Multivariate Normal Distribution

Arnab Maity

NCSU Department of Statistics ~ 5240 SAS Hall ~ 919-515-1937 ~ amaity[at]ncsu.edu

Contents

Review of univariate normal distribution

We say that Y follows a normal distribution, that is, $Y \sim N(\mu, \sigma^2)$, if the pdf of Y is

$$f_Y(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}, \qquad -\infty < y < \infty.$$

We can show that $E(Y) = \mu$ and $Var(Y) = \sigma^2$. PDF and CDF of normal distribution $N(\mu, \sigma^2)$ for different values of μ and σ^2 are shown in Figure 1.

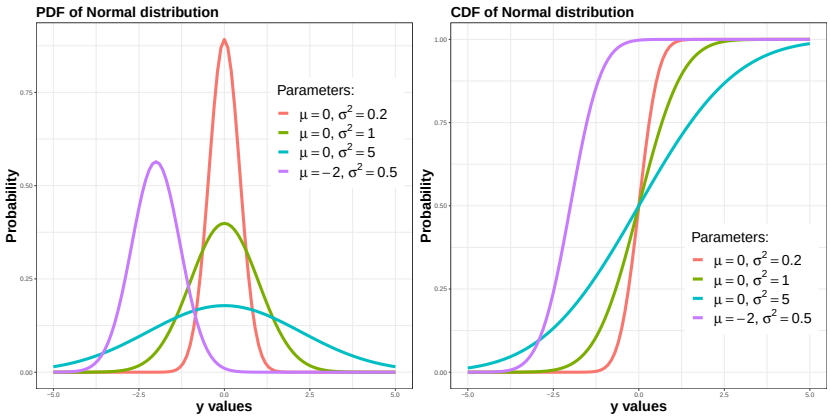


Figure 1: Normal PDF (left panel) and CDF (right panel) for various choices for mean and variance.

Some properties

There are some basic properties of normal distribution:

- **Standard Normal Distribution**: $Z \sim N(0, 1)$. Any normal random variable $Y \sim N(\mu, \sigma^2)$ can be standardized using

$$Z = \sigma^{-1}(Y - \mu).$$

- The function $\phi(\cdot)$ is often used to denote the *pdf of the standard normal distribution*:

$$\phi(t) = (\sqrt{2\pi})^{-1} e^{-t^2/2}.$$

- The function $\Phi(\cdot)$ is often used to denote the *cdf of the standard normal* (no closed form)

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} (\sqrt{2\pi})^{-1} e^{-t^2/2} dt.$$

If for some value z_p, we have $\Phi(z_p) = p$, then z_p is called the *p-quantile* of the standard normal distribution. For example, the *area* of shaded region in Figure 2 is 0.4; this corresponds to Φ(-0.253). Thus -0.253 is the 0.4-quantile.

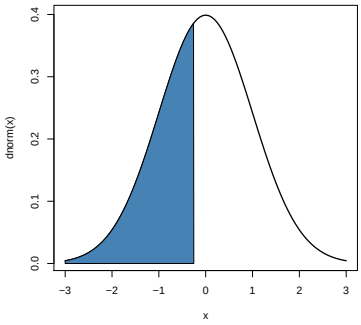


Figure 2: Normal CDF and quantiles

- Any normal distribution can be created from a standard normal distribution using $Y = \mu + Z\sigma$. Specifically, if $Z \sim N(0,1)$ then $\mu + Z\sigma \sim N(\mu, \sigma^2)$.
- Each interval has an associated probality, see Figure 3 for some examples.

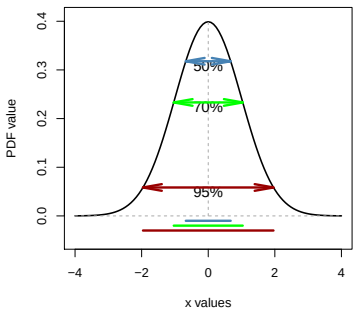


Figure 3: Normal PDF and associated probabilities

R functions

R requires that you specify the mean and standard deviation, rather than mean and variance. The R functions related to normal distribution are

(a) **PDF**: dnorm(x, mean, sd, log = FALSE)
(b) **CDF**: pnorm(q, mean, sd, lower.tail = TRUE, log.p = FALSE)
(c) **Quantiles**: qnorm(p, mean, sd, lower.tail = TRUE, log.p = FALSE)
(d) **Random number**: rnorm(n, mean, sd)

Here the arguments are:

- `x`, `q`: the value at which to compute the probability PMF of CDF
- `mean`: mean μ
- `sd`: standard deviation σ
- `p`: probability, it must be between 0 and 1
- `n`: the number of times to repeat the experiment.
- `log`, `log.p`: logical; if TRUE, probabilities p are given as $log(p)$.
- `lower.tail`: logical; if TRUE (default), probabilities are $P[Y \leq x]$, otherwise, $P[Y > x]$.

Assessing univariate normality

We can use graphical as well as hypothesis testing techniques to assess wheather the normality assumption is reasonable for a dataset. A common graphical technique to check for normality is to create a *normal quantile-quantile plot (Q-Q plot)*.

Normal quantile-quantile plot

A scatterplot of the sorted data, $x_{(1)} \leq \ldots \leq x_{(n)}$, against normal quantiles, $\Phi^{-1}\{(1-0.5)/n\}, \ldots, \Phi^{-1}\{(n-0.5)/n\}$.

If this plot shows a linear pattern, then assumption of normalty is supported.[1]

Let us consider the sepal length of setosa flowers, and create a normal Q-Q plot. We can use the `qqnorm()` function in R.

[1] We can actually create Q-Q plots for *any* distribution. We just need to use the quantiles of that distribution instead of normal.

```r
# Extract only sepal.length of setosa flowers
SL <- iris$Sepal.Length[1:50]

# Q-Q plot to assess normality
qqnorm(SL, pch = 19)
```

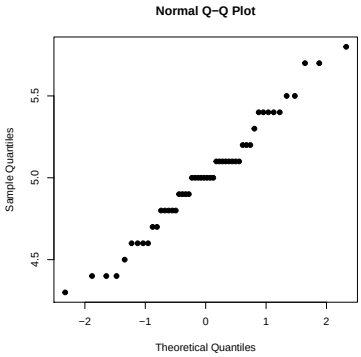


Figure 4: Normal Q-Q plot of sepal length of setosa flowers.

Note that the theoretical quantiles of a $N(0,1)$ distribution are ploted in the x-axis. Since the plot is fairly linear, normality assumption seems reasonable in this case.

We can also employ formal statistical tests to check for normality.

- Shapiro-Wilks test, and Shapiro–Francia test: the later test is a simplification of the former; they show similar power to each other. These two are among the more powerful normality tests.
- Kolmogorov-Smirnov (K-S) test, and Lilliefors corrected K-S test: the later test is usually preferred among the two tests.
- Cramer von Mises test, and Anderson-Darling test (a modification of the CVM test): based on weighted difference between the empirical and theoretical CDFs.
- Person's Chi-squared test: a goodness-of-fit test, not highly recommended for continuous distributions.
- Jarque-Bera test and D'Agostino-Pearson omnibus tests: moment based tests.

The [nortest] package in R implements a few of the tests mentioned above. Overall, Shapiro-Wilk test shows a robust performance against a wide variety of alternatives.[2] We can use the function `shapiro.test()` to perform this test.

[2] See the article [Yap and Sim (2011). Comparisons of various types of normality tests] for a numerical comparison between various tests.

```r
shapiro.test(SL)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SL
## W = 0.9777, p-value = 0.4595
```

Since the p-value is large (e.g., larger than 5%), we can say that normality assumption for the data is pausible.

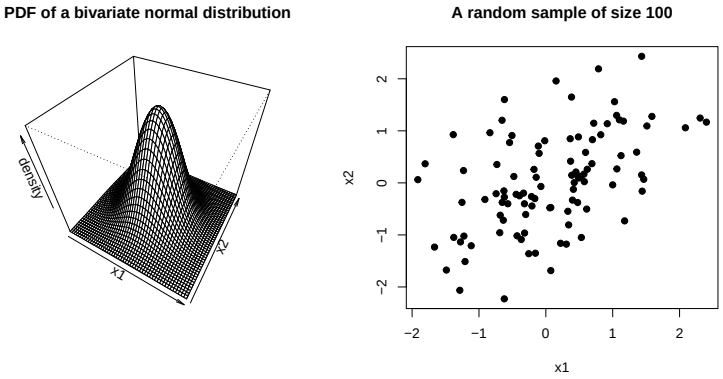
Bivariate and Multivariate normal distributions

The random vector $\boldsymbol{X}_{2\times 1} = (X_1, X_2)^T$ follows a bivariate normal (Gaussian) distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and variance-covariance (positive definite) matrix $\boldsymbol{\Sigma}$ and denoted as $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its probability density function is[3]

$$f(\boldsymbol{x}) = (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\}.$$

[3] Recall the PDF of univariate normal distribution, $N(\mu, \sigma^2)$ is

$$f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$



The shape of the PDF (and that of the scatterplot of a random sample generated from the distribution) is determined by $\boldsymbol{\Sigma}$, the variance-covariance matrix of $\boldsymbol{X}$. An easy was to visualize the PDF of a bivariate distribution is to plot the constant probability density contours.

Constant probability density contours

We define the constant probability density contour (also called constant-density contour) of a bivariate normal PDF to be the set of vectors $\boldsymbol{x}$ such that $f(\boldsymbol{x})$ is constant, that is,

$$\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c\}$$

for a specific c. These sets are ellipses that are centered around $\boldsymbol{\mu}$, and the major and minor axes are $c\sqrt{\lambda_i}\boldsymbol{e}_i$, where λ_i are the eigenvalues and $\boldsymbol{e}_i$ are the corresponding eigenvectors of $\boldsymbol{\Sigma}$.

More generally, a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is said for follow a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a $p \times 1$ vector and $\boldsymbol{\Sigma}$ is positive definite matrix, if the PDF of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\}.$$

We can show that $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and that $cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

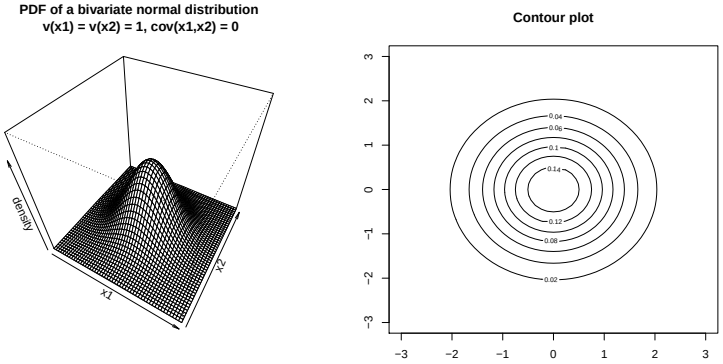


Figure 5: PDF and contours of a bivariate normal distribution with v(x1) = v(x2) = 1, cov(x1,x2) = 0

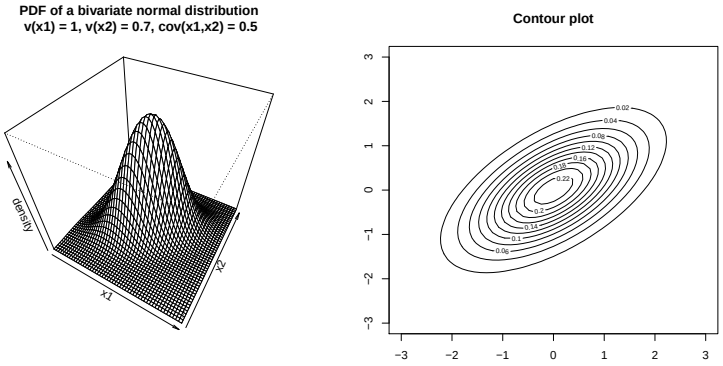


Figure 6: PDF and contours of a bivariate normal distribution with v(x1) = 1, v(x2) = 0.7, cov(x1,x2) = 0.5

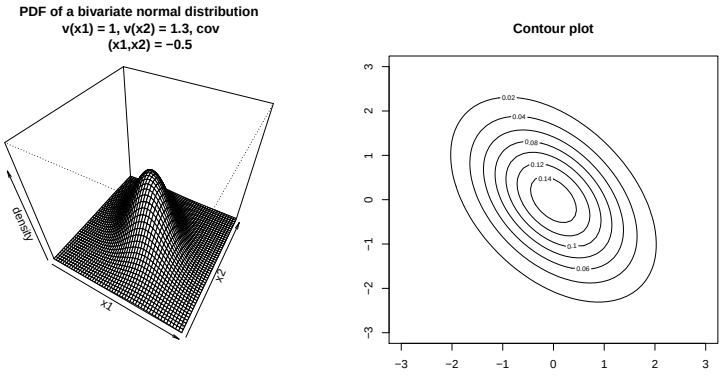


Figure 7: PDF and contours of a bivariate normal distribution with v(x1) = 1, v(x2) = 1.3, cov(x1,x2) = -0.5

Some properties of the multivariate normal distribution

- The constant probability density contours are ellipsoids
- Zero covariance implies the compoments of $\boldsymbol{X}$ are independent (**ONLY** when $\boldsymbol{X}$ is multivariate normal)
- When $\boldsymbol{\mu} = \mathbf{0}_p$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we say that we have a **standard multivariate normal distribution**, $\boldsymbol{Z} \sim N_p(0_p, I_p)$.[4]
- All subsets of $\boldsymbol{X}$ also follow multivariate normal distribution.
- If $\boldsymbol{X}$ follows a multivariate normal distribution, then any linear combination of $\boldsymbol{X}$ follow multivariate normal distribution. Specifically, if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$ for any matrix $\boldsymbol{A}$.
- $(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p$, where χ^2_p is the chi-square distribution with p degrees of freedom.[5]

[4] Compare with univariate standard normal distribution: $Z \sim N(0, 1)$.

[5] Recall that in the univariate case, if $X \sim N(\mu, \sigma^2)$, then

$$(X - \mu)^2 / \sigma^2 \sim \chi^2_1.$$

Mahalanobis distance

The quantity

$$d^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

is called the Mahalanobis squared distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$.

Using the last property, we can compute the probability observing data within any constant-density contours. Specifically, consider the constant-density ellipse $E_c = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c\}$, $c > 0$. Then

$$Pr(\boldsymbol{X} \in E_c) = G_p(c),$$

where $G_p(c)$ is the CDF of a χ^2_p distribution. Figure 8 shows 50% and 90% contours below for two bivariate normal distributions.

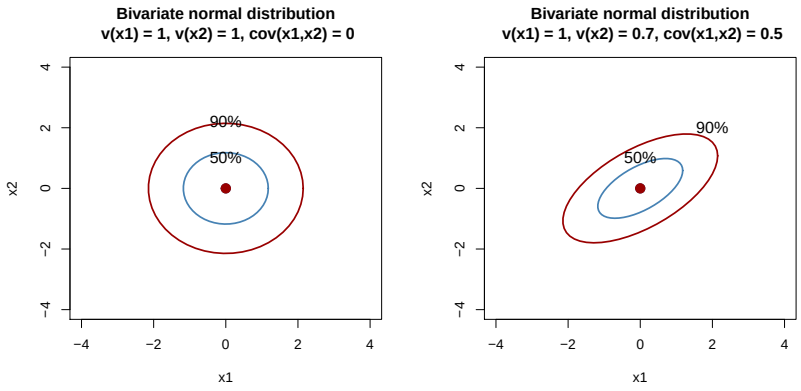


Figure 8: Constant probability density contours (blue:0.50 and red:0.90) for two bivariate normal distributions.

Sampling distribution of $\bar{X}$ and S

Recall that for univariate normal distribution, if $X_1, \ldots, X_n$ form a random sample from $N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \sigma^2/n), \text{ and } \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1},$$

where S^2 is the sample variance. We also know that

$$\bar{X} \text{ and } S^2 \text{ are independent.}$$

We have similar results for multivariare normal distribution.

Exact distribution of $\bar{X}$ and S

Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ form a random sample from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then

- $\bar{\boldsymbol{X}}$ has a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ distribution.
- $(n-1)\boldsymbol{S}$ has a **Wishart** distribution with $n-1$ degrees of freedom (a generalization of χ^2 distribution).
- $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.

Large sample results analogous to univariate normal also exist. Recall that if $X_1, \ldots, X_n$ form a random sample from $N(\mu, \sigma^2)$, then Central Limit Theorem (CLT) says *when n is large enough*

$$\bar{X} \text{ approximately has a } N(\mu, \sigma^2/n) \text{ distribution.}$$

Similar results hold for multivariate normal distribution.

Large sample results

Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ form a random sample from a population (can be different from normal) with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. When the *sample size n is large*,

- $\bar{\boldsymbol{X}}$ has an *approximate* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ distribution (multivariate CLT).
- $(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$ has an *approximate* χ^2_p distribution (also need $n - p$ large; note that we replaced $\boldsymbol{\Sigma}$ with $\boldsymbol{S}$).

Checking multivariate normality

Many of the techniques typically used in multivariate statistics assume that the parent distribution is multivariate normal or that the sample size sufficiently large (in which case the normality assumption is less crucial). However, the quality of the inferences relies on how close the parent distribution is to the multivariate normal. Thus it is essential to validate the normality assumption.

It is difficult to assess multivariate normality. In practice, we investigate the univariate and bivariate distributions to determine how close they are to normality. We describe a few steps for checking multivariate normality below.

Check univariate normality

Usual univariate analysis for each variable, such as normal Q-Q plot and statistical tests for normality can be done. Recall, if $\boldsymbol{X}$ is multivariate normal, then each component is univariate normal as well. If we reject normality for one of the variables, then $\boldsymbol{X}$ can not be multivariate normal.

Let us consider the lumber stiffness dataset[6] where four measures of stiffness $x_1, \ldots, x_4$ are measured of each of the $n = 30$ boards.

[6] Table 4.3 in Johnson and Wichern (2007). Applied Multivariate Analysis.; provided in the course webpage.

```
# Reading the data set
dat <- read.table("data/T4-3.DAT", header = F)
colnames(dat) <- c("x1", "x2", "x3", "x4", "d2")

# Dimensions of the dataset
n <- nrow(dat)
p <- ncol(dat) - 1

# snapshot
head(dat)
```

```
##     x1   x2   x3   x4   d2
## 1 1889 1651 1561 1778 0.60
## 2 2403 2048 2087 2197 5.48
## 3 2119 1700 1815 2222 7.62
## 4 1645 1627 1110 1533 5.21
## 5 1976 1916 1614 1883 1.40
## 6 1712 1712 1439 1546 2.22
```

The first four columns provide the four maesured variables. Let us construct their relative frequency histograms (Figure 9) and normal Q-Q plots (Figure 10).

```
par(mfrow = c(2, 2))
for (ii in 1:4) {
    hist(dat[, ii], probability = T, xlab = paste("x",
        ii, sep = ""), main = paste0("Histogram of x",
        ii))
}
```

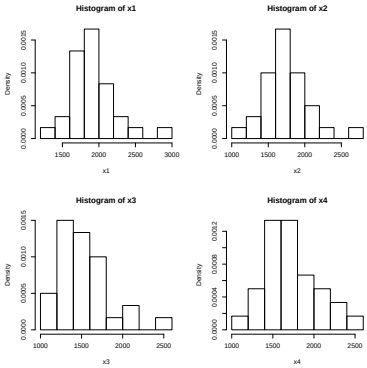


Figure 9: Relative frequency histograms of the four measures of stiffness.

```
par(mfrow = c(2, 2))
for (ii in 1:4) {
    qqnorm(dat[, ii], main = paste0("Q-Q plot of x",
        ii), pch = 19, cex = 1.5)
}
```

These marginal distributions appear somewhat close to normal. However, it seems there might be outliers; notice the point in the upper-right corner of the Q-Q plots.

In general, just checking univariate plots is not enough. Even if individual variables are normally distributed, their joint distribution may not be multivariate normal.

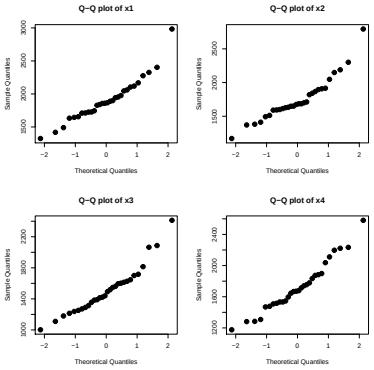


Figure 10: Normal Q-Q plot of the four measures of stiffness.

Check scatterplots

If the data indeed are generated from a normal distribution, the constant-density contours must be ellipses. Thus, the scatterplots should also conform to this structure. Creating scatterplots and `pairs-plot` (pairwise scatterplots) of the variables will also reveal any unusual shape (or outliers) in the data set.

The R function `pairs()` can be used to create pairwise scatterplots. The `pairs-plot` of the dataset is shown in Figure 11

```
pairs(dat[, 1:4], pch = 19)
```

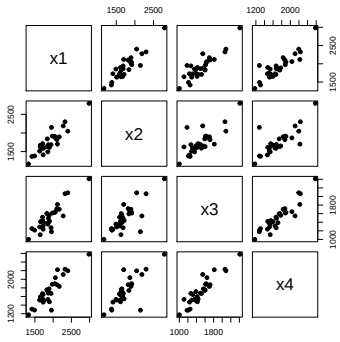


Figure 11: Pairs-plot of the four measures of stiffness.

Overlaying "data ellipses" (constant-density contours estimated from the data assuming normality) on top of scatterplots are useful in this situation. The data ellipses can be drwan using the `dataEllipse` function in the `car` package. Figure 12 shows the 50% and 90% data ellipses overlayed on the scatter plot of X_2 vs. X_1.

```
library(car)

# define x1 and x2
x1 <- dat[, 1]
x2 <- dat[, 2]

# Draw data ellipses
```

```
dataEllipse(x1, x2, xlim = c(1000, 3500), ylim = c(800,
    3000), pch = 19, col = c("steelblue", "#990000"),
    lty = 2, ellipse.label = c(0.5, 0.95), levels = c(0.5,
        0.95), fill = TRUE, fill.alpha = 0.1)
```

By default, the 50% and 95% ellipses are drawn. See the documentation using `?dataEllipse` for more customization options. We can see from Figure 12 that the data cloud does have an elliptical shape. However, there is one point that might be an outlier.

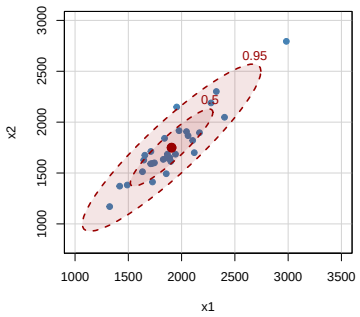


Figure 12: Scatterplot of X2 against X1 with overlayed data ellipses.

Check bivariate distributions

We can also estimate the PDF of each pair of variables. This can be done using the `bkde2D()` function in the `KernSmooth` package. Visualization can be done using `persp()` and `contour()` functions. Figures 13 and 14 show the estimated density funtion and a contour plot of the estimated density, respectively.

```
library("KernSmooth")
# Estimate bivariate density
den.est <- bkde2D(dat[, 1:2], bandwidth = apply(dat[,
    1:2], 2, dpik))

persp(x = den.est$x1, y = den.est$x2, z = den.est$fhat,
    xlab = "x1", ylab = "x2", zlab = "density",
    phi = 45, theta = 30, ticktype = "detailed",
    main = "Estimated PDF of (X1, X2)")


# Contour plot of the estimated density
plot(dat[, 1:2], xlab = "x1", ylab = "x2", pch = 19,
    main = "Contour plot of the estimated PDF")
contour(x = den.est$x1, y = den.est$x2, z = den.est$fhat,
    add = TRUE, col = "#990000")
```

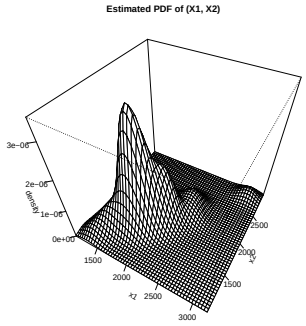


Figure 13: Estimated bivariate density function of X1 and X2

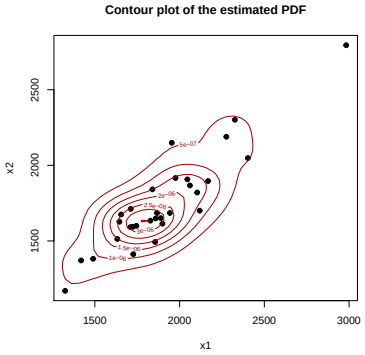


Figure 14: Contour plot of estimated bivariate density function of X1 and X2

Construct a chi-square plot

Given sample data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the chi-square plot is constructed using the following steps:

- For each i, compute the Mahalanobis squared distance

$$d_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{s}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}),$$

  where $\bar{\boldsymbol{x}}$ and $\boldsymbol{s}$ are observed values of the sample mean and covariance matrix, respectively.

- If the data are indeed generted from a normal distribution, then the d_i^2 values should follow a χ_p^2 (in our example, $p = 4$) distribution. Thus, we plot the ordered d_i^2 values,

$$d_{(1)}^2 \leq \ldots \leq d_n^2$$

against the theoretical quantiles of the χ_p^2 distribution

$$q_p(\frac{1-0.5}{n}), \ldots, q_p(\frac{n-0.5}{n}).$$

If the multivariate normality assumption is correct, then the points should follow a straight line. A systematic curved pattern will suggest a departure from normality. One or two points that show large deviations from the linear trend might be outliers and would warrant further investigation.

A function to create such a chi-square plot is shown below.

```
# A function to create a chi-square plot
chisquare.plot <- function(x, mark) {
    # x= a n x p data matrix, mark: number of
    # extreme points to mark

    # p=number of variables, n=sample size
    p <- ncol(x)
    n <- nrow(x)
    # xbar and s
    xbar <- colMeans(x)
    s <- cov(x)
    # Mahalanobis dist, sorted
    x.cen <- scale(x, center = T, scale = F)
    d2 <- diag(x.cen %*% solve(s) %*% t(x.cen))
    sortd <- sort(d2)
    # chi-sq quantiles
    qchi <- qchisq((1:n - 0.5)/n, df = p)
    # plot, mark points with heighest distance
    plot(qchi, sortd, pch = 19, xlab = "Chi-square quantiles",
        ylab = "Mahalanobis squared distances",
        main = "Chi-square Q-Q Plot")
    points(qchi[(n - mark + 1):n], sortd[(n -
        mark + 1):n], cex = 3, col = "#990000")
}
```

Figure 15 shows a chi-square plot for the dataset.

```
# Call the function; mark two top points
chisquare.plot(x = dat[, 1:4], mark = 2)
```

It seems, for the most part, the chi-square plot indeed shows a linear pattern. However, there are one or two points (upper right corner; marked by red circles) show deviation from the linear trend. These points may indicate that there are outliers.

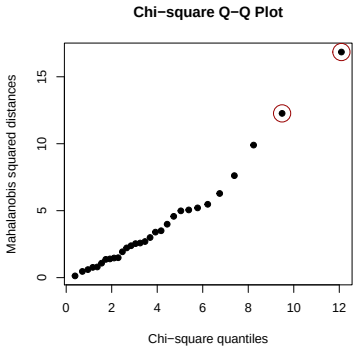


Figure 15: Chi-square plot for the stiffness dataset.

Outlier detection

Outliers can be viewed as unusual data points that do not seem to follow the pattern of variability produced by other observations. Univariate outliers can be detected using a dot plot or boxplot. However, it might be more complicated for multivariate data. The chi-square plot describes above can also be used for outlier detection.

In case that there are suspected outliers, we should inspect the data points corresponding to the top few distance values. We would like to see in what manner the outliers differ from the rest of the dataset. Thus, along with the actual data points, it is also useful to inspect the z-scores for each variable. Recall that, if the assumption of multivariate normality is reasonable, then z-scores of each variable should follow a standard normal distribution. We expect roughly[7] 99% of the z-score values should fall in the interval $[-2.57, 2.57]$. Thus any z-scores outside the interval above can be considered unusual.

[7] We know that

$$P[-2.57 \leq Z \leq 2.57] \approx 0.99.$$

The following table shows data rows with largest Mahalanobis distance values, along with z-scores of each variable.

x1	x2	x3	x4	z1	z2	z3	z4	d2	ID
1954	2149	1180	1281	0.15	1.25	-1.09	-1.38	**16.85**	16
2983	2794	2412	2581	**3.31**	**3.28**	**2.98**	**2.65**	**12.26**	9
2276	2189	1547	2111	1.14	1.38	0.12	1.2	9.9	21
2119	1700	1815	2222	0.66	-0.16	1.01	1.54	7.62	3
2326	2301	2065	2234	1.29	1.73	1.83	1.58	6.28	29
2403	2048	2087	2197	1.53	0.94	1.91	1.46	5.48	2

Table 1: Data rows with largest Mahalanobis distance values.

It seems that observations 16 and 9 are outliers for different reasons. Observation 16 has the highest Mahalanobis distance; however, the z-scores of the individual variables are well within the usual range of $[-2.57, 2.57]$. This type of outliers are quite difficult to detect visually. For example, a scatterplot of X_1 and X_2 in Figure 16. However, observation 16 is hidden within the data cloud and is only visible in the chi-square plot.

In contrast, observation 9 is easy to notice since it is visible in scatterplots as well. This is because even though the observation

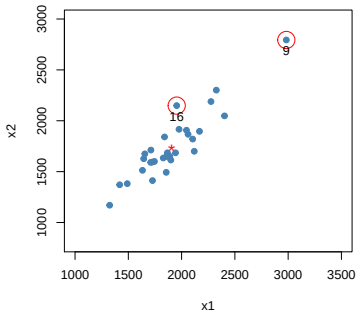


Figure 16: Scatterplot of X1 and X2 with marked outliers (red circles).

follows the overall pattern on the plot (there seems to be a linear relationship between X_1 and X_2, and observation 9 does conform to the relationship), the z-scores are very large in magnitude for all the four measures of stiffness.

Once we find an outlier, we must try to access the real specimens and re-examine them whenever possible to determine the reason behind the unusual observations.

Bivariate boxplot

We discuss two extensions of the univariate boxplot to the bivariate situation. A bivariare analogue of the usual boxplot is proposed by Goldberg and Iglewicz (1992).[8] The `bvbox()` function in the `MVA` package implements this method.

Let us look at the variables x_1 and x_2 from the lumber stiffnes data discussed before. A bivariate boxplot is shown in Figure 17.

[8] Goldberg and Iglewicz (1992). Bivariate Extensions of the Boxplot, Technometrics, 34:3, 307-320.

```
library(MVA)
bvbox(dat[, 1:2], pch = 19, col = "#990000", xlab = "x1",
    ylab = "x2", main = "Bivariate boxplot")
text(dat[c(9, 16), 1], dat[c(9, 16), 2], pos = 1,
    labels = c(9, 16))
```

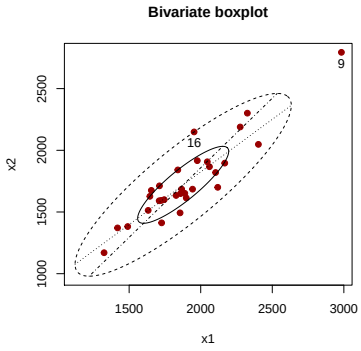


Figure 17: Bivariate boxplot of X1 and X2.

The bivariate boxplot consists of the following:

- Two concentric ellipses, the inner ellipse (called the "hinge") contains 50% of the data, and the outer ellipse (called the "fence") determines potential outliers. These ellipses are drawn based on robust measures of location, scale, and correlation, and a constant, D, that determines the distance of the fence from the hinge. Goldberg and Iglewicz (1992) propose to use $D = 7$ so that the outer ellipse forms an approximate 99% confidence bound.
- Resistant (robust) regression lines of both y on x and x on y are drawn. Their intersection shows the location estimator.

It seems observation 9 is an outlier. However, observation 16 is on the fence.

Bagplot

Another bivariate extension of the usual boxplot, called bagplot, has been suggested by Rousseeuw, Ruts and Tukey (1999).[9] The `bagplot()` function in the `aplpack` package implements this method. Figure 18 shows a bagplot of X_1 and X_2.

[9] Rousseeuw, Ruts and Tukey (1999). The Bagplot: A Bivariate Boxplot, The American Statistician, 53:4, 382-387.

```
# Example of a Bagplot
library(aplpack)
bagplot(dat[, 1], dat[, 2], xlab = "x1", ylab = "x2",
    main = "Bagplot", pch = 19, cex = 1)
text(dat[c(9, 16), 1], dat[c(9, 16), 2], pos = 1,
    labels = c(9, 16))
```

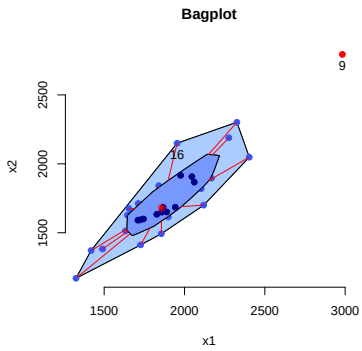


Figure 18: Bagplot of X1 and X2.

The bagplot is based on the concept of *halfspace location depth* of a point relative to a bivariate dataset, which extends the univariate concept of rank. The plot consists of the following:

- An inner convex polygon, called the "bag," containing 50% of the data points (with the largest depth).
- The outer polygon, called the "fence" is created by magnifying the bag by a factor of three. The fence separates inliers from outliers. The fence is not plotted, but the outliers are plotted in red. The observations between the bag and the fence are shown using a lighter color.

The bagplot visualizes the location, spread, correlation, skewness, and tails of the data. It is not limited to elliptical (e.g., multivariate normal) distributions.

Steps to detect outliers

Johnson and Wichern (2007)[10] suggests the following steps for detecting outliers:

[10] Applied multivariate statistical analysis by Richard A. Johnson, Dean W. Wichern. Prentice-Hall.

- Construct dotplot/boxplot/qqplot of each variable
- Make scatterplots for each pair of variables
- Calculate standardized values for each variable

$$z_k = \frac{x_k - \bar{x}_k}{\sqrt{s_k^2}}.$$

  Examine the standardized values for extreme points. This depends of the sample size as well as number of variables. Even if the data came from a normal distribution, we can expect 1% absolute values of the z-scores to exceed 2.57.

- Calculate the Mahalanobis squared distances $(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{s}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ and create chi-square plot. Examine the points with unusually large distance values.

We reiterate that that once we find an outlier, we must try to access the real specimens and re-examine them whenever possible to determine the reason behind the unusual observations.