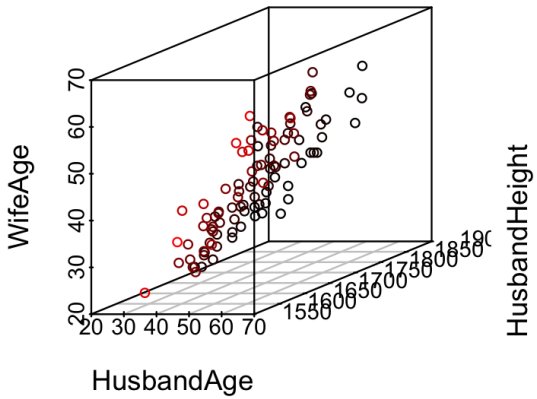
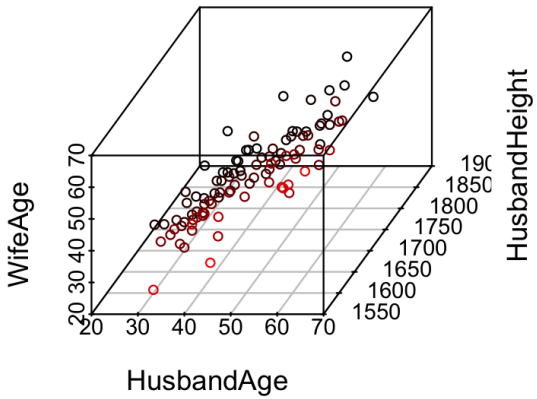
3d scatterplots

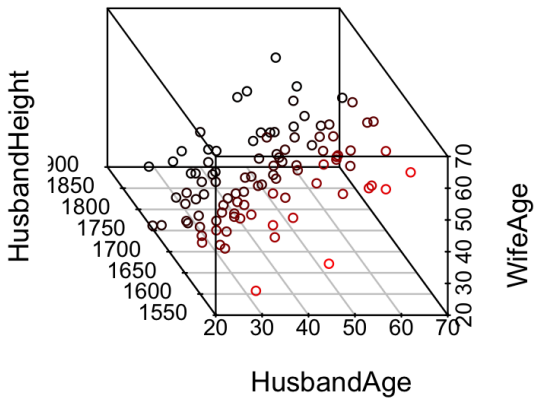
You can also make 3d scatterplots, although these are less common than scatterplot matrices.

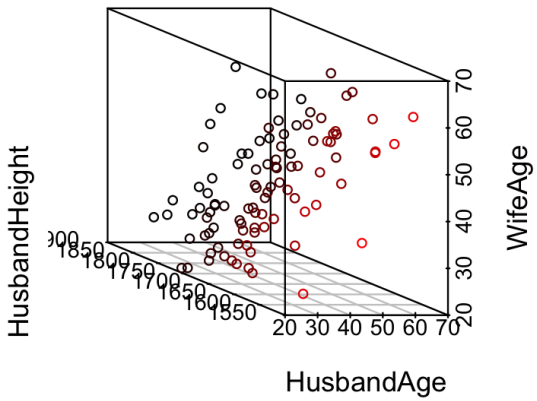
```
> library(scatterplot3d)
> y <- x[,1:3]
> par(mfrow=c(2,2))
> scatterplot3d(y,highlight.3d=T,angle=20)
> scatterplot3d(y,highlight.3d=T,angle=60)
> scatterplot3d(y,highlight.3d=T,angle=120)
> scatterplot3d(y,highlight.3d=T,angle=160)
```

3d scatterplots









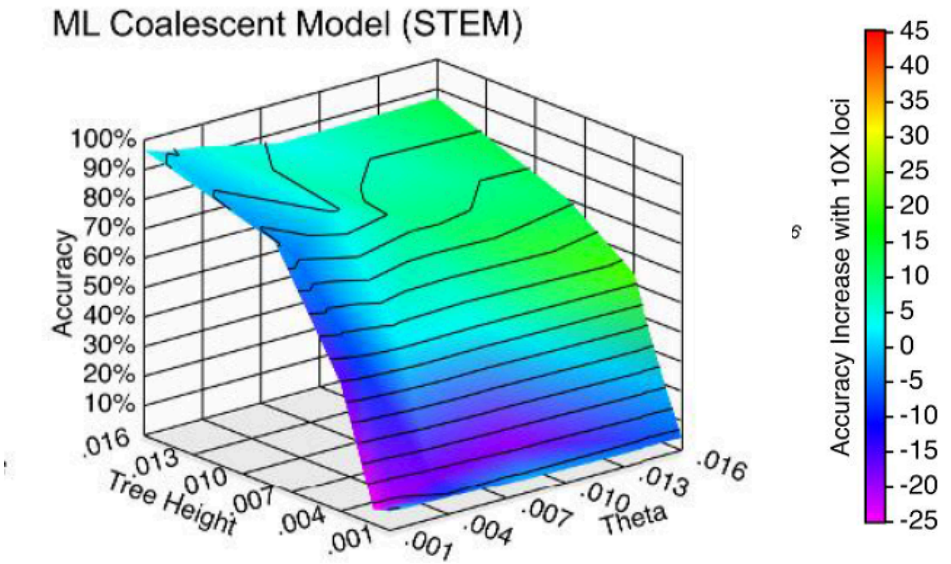
3d scatterplots

If you want to be really fancy you could use color to encode a fourth variable. This would work better if the fourth dimension has a small number of values (e.g., keeping track of different populations, so that color could represent say, country where the individual is from).

A fifth variable could encoded by shape....but putting too much information in one plot is often hard to read.

4d

Here's an exmaple that used color to encode a fourth dimension. The example actually has 4 continuous dimensions.



More creative ways of displaying multiple dimensions

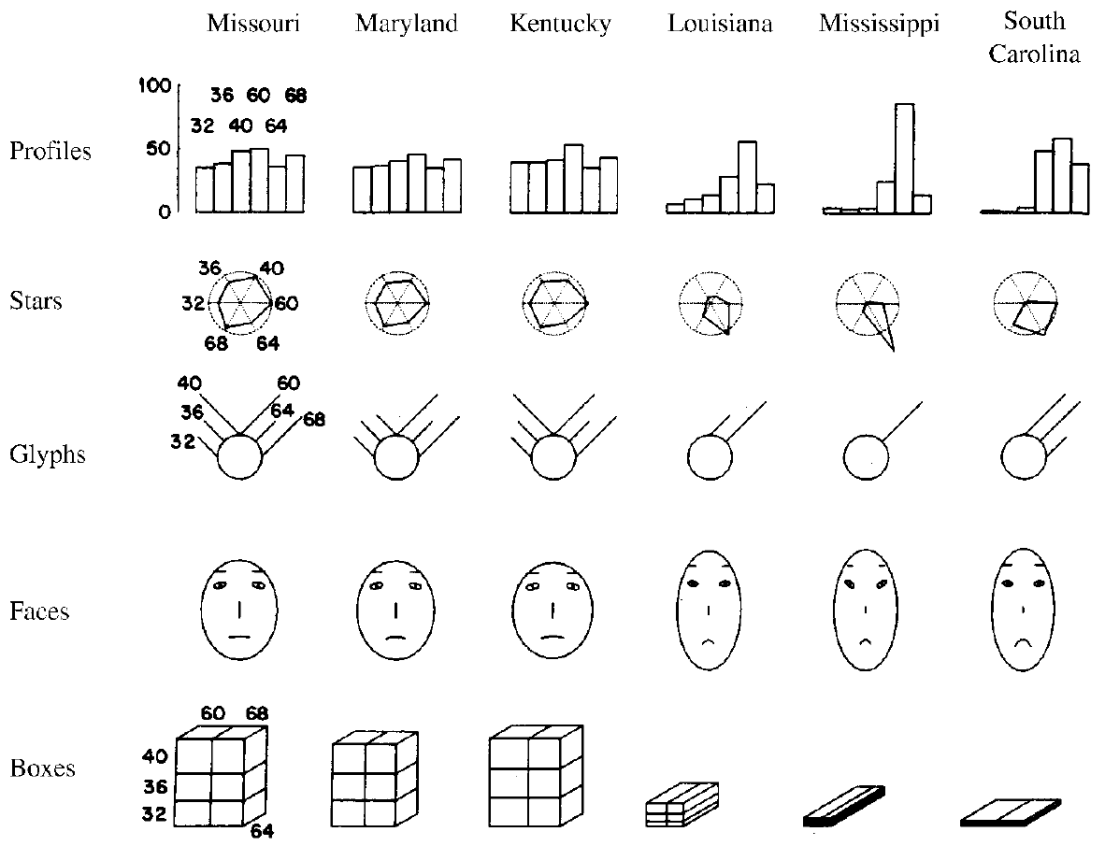


Figure 3.8. Profiles, stars, glyphs, faces, and boxes of percentage of Republican votes in six presidential elections in six southern states. The radius of the circles in the stars is 50%. Assignments of variables to facial features are 1932, shape of face; 1936, length of nose; 1940, curvature of mouth; 1960, width of mouth; 1964, slant of eyes; and 1968, length of eyebrows. (From the *Journal of the American Statistical Association*, 1981, p. 262.)

Correlation and Covariance

Graphs are useful for showing how variables tend to be related. Correlation quantifies the strength of the linear relationship between two variables and puts it on a scale of -1 to 1 with -1 being a perfect linear negative relationship, $+1$ being a perfect linear positive relationship, and 0 being no linear relationship.

Note that nonlinear relationships can have a correlation of 0. For example, if $x = -2, -1, 0, 1, 2$, and $y = x^2$, then x and y are related by a parabola, but have a correlation of 0.

Expectation, Correlation, and Covariance

We also want to be express correlations and covariances in terms of vector and matrix notation. We'll start with expectation.
If you have a vector of random variables

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{pmatrix}$$

then

$$E(\mathbf{y}) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \dots \\ E(y_p) \end{pmatrix}$$

An important property of expectations for a, b, c constant and x, y random:

$$E(ax + by + c) = aE(x) + bE(y) + c$$

Expectation

We use p (rather than n) as the dimension in the previous example because we are interested in p variables that are being measured. Thus the population mean vector is

$$\boldsymbol{\mu} = E(\mathbf{y}) = \begin{pmatrix} \mathbf{y}_1 \\ \dots \\ \mathbf{y}_p \end{pmatrix}$$

Covariance

For two random varianbles, x and y, their covariance is

$$E[(x - E(x))(y - E(y))]$$

The sample covariance is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Properties of the variance and covariance

For a, b, c constants and x, y random,

$$Var(ax + by + c) = a^2 Var(x) + b^2 Var(y) + 2abCov(x, y)$$

If x and y are independent, then their covariance is 0. However, if the covariance is 0, it doesn't follow that x and y are independent.

Correlation

The theoretical correlation between two random variables x and y is

$$\frac{cov(x, y)}{\sigma_x \sigma_y}$$

where σ_x is the standard deviation of x, equal to $\sqrt{E[(x - E(x))^2]}$, and σ_y is the standard deviation of y.
The correlation is estimated by

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

where $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is the sample standard deviation. The correlation can be interpreted as the cosine of the angle between two vectors, which gives a mathematical explanation of why it is always between -1 and 1.

Covariance

We want to generalize the idea of the covariance to multiple (more than two) random variables. The idea is to create a matrix $\mathbf{\Sigma}$ for theoretical covariances and $\mathbf{S}$ for sample covariances of pairwise covariances. Thus for a vector of random variables $\mathbf{y}$, the ijth entry of $\mathbf{S}$ is covariance between variables $\mathbf{y}_i$ and $\mathbf{y}_j$. thus,

$$s_{ij} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y}_i)(y_{ik} - \overline{y}_k) = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_{ij}y_{ik} - n\overline{y}_i\overline{y}_j\right)$$

The diagonal entries of $\mathbf{S}$ are the sample variances. Analogous statements hold for the theoretical covariance matrix $\mathbf{\Sigma}$.

Note that if you plug in $y = x$ for the two-variable covariance (either theoretical or sample-based), you end up with the variance. The covariance formulas generalize the variance formulas.

Covariance

The sample covariance matrix can also be expressed in terms of the observations (rows of the data matrix) as follows:

$$\begin{aligned}\mathbf{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - n\,\overline{\mathbf{y}}\,\overline{\mathbf{y}}'\right)\end{aligned}$$

where $\overline{\mathbf{y}}$ consists of the p column averages of $\mathbf{Y}$, so

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i' = \begin{pmatrix}\overline{y}_1 \\ \overline{y}_2 \\ \dots \\ \overline{y}_p\end{pmatrix}$$

Covariance

The notation here is kind of confusing in terms of the number of dimensions. Here $\mathbf{y}_i$ refers to ith observation vector, meaning the ith row of the data matrix. This is confusing because we used $\mathbf{a}_i$ to refer to the ith column of a matrix $\mathbf{A}$.

We'll work through an example.

Covariance: example

Suppose we have data one y_1 =available soil calcium, y_2 =exchangeable soil calcium, y_3 =turnip green calcium.

Location	y_1	y_2	y_3
1	35	3.5	2.8
2	35	4.9	2.7
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	35	4.6	2.88
8	35	10.9	2.90
9	35	8.0	3.28
10	30	1.6	3.20

Covariance: example

To calculate the sample covariance matrix, we can calculate the pairwise covariances between each of the three variables. Then $s_{i,j} = cov(y_i, y_j)$. There are only three covariances to calculate and three variances to calculate to determine the entire matrix $\mathbf{S}$.

However, let's also try this with using vector notation. The quantities $\mathbf{y}_i$ will be the ith observation, so that $\mathbf{y}_2 = (35, 4.9, 2.7)'$. We write this with the transpose because we think of $\mathbf{y}_i$ as a column vector even though it is the ith row in the data set. Consequently, $\mathbf{y}_i$ is 3×1 for $i = 1, \ldots, 10$.

$\overline{\mathbf{y}}$ is the 3×1 vector of means of the three columns. Thus $\overline{\mathbf{y}}' = (28.1, 7.18, 3.089)'$, and we also think of $\overline{\mathbf{y}}'$ as a column vector.

Covariance: example

To apply the formula in the first form, we have

$$\mathbf{S} = \frac{1}{9}\left[\begin{pmatrix}35\\3.5\\2.8\end{pmatrix} - \begin{pmatrix}28.1\\7.18\\3.089\end{pmatrix}\right]\left[\begin{pmatrix}35\\3.5\\2.8\end{pmatrix} - \begin{pmatrix}28.1\\7.18\\3.089\end{pmatrix}\right]'$$

$$= + \cdots +$$

$$\frac{1}{9}\left[\begin{pmatrix}30\\1.6\\3.2\end{pmatrix} - \begin{pmatrix}28.1\\7.18\\3.089\end{pmatrix}\right]\left[\begin{pmatrix}30\\1.6\\3.2\end{pmatrix} - \begin{pmatrix}28.1\\7.18\\3.089\end{pmatrix}\right]'$$

Covariance: example

Using R as a calculator (which I think is a good way to see some of the details of the matrix calculations),

```
> y1 <- c(35,3.5,2.8)
> ybar <- c(28.1,7.18,3.089)
> y1 - ybar
[1]  6.900 -3.680 -0.289
> t(y1 - ybar)
     [,1]  [,2]   [,3]
[1,]  6.9 -3.68 -0.289
> (y1-ybar) %*% t(y1 - ybar)
         [,1]      [,2]      [,3]
[1,]  47.6100 -25.39200 -1.994100
[2,] -25.3920  13.54240  1.063520
[3,]  -1.9941   1.06352  0.083521
> y10 <- c(30,1.6,3.2)
> (y10-ybar) %*% t(y10 - ybar)
         [,1]      [,2]      [,3]
[1,]   3.6100 -10.60200  0.210900
[2,] -10.6020  31.13640 -0.619380
[3,]   0.2109  -0.61938  0.012321
>
```

Covariance: example

Adding up all of these components and dividing by $n - 1$ (in this case 9) results in the covariance matrix

```
> ((y1-ybar) %*% t(y1-ybar) + (y2-ybar) %*% t(y2-ybar) + (y3-ybar) %*% t(y3-ybar)+
(y4-ybar) %*% t(y4-ybar) + (y5-ybar) %*% t(y5-ybar) + (y6-ybar) %*% t(y6-ybar)+
 (y7-ybar) %*% t(y7-ybar) + (y8-ybar) %*% t(y8-ybar) + (y9-ybar) %*% t(y9-ybar)+
 (y10-ybar) %*% t(y10-ybar))/9
           [,1]      [,2]      [,3]
[1,] 140.544444 49.680000 1.9412222
[2,]  49.680000 72.248444 3.6760889
[3,]   1.941222  3.676089 0.2501211
> cov(t(Y))
           [,1]      [,2]      [,3]
[1,] 140.544444 49.680000 1.9412222
[2,]  49.680000 72.248444 3.6760889
[3,]   1.941222  3.676089 0.2501211
```

Covariance: example

On the previous slide, I computed the covariance directly in R using the `cov` function applied to the matrix `Y`. This matrix could be typed in directly or can be created by "glueing" together the y vectors. Here I typed in all of the y vectors and used the column bind function `cbind` in R:

```
Y <- cbind(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10)
> Y
      y1   y2    y3    y4   y5    y6    y7   y8    y9  y10
[1,] 35.0 35.0 40.00 10.00 6.00 20.00 35.00 35.0 35.00 30.0
[2,]  3.5  4.9 30.00  2.80 2.70  2.80  4.60 10.9  8.00  1.6
[3,]  2.8  2.7  4.38  3.21 2.73  2.81  2.88  2.9  3.28  3.2
```

which looks like the transpose of the original data. Note that for a data matrix like this `cov(Y)` and `cov(t(Y))` result in different matrices because one will be 3×3 and the other is 10×10. One is getting the covariance of the column vectors and the other is getting the covariance of the row vectors.

Covariance

A matrix formula for the covariance is

$$\mathbf{S} = \frac{1}{n-1}\left[\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\left(\frac{1}{n}\mathbf{J}\right)\mathbf{Y}\right] = \frac{1}{n-1}\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$$

This expression is mathematically concise but uses a large matrix since $\mathbf{I} - \frac{1}{n}\mathbf{J}$ is $n \times n$, so would require a lot of memory in the computer for large data sets. (Mathematically equivalent expressions are not necessarily computationally equivalent...)

A matrix expression for the theoretical covariance is

$$\mathbf{\Sigma} = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = E(\mathbf{y}\mathbf{y}') - \boldsymbol{\mu}\boldsymbol{\mu}'$$

Correlation matrices

The correlation matrix $\mathbf{R}$ has entries

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

We can take advantage of matrices to write

$$\mathbf{D}_s = \text{diag}(\sqrt{s_{11}}, \cdots, \sqrt{s_{pp}}) = \text{diag}(s_1, \ldots, s_p)$$

and

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1}$$

since one multiplication divides the ith row by s_i and the other divides the jth column by s_j. From this, we also obtain

$$\mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s$$

Therefore you can obtain the correlation matrix from the covariance matrix and vice versa (if you know the sample variances).

Partitioned vectors and matrices

If a vector is partitioned into two sets of variables, say $\mathbf{y}_1, \ldots, \mathbf{y}_p, \mathbf{x}, \ldots, \mathbf{x}_q$, then we can compute expected values, covariances and correlations using partitioned matrices. We have

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} = \begin{pmatrix}E(\mathbf{y})\\ E(\mathbf{x})\end{pmatrix}$$

$$cov\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix} = \mathbf{\Sigma} = \begin{pmatrix}\mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yx}\\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx}\end{pmatrix}$$

Here $\mathbf{\Sigma}_{yy}$ is the covariance matrix of the y variables only, $\mathbf{\Sigma}_{yx}$ has the covariances of the y variables versus the x variables, and $\mathbf{\Sigma}_{xx}$ is the covariance matrix of the x variables. For this matrix, $\mathbf{\Sigma}_{yx} = \mathbf{\Sigma}'_{xy}$.

Partitioned vectors

Note that if $\mathbf{x}$ and $\mathbf{y}$ are independent, then $\mathbf{\Sigma}_{xy} = \mathbf{0}$, and the covariance matrix has block diagonal structure.

This structure could arise when you have measurements on different families, where different families are assumed to be independent, but individuals within families are not independent of one another for measurements like height, weight, age, blood pressure, etc.

For more than two families, you'd want to partition your random variables into multiple sets (one for each family), and you can work with partitioned vectors and matrices that generalizes the case of two partitions.

Linear combinations of variables

Sometimes we want to work with a linear combination of variables. For example, in Galton's data set, you could have each row contain a family and the columns could be the y_1 = height of the son, y_2 = height of the father, and y_3 = height of the mother. Galton used the average of the father's height and 1.08 times the mother's height. So Galton was interested in the relationship between y_1 and $(y_2 + 1.08y_3)/2$.

More generally, if you have variables $y_1, \ldots, y_p$, then the linear combination

$$z = a_1 y_1 + \cdots a_p y_p$$

can be written as

$$z = \mathbf{a}'\mathbf{y}$$

Linear combinations of variables

If we let the ith observation and jth variable be y_{ij}, then the ith linear combination is

$$z_i = \sum_{j=1}^{p} a_j y_{ij} = \mathbf{a}' \mathbf{y}_i$$

where $\mathbf{y}_i$ is the ith observation vector (row in the data set), but is treated as a column vector. Since $\mathbf{a}$ and $\mathbf{y}_i$ are $p \times 1$, z_i is 1×1.

Linear combinations of variables

The average of the linear combinations is

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \mathbf{a}'\bar{\mathbf{y}}$$

Recall that $\bar{\mathbf{y}}$ is a vector of the p column averages of the data set. Thus, we can compute the average using the z_i terms or directly from the original data.
Similarly, the sample variance of the z_i terms is

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$$

Linear combinations of variables

Here, if you have a univariate random variable y, then $var(ay) = a^2 var(y)$. This true for both theoretical variances and sample variances (if you multiple all samples by a, the sample variance is multiplied by a factor a^2). The matrix analogue for this is that

$$\text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\text{var}(\mathbf{y})\mathbf{a}$$

Again, the result is a scalar (single real number), and the same idea applies to both the sample variance and theoretical variance, so you can replace var($\mathbf{y}$) with either $\mathbf{S}$ or $\mathbf{\Sigma}$.

Linear combinations of variables

Because $s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 \geq 0$, it's also the case that $\mathbf{a}'\mathbf{S}\mathbf{a} \geq 0$, and therefore this shows that $\mathbf{S}$ is positive semidefinite, and positive deniteness holds (meaning the inequality is strict) with probability 1 if $n - 1 > p$ and the variables are continuous.

In practice, the probability 1 result won't hold because our measurements have finite precision, so you could get unlikely and end up with a covariance matrix that wasn't full rank. Statements claiming that something holds with probability 1 mean that there are outcomes in the sample space where the condition wouldn't hold, but this events have probability 0. For example, if x and y are both standard normal, then the event that $x = 0$ and $y = 0$ could occur based on the underlying sample space, yet $P(x = y = 0) = 0$.

Linear combinations of variables

If $w = \mathbf{b}'\mathbf{y}$ is a second linear combination, then z and w are two distinct random variables which might or might not be correlated. Their covariance is

$$s_{zw} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(w_i - \overline{w})^2}{n-1} = \mathbf{a}'\mathbf{S}\mathbf{b}$$

and the sample correlation is

$$r_{zw} = \frac{s_{zw}}{s_z s_w} = \frac{\mathbf{a}'\mathbf{S}\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{S}\mathbf{b}}}$$

Extensions to more than two variables

If you want to transform your $n \times p$ data matrix into something that is, say $n \times k$ with $k < p$, then you can transform your data with a matrix multiplication. Let

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i$$

Here $\mathbf{A}$ is $k \times p$ and $\mathbf{y}_i$ is $p \times 1$, so $\mathbf{z}_i$ is $k \times 1$. The point of doing this is to reduce the dimension of your statistical problem, or to remove collinearity in your variables that could cause problems for some statistical procedures. In some cases we also use $k = p$, to rotate the data in a useful way. You could also use

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i + \mathbf{b}$$

for a more general transformation.

Extensions to more than two variables.

The previous results generalize well to this type of problem with

$$\bar{\mathbf{z}} = \mathbf{A}\bar{\mathbf{y}} + \mathbf{b}$$

$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'$$

Theoretical covariances can replace sample covariances by substituting $\mathbf{\Sigma}$ for $\mathbf{S}$:

$$E(\mathbf{A}\mathbf{y} + \mathbf{b}) = \mathbf{A}E(\mathbf{y}) + \mathbf{b}$$

$$cov(\mathbf{A}\mathbf{y} + \mathbf{b}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}'$$

Measures of overall variability

To measure the overall variability in a sample, you can use the **generalized sample variance**, $\det(\mathbf{S}) = |\mathbf{S}|$ and the trace, $\text{tr}(\mathbf{S})$.

The trace gives the sum of the individual sample variances, and therefore ignores covariances. On the other hand $|\mathbf{S}|$ will be 0 (or near 0) if there is multicollinearity (or near multicollinearity) in the data. Otherwise, large values of either of these statistics suggest that the observations are not close to the mean vector $\overline{\mathbf{y}}$. The book says that $\mathbf{y}_1, \ldots, \mathbf{y}_p$ are not close to $\overline{\mathbf{y}}$, but this appears to be a typo, and it should say $\mathbf{y}_1, \ldots, \mathbf{y}_n$ instead, since the issue is whether the observations are close to the average observation.

Distance between vectors

In univariate statistics, you might be interested in seeing whether a value is unusual. For example, to see how unusual it is for someone to be say, 76 inches tall (6 foot 4), knowing the mean value (say 70 inches) only tells us that the person is 6 inches above average in height, but not how unusual this is.

To measure how unusual an observation is, we get the z-score, which standardizes the observation:

$$z = \frac{x - \mu}{\sigma} = (x - \mu)\sigma^{-1}$$

or we estimate using

$$z = \frac{x - \overline{x}}{s} = (x - \overline{x})s^{-1}$$

Distance between vectors

For vectors, we want to know how unusual vectors are. This is also done by standardizing, but now using the covariance matrix to get a squared distance

$$D^2 = (\mathbf{y} - \bar{\mathbf{y}})' S^{-1} (\mathbf{y} - \bar{\mathbf{y}})$$

We might replace $\bar{\mathbf{y}}$ with $\boldsymbol{\mu}$ if the mean vector is known and S^{-1} with $\boldsymbol{\Sigma}$ if the covariance matrix is known. More generally, we might define the squared distance between two vectors $\mathbf{y}_i$ and $\mathbf{y}_j$ as

$$D^2 = (\mathbf{y}_i - \mathbf{y}_j)' S^{-1} (\mathbf{y}_i - \mathbf{y}_j)$$

This generalizes the Euclidean squared distance between two vectors:

$$d^2(\mathbf{y}_i, \mathbf{y}_j) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j)$$

Mahalanobis distance

The squared distances are often called (squared) Mahalanobis distances, who proposed them in 1936. The idea is that the similarity between two vectors can take into account the variability in some of the dimensions not being equal. For example, if I measure blood pressure on individuals, then the systolic blood pressure (the higher number) might be more variable than the lower number. So, if I compare three blood pressures:

Individual	systolic	diastolic
1	120	80
2	130	75
3	140	85
4	150	90

Mahalanobis distance

In this case $\bar{\mathbf{y}} = (135, 82.5)'$. Individuals 1 and 2 have squared Euclidean distance from the mean of

$$\sqrt{(120 - 135)^2 + (80 - 82.5)^2} = 15.2$$

$$\sqrt{(130 - 135)^2 + (75 - 82.5)^2} = 12.5$$

Based on their Mahalanobis distances, the two individuals are equally distant from the mean vector $\bar{\mathbf{y}} = (135, 82.5)'$. The reason is that the diastolic blood pressure is more variable and being different from the mean diastolic pressure counts less than being different from the mean systolic blood pressure.

Mahalanobis distance in R

Here's how to calculate the Mahalanobis distances in R for this example:

```
> x <- c(120,130,140,150,80,75,85,90)
> x <- matrix(x,ncol=2,byrow=F)
> mu <- c(mean(x[,1]),mean(x[,2]))
> sqrt(t(x[1,]-mu) %*% solve(cov(x)) %*% (x[1,]-mu))
         [,1]
[1,] 1.47196
> sqrt(t(x[2,]-mu) %*% solve(cov(x)) %*% (x[2,]-mu))
         [,1]
[1,] 1.47196
> sqrt(t(x[3,]-mu) %*% solve(cov(x)) %*% (x[3,]-mu))
           [,1]
[1,] 0.4082483
> sqrt(mahalanobis(x,mu,cov(x)))
[1] 1.4719601 1.4719601 0.4082483 1.2247449
```

Distances between vectors

Distances are obtained by taking square roots of these squared distances. Mathematically a distance function d satisfies

$$d(x, y) \geq 0 \text{ with equality if and only if } x = y$$
$$d(x, y) = d(y, x) \text{ (symmetry)}$$
$$d(x, y) + d(y, z) \geq d(x, z) \text{ (triangle inequality)}$$

Note that squared distances are not always distances because they can fail to satisfy the triangle inequality. Sometimes functions that fail the triangle inequality but satisfy the other two properties are called **dissimilarity measures** and are useful in statistics also.

Example of a squared distance not being a distance

Let $f(a,b) = (a-b)^2$, where a and b are real numbers. Then let $a = 0.1$, $b = 0.2$, $c = 0.4$. Then

$$f(a,b) + f(b,c) = 0.01 + 0.04 = 0.05, \quad f(a,c) = 0.09 > f(a,b) + f(b,c)$$

Since f doesn't satisfy the triangle inequality, it is not a distance.
However, it could still be used as a measure of dissimilarity.
s On the other hand,

$$g(a,b) = |\log(a) - \log(b)|$$

does define a metric.

Chapter 4: Multivariate Normal Distribution

Just as many statistical methods developed for a univariate response assume normality, many multivariate methods extend the univariate normal distribution to a multivariate normal distribution.

The simplest case is the bivariate normal distribution, in which there are two random variables, say y_1 and y_2, which are correlated. In the bivariate case, the joint distribution of y_1 and y_2 has 5 parameters: μ_1, μ_2, σ_1 (or σ_1^2), σ_2 (or σ_2^2), plus ρ, the correlation between y_1 and y_2.

We'll quickly want to generalize this to more than two random variables. Instead of a correlation, a vector of random variables $\mathbf{y}$ is characterized by a vector of means $\boldsymbol{\mu}$, and a covariance matrix Σ, which includes the variances of the individual variables as well as the covariances.

The normal distribution

The univariate normal has density function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/(2\sigma^2)}$$

It is very common to see this written as

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

without the parentheses in the denominator in the exponent. It is common to see both forms, and the textbook uses the first form. However, the order of operations here isn't very clear. Keep in mind that if you type

```
exp(-(y-mu)^2/2*sigma^2)
```

In R, you will get the wrong number, and

```
exp(-(y-mu)^2/(2*sigma^2))
```

will be correct.

The normal distrbution

The normal distribution goes by several names: The normal, the bell curve, the Gaussian, and it is related to the error function

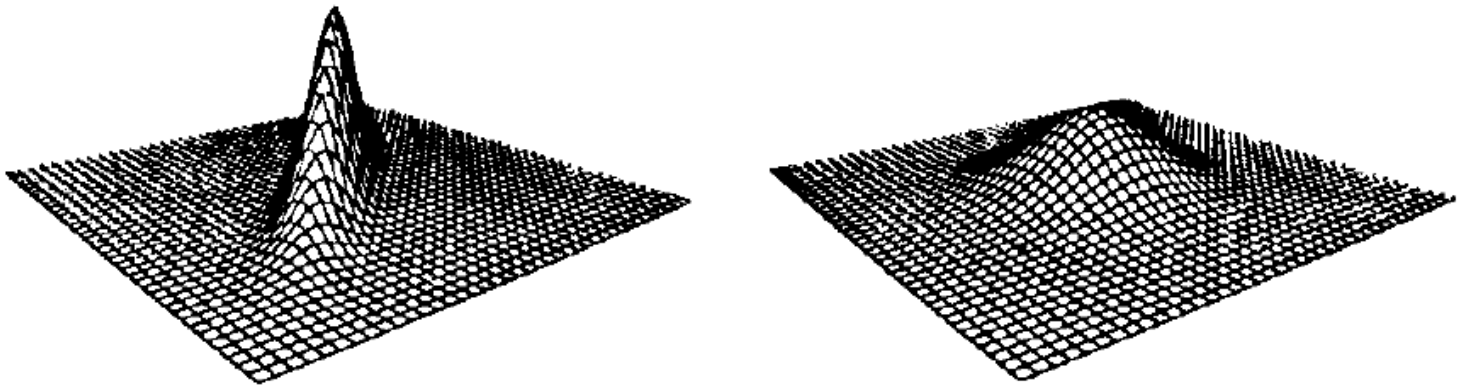
$$\text{erf}(t) = \int_0^x e^{-t^2}\, dt$$

Different names are common in different disciplines; for example Gaussian is often used in engineering. I see this as being like Gandalf from the Lord of the Rings and The Hobbit. He is a central character in the stories and many things to different people, so he gets many names: Gandalf the Grey, Gandalf the White, Mithrandir, Greybeard, Olórin, ...

The normal distribution is like the Gandalf of Statistics...

The bivariate normal distribution

The bivariate normal looks like a hill. The total variance, $|\Sigma|$ afffects how peaked or spread out the distribution is, while the correlation (or covariance) terms affect how symmetrical the distribution is.



(a) small $|\Sigma|$ (b) large $|\Sigma|$

Figure 4.2. Bivariate normal densities.

The bivariate normal distribution

The bivariate normal looks like a hill. The total variance, $|\Sigma|$ afffects how peaked or spread out the distribution is, while the correlation (or covariance) terms affect how symmetrical the distribution is.

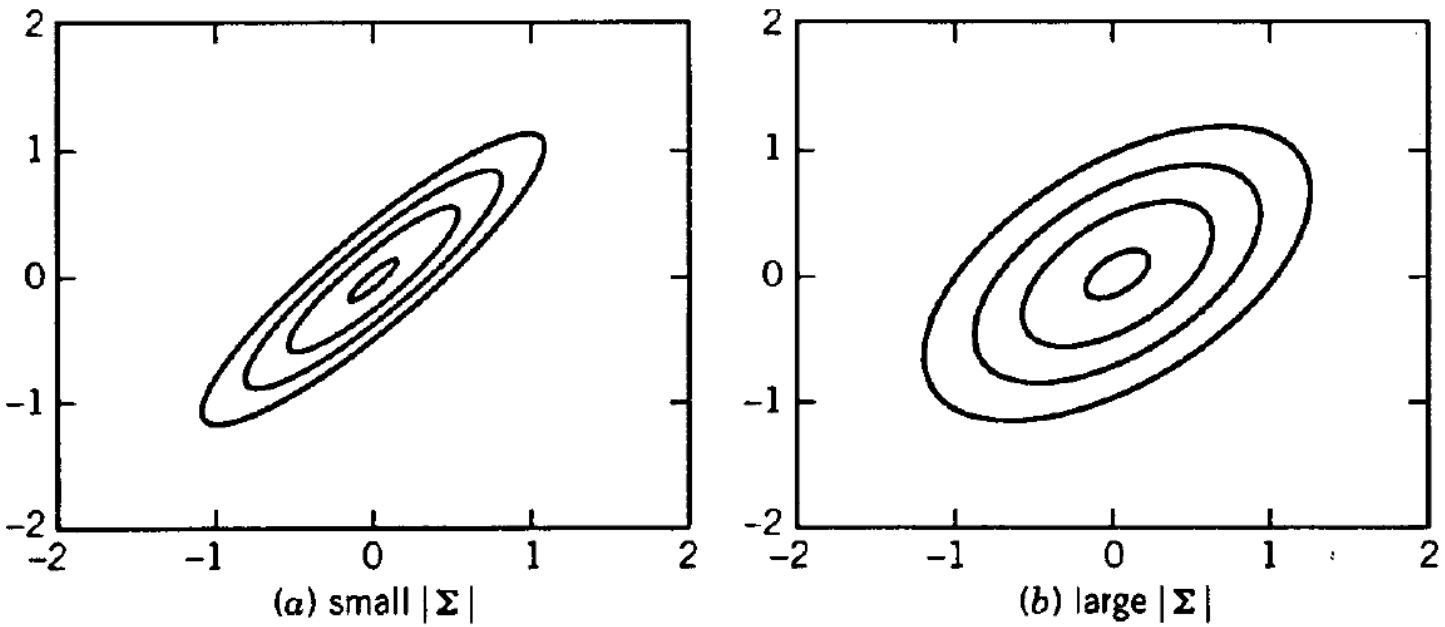


Figure 4.3. Contour plots for the distributions in Figure 4.2.

The bivariate normal distribution

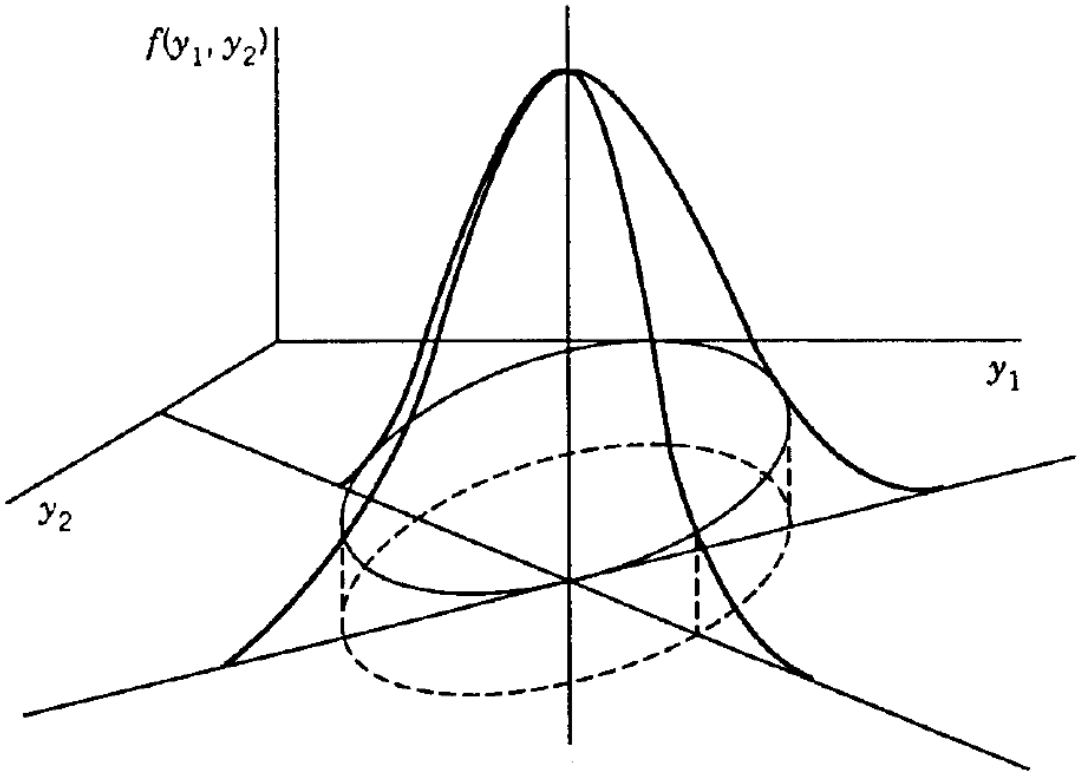


Figure 4.4. Constant density contour for bivariate normal.

The bivariate normal distribution

The bivariate normal density is manageable to write down without using matrices. For more than two variables, we'll use matrix notation, which can also be used in the two variable case. For two variables, the density is

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{z}{2(1-\rho^2)}\right]$$

where

$$z = \frac{(y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}$$

and

$$\rho = cov(y_1, y_2)/(\sigma_1\sigma_2)$$

The bivariate normal distribution

If you let $\tilde{y}_i = \frac{y_i - \mu_i}{\sigma_i}$, then you can write the bivaraite density using

$$z = \tilde{y_1}^2 - 2\rho\, \tilde{y}_1 \tilde{y}_2 + \tilde{y_2}^2$$

so

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[\frac{\tilde{y_1}^2 - 2\rho\, \tilde{y}_1\tilde{y}_2 + \tilde{y_2}^2}{2(1-\rho^2)}\right]$$

The multivariate normal distribution

To express this in matrix and vector form, and to generalize to more than two variables, we write

$$f(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left[-(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})/2\right]$$

Note that

$$(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

is the square of the Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}$.

The multivariate normal and bivariate normal

We might want to check that this is equivalent to the expression we had for the bivariate normal with $p = 2$. To do this, recall that for a 2×2 matrix

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

so

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}$$

The multivariate normal and bivariate normal

After multiplying, we get

$$
\begin{aligned}
&(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})/2 \\
&= \frac{\sigma_2^2(y_1-\mu_1)^2 - 2\sigma_{12}(y_1-\mu_1)(y_2-\mu_2) + \sigma_1^2(y_2-\mu_2)^2}{2(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}
\end{aligned}
$$

With a little manipulation, we can show that this is equivalent to the first version of the bivariate density that we showed.

Properties of the multivariate normal

We can write that a vector is multivariate normal as $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Some important properties of multivariate normal distributions include

1. Linear combinations of the variables $y_1, \ldots, y_p$ are also normal with
Note that for some distributions, such as the Poisson, sums of independent (but not necessarily identically distributed) random variables stay within the same family of distributions. For other distributions, they don't stay in the same family (e.g., exponential random variables). However, it is not clear that sums of two correlated Poissons will still be Poisson. Also, differences between Poisson random variables are not Poisson. For the normal, we have the nice property that even if two (or more) normal random variables are correlated, any linear combinations will still be normal.

Properties of multivariate normal distributions

2. If $\mathbf{A}$ is constant (entries are not random variables) and is $q \times p$ with rank $q \leq p$, then

$$\mathbf{Ay} \sim N_q(\mathbf{A}\mu, \mathbf{A\Sigma A}')$$

What happens if $q > p$?

Properties of multivariate normal distributions

3. A vector $\mathbf{y}$ can be standardized using either

$$\mathbf{z} = (\mathbf{T}')^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

where $\mathbf{T}$ is obtained using the Cholesky decomposition so that $\mathbf{T}'\mathbf{T} = \boldsymbol{\Sigma}$, or

$$\mathbf{z} = (\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

This standardization is similar to the idea of z-scores; however, just taking the usual z-scores of the individual variables in $\mathbf{y}$ will still leave the variables correlated. The standardizations above result in

$$z \sim N_p(\mathbf{0}, \mathbf{I})$$

Properties of multivariate normal distributions

4. Sums of squares of p independent standard normal random variables have a χ^2 distribution with p degrees of freedom. Therefore, if $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\begin{aligned}
(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu}) &= (\mathbf{y}-\boldsymbol{\mu})'\left(\Sigma^{1/2}\Sigma^{1/2}\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}) \\
&= (\mathbf{y}-\boldsymbol{\mu})'\left(\Sigma^{1/2}\right)^{-1}\left(\Sigma^{1/2}\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}) \\
&= \left[\left(\Sigma^{1/2}\right)^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]'\left(\Sigma^{1/2}\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}) \\
&= \mathbf{z}'\mathbf{z}
\end{aligned}$$

Here the vector $\mathbf{z}$ consists of i.i.d. standard normal vectors according to property 3, so $\mathbf{z}'\mathbf{z}$ is a sum of squared i.i.d. standard normals, which is known to have χ^2 distribution. Therefore

$$(\mathbf{y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu}) \sim \chi^2_p$$

Properties of multivariate normal distributions

5. Normality of marginal distributions

If $\mathbf{y}$ has p random variables and is multivariate normal, then any subset $y_{i_1}, \ldots, y_{i_r}$, $r < p$, is also multivariate normal. We can assume that the r variables of interested are listed first so that

$$\mathbf{y}_1 = (y_1, \ldots, y_r)', \quad \mathbf{y}_2 = (y_{r+1}, \ldots, y_p)'$$

Then we have

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

and

$$\mathbf{y}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Marginal distributions

If you have a collection of random variables, and you ignore some of them, the distribution of the remaining is a marginal distribution. For a bivariate random variable $\mathbf{y} = (y_1, y_2)'$, the distribution of y_1 is a marginal distribution of the distribution of $\mathbf{y}$.

In non-vector notation, the joint density for two random variables is often written

$$f_{12}(y_1, y_2)$$

and the marginal distribution can be obtained by

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\, dy_2$$

The joint density for $\mathbf{y}_1$ is

$$f_1(y_1, \ldots, y_r) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, \ldots, y_p)\, dy_{r+1} \cdots dy_p$$

And this is why it is called a marginal density.

Plotting marginal densities in R

```
> install.packages("ade4")
> library(ade4)
> x <- rnorm(100)
> y <- x+ rnorm(100)
> d <- data.frame(x,y)
> s.hist(d)
```

Plotting marginal densities in R