging and Ladder Rules

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

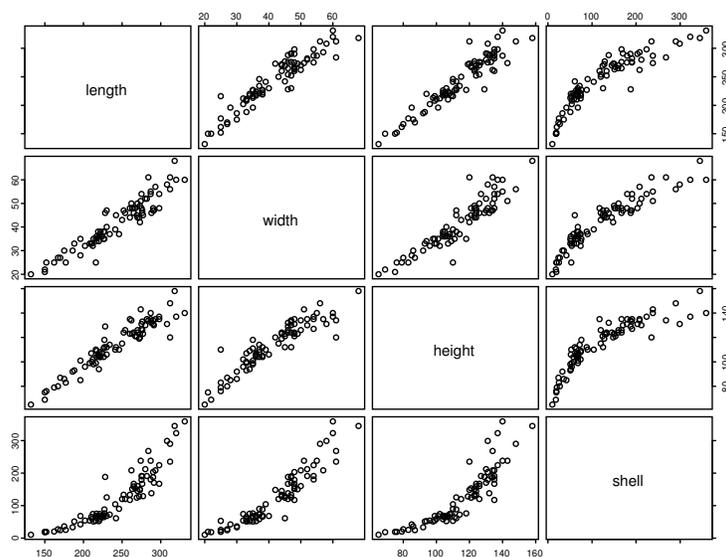


Figure 2.3: Scatterplot Matrix for Original Mussel Data Predictors

Example 2.2: Mussel Data. Cook and Weisberg (1999a, p. 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 and *height* H of the shell in mm, the *shell width* W 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 labelled from 150 to 300 on the bottom of the

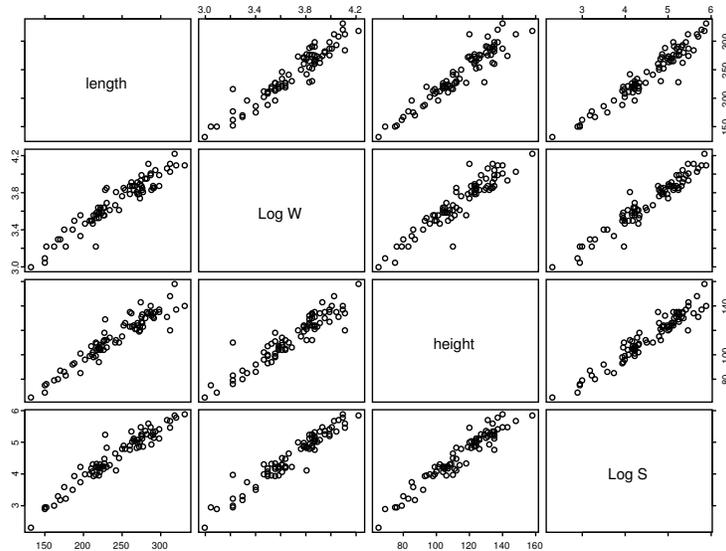


Figure 2.4: Scatterplot Matrix for Transformed Mussel Data Predictors

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 $\log(W)$. 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) $E(\mathbf{x}) = \boldsymbol{\mu} = (E(x_1), \dots, E(x_p))^T$.
- 2) The $p \times p$ population covariance matrix $\text{Cov}(\mathbf{x}) = E(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T = ((\sigma_{i,j})) = \boldsymbol{\Sigma}_x$.
- 3) The $p \times p$ population correlation matrix $\text{Cor}(\mathbf{x}) = \boldsymbol{\rho} = ((\rho_{i,j}))$.
- 4) The population covariance matrix of \mathbf{x} with \mathbf{y} is $\text{Cov}(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Sigma}_{x,y} = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{y} - E(\mathbf{y}))^T]$.
- 5) Let the $p \times p$ matrix $\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$. Then $\boldsymbol{\Sigma}_x = \boldsymbol{\Delta}\boldsymbol{\rho}\boldsymbol{\Delta}$, and $\boldsymbol{\rho} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_x\boldsymbol{\Delta}^{-1}$.
- 6) The $n \times p$ data matrix

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

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

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

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

- 8) The **sample covariance matrix**

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

- 9) $(n-1)\mathbf{S} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T - \bar{\mathbf{x}} \bar{\mathbf{x}}^T = (\mathbf{W} - \mathbf{1}\bar{\mathbf{x}}^T)^T (\mathbf{W} - \mathbf{1}\bar{\mathbf{x}}^T) = \mathbf{W}^T \mathbf{W} - \frac{1}{n} \mathbf{W}^T \mathbf{1} \mathbf{1}^T \mathbf{W}$. Hence if the *centering matrix* $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T$, then $(n-1)\mathbf{S} = \mathbf{W}^T \mathbf{H} \mathbf{W}$.
- 10) The **sample correlation matrix** $\mathbf{R} = ((r_{ij}))$.
- 11) Let the $p \times p$ sample standard deviation matrix $\mathbf{D} = \text{diag}(\sqrt{S_{11}}, \dots, \sqrt{S_{pp}})$. Then $\mathbf{S} = \mathbf{D} \mathbf{R} \mathbf{D}$, and $\mathbf{R} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1}$.
- 12) The spectral decomposition $\mathbf{A} = \sum_{i=1}^p \lambda_i \mathbf{e}_i \mathbf{e}_i^T = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \dots + \lambda_p \mathbf{e}_p \mathbf{e}_p^T$.
- 13) Let $\mathbf{A} = \sum_{i=1}^p \lambda_i \mathbf{e}_i \mathbf{e}_i^T$ be a positive definite $p \times p$ symmetric matrix. Let $\mathbf{P} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \dots \ \mathbf{e}_p]$ be the $p \times p$ orthogonal matrix with i th column \mathbf{e}_i . Let $\mathbf{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$. The *square root matrix* $\mathbf{A}^{1/2} = \mathbf{P} \mathbf{\Lambda}^{1/2} \mathbf{P}^T$ is a positive definite symmetric matrix such that $\mathbf{A}^{1/2} \mathbf{A}^{1/2} = \mathbf{A}$.
- 14) The population squared Mahalanobis distance $D_{\mathbf{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$.
- 15) The sample squared Mahalanobis distance $D_{\mathbf{x}}^2(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})$.
- 16) The *generalized sample variance* $= |\mathbf{S}| = \det(\mathbf{S})$.
- 17) The hyperellipsoid $\{\mathbf{x} | D_{\mathbf{x}}^2 \leq h^2\} = \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \leq h^2\}$ is centered at $\bar{\mathbf{x}}$ and has volume is

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

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

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**

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

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.

2.6 Complements

Section 2.3 will be useful for principal component analysis and for prediction regions.

2.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

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

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

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

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

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

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

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

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

Problems using ARC

To quit *Arc*, move the cursor to the \mathbf{x} in the northeast corner and click.

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

$$\Sigma_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

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 2.4a.” 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 HW2d4.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/Splus Problems

2.10. Use the following *R/Splus* 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.)