Robust estimators of covariance for examination of inter-laboratory study data

Stephen L R Ellison
LGC Limited, UK

Introduction

- **Interlaboratory study data**
- **Covariance – a reminder**
- **Two robust covariance estimators**
- **Some applications in interlaboratory data review**
  - > Improved Youden plots
  - > Sharper outlier detection using a multivariate distance measure
  - > Robust principal component analysis

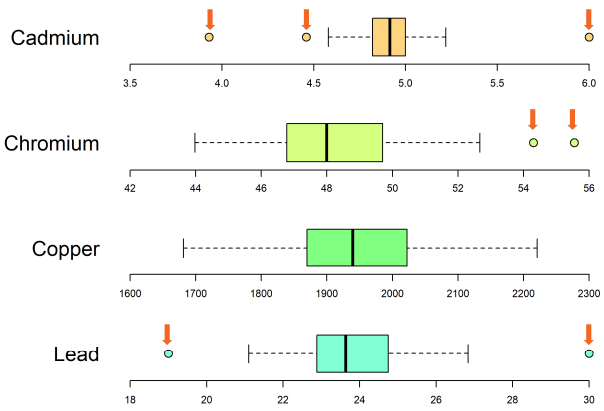
Environmental RM certification study: Metals in drinking water

- 23 elements
- 25 laboratories
- Natural candidate RM (5 reps)
- Synthetic QC material (3 reps)
- ~4600 data points
- Up to 1150 lab means/medians

Lab	Aluminium	Antimony	Arsenic	Barium	Beryllium	B
1	194.0600	5.470000	9.63000	113.0300	5.190000	1055.
1	211.1100	5.670000	9.86000	113.6000	4.920000	1021.
1	205.6900	5.380000	9.45000	113.1400	5.340000	1035.
1	197.0600	5.530000	9.37000	112.7300	5.430000	1022.
1	201.8300	5.500000	9.83000	112.3800	5.360000	1028.
2	200.0000	4.708000	9.77000	118.0000	NA	1010.
2	201.0000	4.920000	9.91000	117.0000	NA	1010.
2	199.0000	5.002000	9.71000	116.0000	NA	1030.
2	200.0000	4.916000	9.80000	116.0000	NA	1020.
2	205.0000	4.980000	9.75000	115.0000	NA	1010.
3	199.0000	5.450000	10.50000	117.0000	5.170000	964.
3	199.0000	5.450000	10.60000	118.0000	5.050000	983.
3	200.0000	5.390000	10.40000	117.0000	4.930000	999.
3	200.0000	5.410000	10.40000	118.0000	5.010000	1005.

Lab	Aluminium	Antimony	Arsenic	Barium	Beryllium	Boron	Cadmium	Chromium	Cobalt	Copper	Iron	Lead
1	226.8000	3.800000	9.90000	134.1800	5.630000	826.2300	6.220000	40.66000	NA	1755.480	202.4600	7.980000
1	226.1600	3.860000	10.51000	133.5800	6.140000	807.0200	6.360000	40.39000	NA	1783.190	204.4700	8.030000
1	228.2300	3.820000	10.91000	133.8100	5.930000	797.8100	6.170000	40.50000	NA	1777.830	202.8600	7.870000
2	230.0000	3.669000	10.90000	138.0000	NA	793.0000	5.810000	40.70000	4.050000	1790.000	215.0000	8.280000
2	233.0000	3.684000	11.00000	142.0000	NA	786.0000	5.840000	41.70000	4.110000	1810.000	218.0000	8.430000
2	233.0000	3.708000	11.00000	142.0000	NA	814.0000	5.850000	42.70000	4.140000	1790.000	219.0000	8.350000
3	226.0000	3.830000	11.30000	139.0000	6.140000	741.9800	6.030000	40.90000	3.840000	1800.000	211.0000	8.230000
3	221.0000	3.750000	11.20000	141.0000	5.800000	732.3800	6.030000	41.00000	3.880000	1790.000	206.0000	8.250000
3	223.0000	3.760000	11.40000	140.0000	5.910000	756.8500	6.000000	40.60000	3.860000	1790.000	209.0000	8.280000
4	251.9700	4.310000	12.77000	150.4900	NA	772.6200	6.540000	45.16000	NA	1631.750	209.5000	8.670000

Univariate outliers - one variable at a time



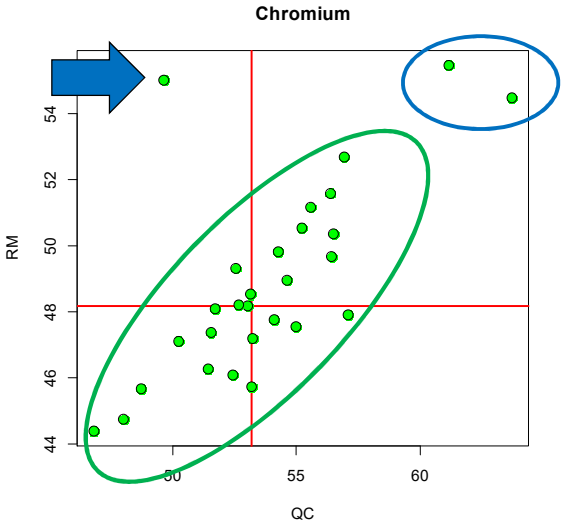
Lab	Aluminium	Antimony	Arsenic	Barium	Beryllium	B
1	194.0600	5.470000	9.63000	113.0300	5.190000	1055.
1	211.1100	5.670000	9.86000	113.6000	4.920000	1021.
1	205.6900	5.380000	9.45000	113.1400	5.340000	1035.
1	197.0600	5.530000	9.37000	112.7300	5.430000	1022.
1	201.8300	5.500000	9.83000	112.3800	5.360000	1028.
2	200.0000	4.708000	9.77000	118.0000	NA	1010.
2	201.0000	4.920000	9.91000	117.0000	NA	1010.
2	199.0000	5.002000	9.71000	116.0000	NA	1030.
2	200.0000	4.916000	9.80000	116.0000	NA	1020.
2	205.0000	4.980000	9.75000	115.0000	NA	1010.
3	199.0000	5.450000	10.50000	117.0000	5.170000	964.
3	199.0000	5.450000	10.60000	118.0000	5.050000	983.
3	200.0000	5.390000	10.40000	117.0000	4.930000	999.
3	200.0000	5.410000	10.40000	118.0000	5.010000	1005.
3	199.0000	5.440000	10.50000	118.0000	5.080000	976.
4	224.1800	6.010000	11.44000	126.2400	NA	954.
4	216.2600	6.130000	11.60000	124.9400	NA	978.
4	211.8700	6.200000	12.05000	125.1500	NA	992.
4	216.8000	5.970000	11.79000	124.9300	NA	991.
4	218.6400	6.160000	11.70000	125.6800	NA	978.
5	180.0000	5.600000	11.00000	130.0000	5.400000	1100.
5	180.0000	5.700000	10.00000	120.0000	5.800000	1000.
5	180.0000	5.900000	10.00000	130.0000	5.600000	1100.
5	190.0000	5.900000	10.00000	130.0000	5.700000	1100.
5	180.0000	5.700000	10.00000	120.0000	5.700000	1100.

A different anomaly

Youden plot

- One measurand against another
- Commonly used to identify significant laboratory effects

- Here: Candidate RM plotted against QC material



Univariate outliers

5

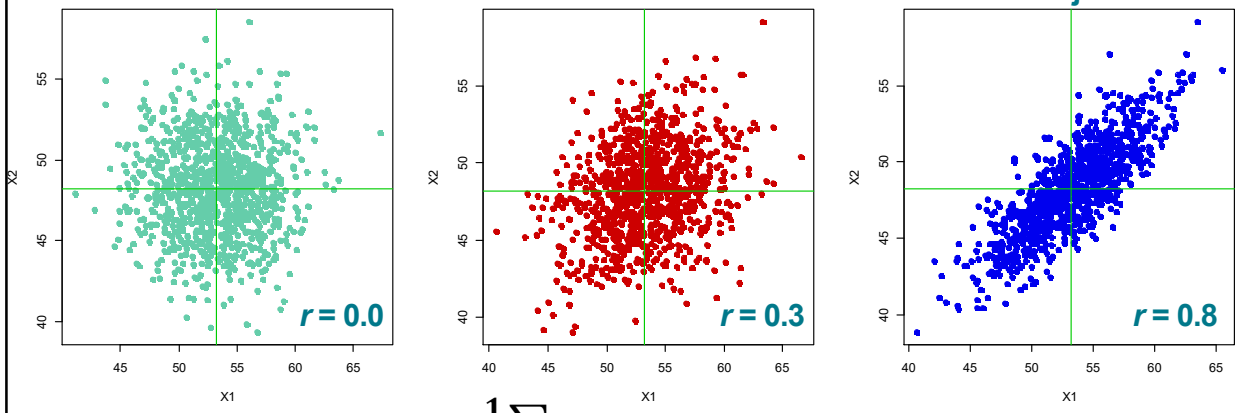
Describing multivariate data -

Why we need covariance

6

Covariance

These three data sets have the same standard deviations on each major axis



$$cov(X,Y) = \frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y}) = r\, s(x)s(y)$$

7

Univariate, bivariate and multivariate data

Univariate data

- One location
- One standard deviation or variance

Bivariate data

- Two locations
- Two standard deviations/variances
- One covariance
- A (small) *covariance matrix*:

$$\begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$$

Multivariate (n) data

- n locations
- n variances
- $n(n-1)/2$ covariances
- An $n \times n$ covariance matrix

$$\begin{bmatrix} var(x_1) & \dots & cov(x_1,x_n) \\ \vdots & \ddots & \vdots \\ cov(x_1,x_n) & \dots & var(x_n) \end{bmatrix}$$

8

Calculating robust covariance

9

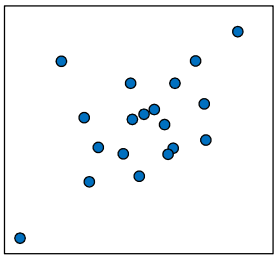
Two approaches to robust covariance

- **Gnanadesikan and Kettenring (GK)** (1972)
  - > **Fact:** $\mathbf{cov}(\mathbf{x},\mathbf{y}) = \left[s(x+y)^2 - s(x-y)^2\right]/4$
  - > **GK proposal:** $cov^*(\mathbf{x},y) = \left[s^*(x+y)^2 - s^*(x-y)^2\right]/4$
  - > **Where s^* is any robust estimator for standard deviation**

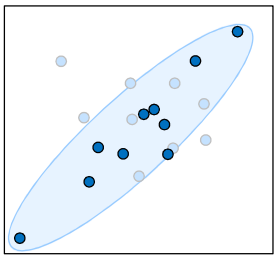
- **Minimum covariance determinant (MCD)**
  (Rousseeuw *et al* 1992; Maronna and Zamar, 2002)
  - > Takes the covariance matrix for the subset of data with the minimum covariance determinant, corrected for subset selection.

10

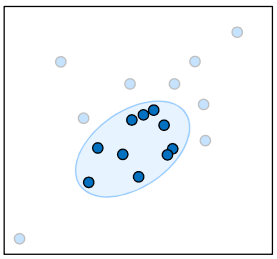
MCD - principle



- Start with a multivariate data set



- Choose $m\ (\geq n/2)$
- Calculate complete covariance matrix



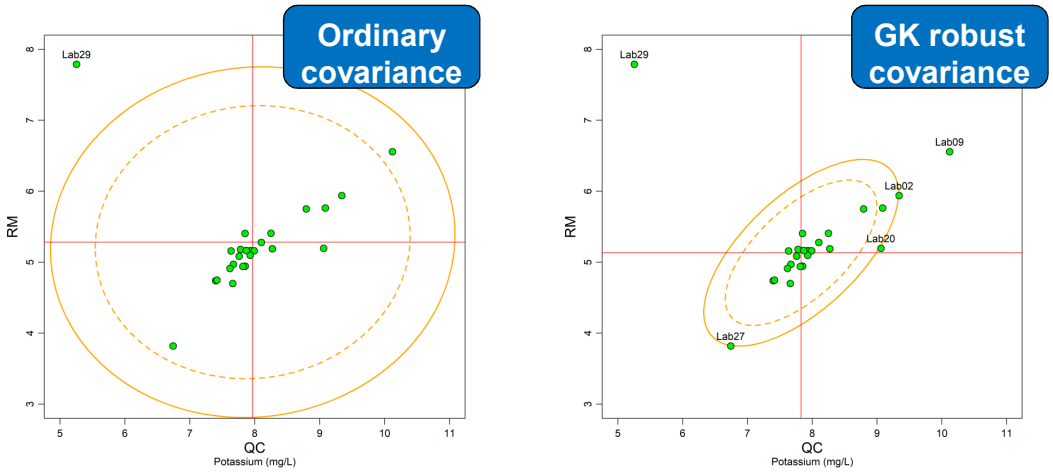
- Find set with smallest covariance determinant
- Correct variances for subset selection

11

Some applications of robust covariance

12

Applications
1. Robust confidence region for Youden plots

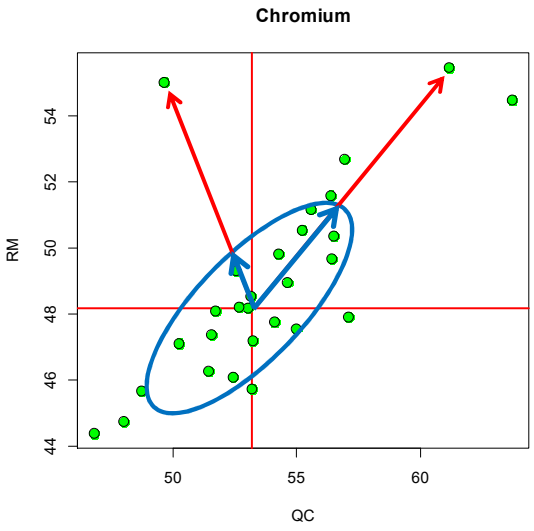


13

Applications
2. Robust Mahalanobis distance

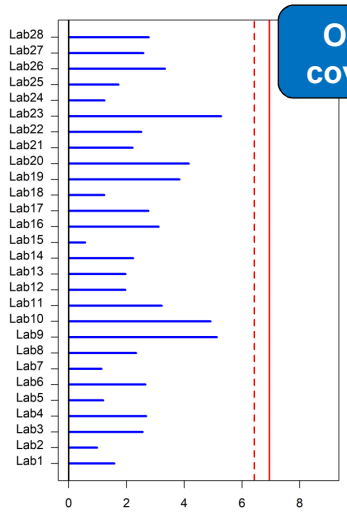
- **What's a Mahalanobis distance?**
  - > A scaled distance from the 'centre' of a data set
  - > Calculated using the (inverse of the) covariance matrix
  - > For multivariate Normal data with n variables*, M^2 is distributed as $\chi^2(n)$

*For known location and covariance

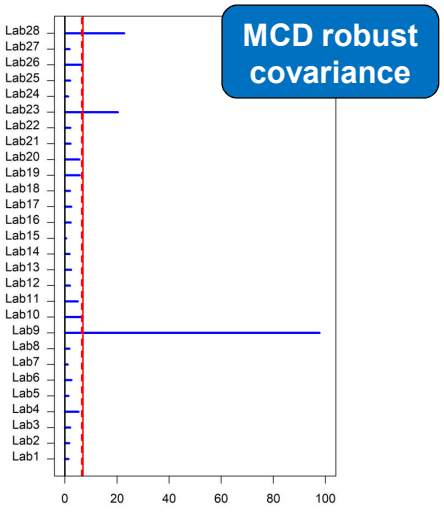


14

Applications
2. Robust Mahalanobis distance



- MHD all look OK under normal covariance
- Robust covariance shows effect of outlying values.



15

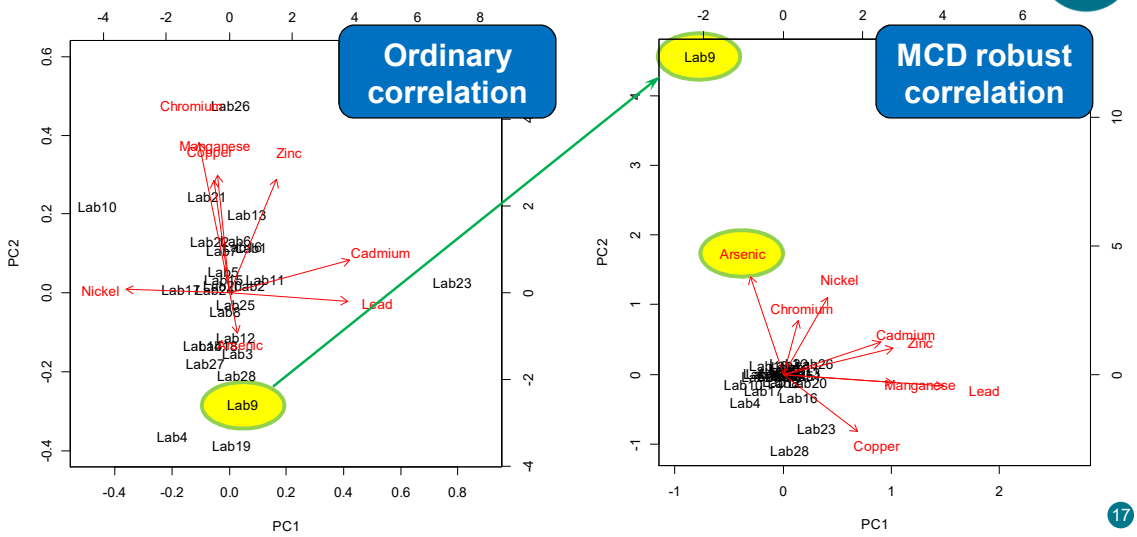
Applications
3. Robust Principal Component Analysis

- **PCA is a useful tool for inspecting multivariate data**
  - > 'Dimensionality reduction' allows simplified plotting
  - > Identifies appreciable differences in 'profile'

- **PCA is related to correlation and covariance**
  - > PC's are eigenvectors of a covariance or correlation matrix
  - > Use of a robust correlation or covariance matrix gives robust PCA

16

Applications
3. Robust Principal Component Analysis



17

Conclusions

- **Several robust covariance estimators are now available**
- **Robust covariance underpins useful tools for outlier detection in multivariate data**
  - > Outlier-resistant confidence regions
  - > Robust measures of multivariate distance
  - > Robust PCA

Credits: Mandel and Youden plots produced using the metRology package for R
Robust PCA produced using the rrcov package

18