Principal Components Analysis

Today . . .

- Identification of outliers
- Verifying multivariate normality

Principal components analysis can be used to screen the data for outliers, especially cases that may not be univariate outliers but are unusual in the multivariate sense.

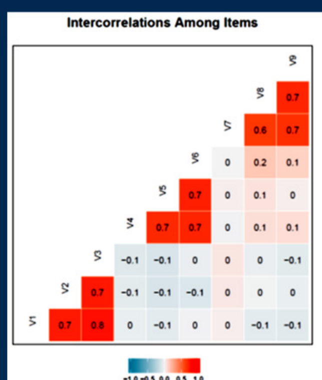
To provide a comparison, we will first examine data that does not contain outliers. The data (N = 250) are generated from a multivariate normal distribution with 9 variables. Later we will replace the last case with a multivariate outlier.

```r
means <- matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0))
sigma <- matrix(c(1, 0.7, 0.7, 0, 0, 0, 0, 0, 0, 0.7, 1, 0.7, 0, 0,
    0, 0, 0, 0, 0.7, 0.7, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0.7, 0.7,
    0, 0, 0, 0, 0, 0, 0.7, 1, 0.7, 0, 0, 0, 0, 0, 0, 0.7, 0.7, 1,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0.7, 0.7, 0, 0, 0, 0, 0, 0, 0.7,
    1, 0.7, 0, 0, 0, 0, 0, 0, 0.7, 0.7, 1), nrow = 9, ncol = 9)
Data <- mvrnorm(250, means, sigma)
```

The generated data will fit this pattern, especially as sample size increases.

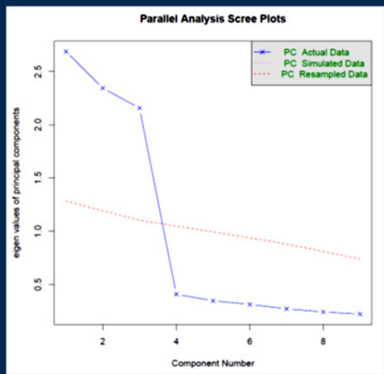
$$R = \begin{bmatrix} 1.0 & 0.7 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.7 & 1.0 & 0.7 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.7 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.7 & 0.7 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.7 & 1.0 & 0.7 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.7 & 0.7 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.7 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7 & 1.0 & 0.7 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7 & 0.7 & 1.0 \end{bmatrix}$$

The intended correlations exist in the sample and the pattern of correlations suggests that three independent linear combinations likely will account for the 9 variables.




```r
R <- cor(Data)
KMO(R)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R)
## Overall MSA =  0.72
## MSA for each item =
##    V1   V2   V3   V4   V5   V6   V7   V8   V9
## 0.74 0.76 0.73 0.75 0.71 0.69 0.71 0.72 0.69

cortest.bartlett(R = R, n = length(Data[, 1]))

## $chisq
## [1] 1154
##
## $p.value
## [1] 6.233e-219
##
## $df
## [1] 36
```

Both the KMO and Bartlett tests indicate that the correlation matrix departs from an identity matrix.



The scree test indicates that three components should be extracted from the data. Beyond three components, the eigenvalues suggest randomness.

```
PCA_1 <- principal(Data, nfactors = 3, rotate = "none", residuals = TRUE)
PCA_1
```

```
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1  PC2   PC3   h2   u2 com
V1 -0.62 0.58  0.32 0.83 0.17 2.5
V2 -0.63 0.59  0.24 0.81 0.19 2.3
V3 -0.64 0.59  0.27 0.84 0.16 2.4
V4  0.59 0.24  0.61 0.78 0.22 2.3
V5  0.61 0.21  0.62 0.79 0.21 2.2
V6  0.60 0.30  0.59 0.81 0.19 2.5
V7  0.28 0.62 -0.54 0.75 0.25 2.4
V8  0.41 0.61 -0.49 0.78 0.22 2.7
V9  0.41 0.59 -0.53 0.80 0.20 2.8

                       PC1  PC2  PC3
SS loadings           2.69 2.34 2.16
Proportion Var        0.30 0.26 0.24
Cumulative Var        0.30 0.56 0.80
Proportion Explained  0.37 0.33 0.30
Cumulative Proportion 0.37 0.70 1.00
```

The components account for 80% of the original score variance.

Note that the principal component loadings do not suggest simple interpretations; that will require an additional transformation.

```
# Create a correlation matrix of the residuals by replacing the
# main diagonal with ones.
R1 <- diag(PCA_1$residual)
R2 <- diag(R1)
R3 <- PCA_1$residual - R2
R4 <- diag(9) + R3

# Assess the factorability of the residual correlation matrix.
KMO(R4)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = R4)
## Overall MSA =  0.46
## MSA for each item =
##   V1   V2   V3   V4   V5   V6   V7   V8   V9
## 0.46 0.47 0.46 0.46 0.47 0.46 0.47 0.45 0.44

cortest.bartlett(R = R4, n = length(Data[, 1]))
## $chisq
## [1] 26.29
##
## $p.value
## [1] 0.8822
##
## $df
## [1] 36
```

Once the three components are removed, there is no additional meaningful variability.

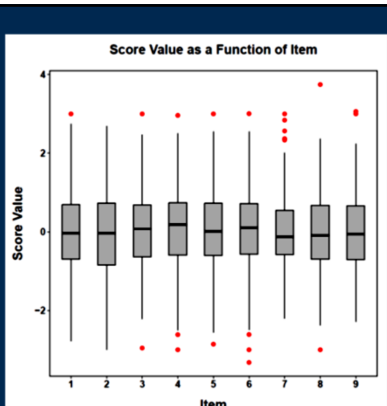
The data are in standard score form to make interpretation easy. The last case was replaced with a profile that made it unusual in the multivariate sense, though not terribly deviant in the univariate sense:

Variable 1:	3
Variable 2:	-3
Variable 3:	3
Variable 4:	-3
Variable 5:	3
Variable 6:	-3
Variable 7:	3
Variable 8:	-3
Variable 9:	3

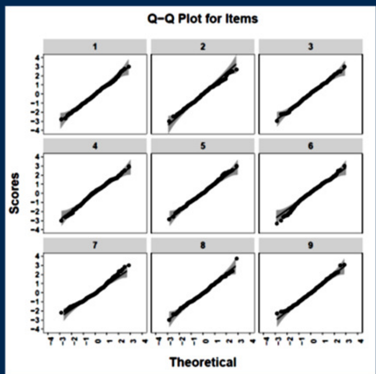
In a sample this large, such values would be expected, but probably not for the same case and certainly not in this pattern.

```
describe(Data)

##    vars   n  mean   sd median trimmed  mad   min  max range
## V1    1 250 -0.05 1.01  -0.04   -0.05 1.04 -2.78 3.00  5.78
## V2    2 250 -0.06 1.01  -0.04   -0.05 1.14 -3.00 2.68  5.68
## V3    3 250  0.02 1.01   0.08    0.03 0.95 -2.96 3.00  5.96
## V4    4 250  0.09 1.03   0.19    0.12 0.96 -3.00 2.96  5.96
## V5    5 250  0.07 1.01   0.01    0.05 1.00 -2.86 3.00  5.86
## V6    6 250  0.03 1.03   0.10    0.08 0.94 -3.32 3.01  6.33
## V7    7 250  0.03 0.94  -0.12   -0.03 0.91 -2.20 3.00  5.20
## V8    8 250 -0.04 0.99  -0.09   -0.06 1.03 -3.00 3.75  6.75
## V9    9 250 -0.01 1.00  -0.06   -0.03 1.00 -2.28 3.06  5.35
##     skew kurtosis   se
## V1  0.06    -0.02 0.06
## V2 -0.07    -0.25 0.06
## V3 -0.05    -0.05 0.06
## V4 -0.31    -0.01 0.07
## V5  0.16     0.01 0.06
## V6 -0.39     0.43 0.07
## V7  0.53     0.13 0.06
## V8  0.24     0.37 0.06
## V9  0.23    -0.16 0.06
```

The univariate outliers do not distort the descriptive statistics in any obvious way. The case with the odd profile does not have the most extreme scores for some of the variables.



Visual displays do not reveal any particular problems.



The data look embarrassingly normal.

```
PCA_2 <- principal(Data, nfactors = 9, rotate = "none", residuals = TRUE,
    scores = TRUE)
PCA_2

## Principal Components Analysis
## Call: principal(r = Data, nfactors = 9, residuals = TRUE, rotate = "none",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1  PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9 h2
## V1 -0.64 0.53  0.36  0.16 -0.14 -0.06 -0.01  0.21 -0.29  1
## V2 -0.62 0.50  0.32 -0.33  0.25  0.07  0.22 -0.18 -0.03  1
## V3 -0.67 0.54  0.31  0.13 -0.12 -0.02 -0.17 -0.02  0.33  1
## V4  0.60 0.21  0.62 -0.16  0.29 -0.26 -0.12  0.17  0.04  1
## V5  0.58 0.22  0.58  0.39  0.02  0.24  0.25  0.06  0.08  1
## V6  0.61 0.27  0.60 -0.15 -0.27  0.02 -0.14 -0.26 -0.11  1
## V7  0.21 0.68 -0.51  0.31  0.26  0.11 -0.22 -0.12 -0.09  1
## V8  0.37 0.63 -0.42 -0.41 -0.12  0.25 -0.01  0.21  0.05  1
## V9  0.34 0.65 -0.50  0.08 -0.13 -0.35  0.23 -0.06  0.04  1
```

Here we extract all 9 principal components to see if the multivariate outlier reveals itself in these linear combinations. Our goal is not data reduction.

A principal components analysis will seek linear combinations that capture the major sources of variance in the data. Most of these will be governed by the "well-behaved" data. But, once that variation is captured, especially deviant multivariate cases may dominant the smaller components and emerge more readily.

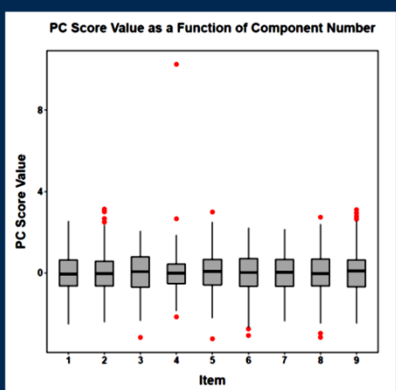
For outlier detection, all components are derived and component scores are produced. Then diagnostics are performed on the component scores.

```
Data_PC <- as.data.frame(PCA_2$scores)
```

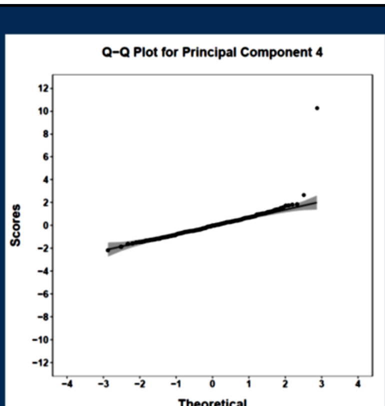
```
describe(Data_PC)

##     vars   n mean sd median trimmed  mad   min   max range  skew
## PC1    1 250    0  1  -0.06   -0.02 0.96 -2.54  2.54  5.08  0.17
## PC2    2 250    0  1  -0.05   -0.02 0.91 -2.42  3.13  5.55  0.21
## PC3    3 250    0  1   0.06    0.02 1.11 -3.18  2.05  5.23 -0.25
## PC4    4 250    0  1  -0.01   -0.05 0.73 -2.17 10.26 12.43  4.34
## PC5    5 250    0  1   0.07    0.00 0.95 -3.25  2.99  6.24 -0.03
## PC6    6 250    0  1   0.01    0.03 1.01 -3.09  2.21  5.30 -0.27
## PC7    7 250    0  1   0.02    0.02 0.96 -2.38  2.15  4.53 -0.19
## PC8    8 250    0  1  -0.04    0.02 0.96 -3.18  2.73  5.92 -0.17
## PC9    9 250    0  1   0.09   -0.01 0.92 -2.50  3.10  5.61  0.21
##     kurtosis   se
## PC1    -0.40 0.06
## PC2     0.30 0.06
## PC3    -0.42 0.06
## PC4    42.39 0.06
## PC5     0.10 0.06
## PC6    -0.26 0.06
## PC7    -0.44 0.06
## PC8     0.15 0.06
## PC9     0.34 0.06
```

Problems are now readily apparent.

Approximate standard errors for skew and kurtosis are $(6/N)^{1/2}$ and $(24/N)^{1/2}$, respectively (.15 and .31).



A visual display of the component scores shows a clear problem with Component 4.



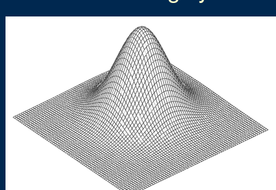
A visual display of the component scores shows a clear problem with Component 4.

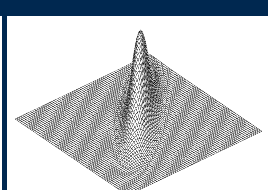
Once identified, a multivariate outlier must be dealt with in some way.

- Transformation (probably will not help)
- Elimination (requires strong justification)
- Sensitivity analysis

In univariate statistics, the normality assumption underlies significance testing. It is with reference to sampling from some theoretical distribution that we can make claims about the likelihood of results occurring "by chance" or "under the null hypothesis." Similarly, the establishment of confidence intervals depends on distributional assumptions.

Underlying many multivariate procedures is the assumption of *multivariate normality*. This assumption extends the idea of bivariate normality to more than two dimensions. In bivariate normality, the distribution of one variable is normal for all values of the other variable, even when the variables are highly correlated.





When multivariate normality holds:

- All marginal distributions will be normal.
- All pairs of variables will be bivariate normal.
- All linear combinations will be normal.
- All pairs of linear combinations will be bivariate normal.
- Squared distances from the population centroid will be chi-square distributed with *k* (*k* =number of variables) degrees of freedom.

Violating any of these is a violation of multivariate normality.

All linear combinations will be normal

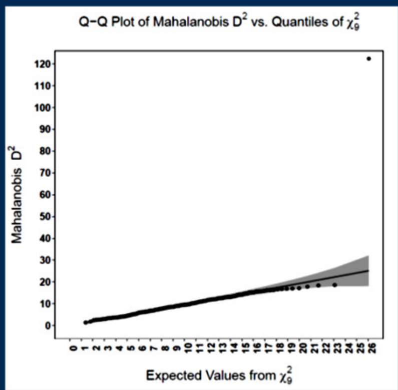
It is not practical to test all linear combinations—there are an infinite number of them. But, testing a small number of commonly used linear combinations is important. The most commonly tested:

- Sum of all measures
- Pair-wise differences
- ***Principal components***

The distance of each case from the multivariate mean is indicated by the Mahalanobis distance:

$$X = [X_1, X_2, X_3, \ldots]$$
$$\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \ldots]$$
$$D_X = \sqrt{[X - \boldsymbol{\mu}]' \sum^{-1} [X - \boldsymbol{\mu}]}$$

Mahalanobis distance squared is χ^2 distributed with degrees of freedom equal to the number of measures (here df = 9).



Mahalanobis distance provides a clear indication that something is amiss.

The general formula for Mahalanobis distance:

$$X = [X_1, X_2, X_3, \ldots]$$
$$\mu = [\mu_1, \mu_2, \mu_3, \ldots]$$
$$D_X = \sqrt{[X - \mu]' \sum^{-1} [X - \mu]}$$

How is this simplified when applied to principal component scores?

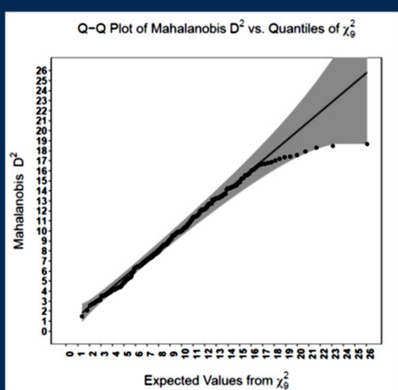
```
mvn(Data, mvnTest = "mardia", multivariatePlot = "qq", multivariateOutlierMethod = "quan",
    showOutliers = TRUE)
```

```
$multivariateNormality
             Test        Statistic               p value Result
1 Mardia Skewness  1380.9251635341 7.25821195066503e-191     NO
2 Mardia Kurtosis 27.0433388155401                     0     NO
3             MVN             <NA>                  <NA>     NO
```

Several direct tests of multivariate normality are available (in the MVN package), but they need not agree and are overly sensitive with large N.

The original data, without the outlier, passes the multivariate normality test. As with other significance tests, minor violations will be detected as significant as sample size increases.

```
$multivariateNormality
             Test          Statistic           p value Result
1 Mardia Skewness  146.823933463264 0.842021240357815    YES
2 Mardia Kurtosis -1.41439294192397 0.157246561287253    YES
3             MVN              <NA>              <NA>    YES
```



The original data, without the outlier, passes the multivariate normality test.

Violations of multivariate normality can be handled in multiple ways:

- Transformations
- Robust methods
- Resampling

Next time . . .

Simplified composites, group contamination, reducing multicollinearity, and some PCA-related methods.