Data standardisation

Stephen Coleman

February 4, 2020

1 Standardisation

Often we are interested in clustering by covariance (consider defining gene sets by co-expression). We are interested in the common variation across experimental conditions rather than in the magnitude of expression; in this case we *standardise* the data.

Consider a dataset of n observations, $X = (x_1, \ldots, x_n)$. Within this, consider a sample (for example a row encoding the expression of a gene or some measurement such as height), $X_i = (x_{i1}, \ldots, x_{ip})$. Standardisation of X_i is the transform from X_i to $X_i' = (x_{i1}', \ldots, x_{ip}')$ defined by the sample mean, $\bar{x}_i$, and sample standard deviation, s_i. Then:

$$\bar{x}_i = \frac{1}{p}\sum_{j=1}^{p} x_{ij} \tag{1}$$

$$s_i^2 = \frac{1}{p-1}\sum_{j=1}^{p} (x_{ij} - \bar{x}_i)^2 \tag{2}$$

$$x_{ij}' = \frac{x_{ij} - \bar{x}_i}{s_i} \; \forall \; j \in (1, \ldots, p) \tag{3}$$

I refer to X_i' as the standardised form of X. If one is given a dataset $X = (X_1, \ldots, X_n)$ where each X_i is a p-vector of observations of the form referred to above, then in referring to the standardised form

of X, I mean the dataset $X' = (X'_1, \ldots, X'_n)$ where each X'_i is the standardised form of X_i.

Standardisation moves the values observed for each X_i to a common scale where each vector has an observed mean and standard deviation of 0 and 1 respectively. A motivating example in the context of gene expression data is described in section 1.1.

1.1 Example: Standardising gene expression data

If one considers table 1 which contains an example of expression data for some genes A, B, C, D and E across people 1 to 4. One can

Genes	Person 1	Person 2	Person 3	Person 4
A	5.1	5.2	4.9	5.0
B	5.1	4.9	5.2	5.4
C	1.4	1.5	1.2	1.3
D	1.4	1.2	1.5	1.7
E	1.4	1.5	1.4	1.5

Table 1: Example gene expression data.

see that genes A and C have similar patters in variation across the people, as do genes B and D. Gene E is not consistent with any other gene here. However, as this relative variation is of interest rather than the magnitude of expression, one can see that standardising the data is required.

If one were to cluster the data as represented in table 1, one would place genes A and B in one cluster and genes C, D and E in another as their absolute expression levels are similar (as can be seen in figure 1). However, if the expression level of each gene is standardised as per section 1, the data is then as represented in table 2. The data are now on the same scale and thus the characteristic that will determine a clustering is the variation of expression across people. As we want genes with similar patterns of variation (i.e. that are co-expressed) this enables us to cluster under our objective of defining gene sets. In

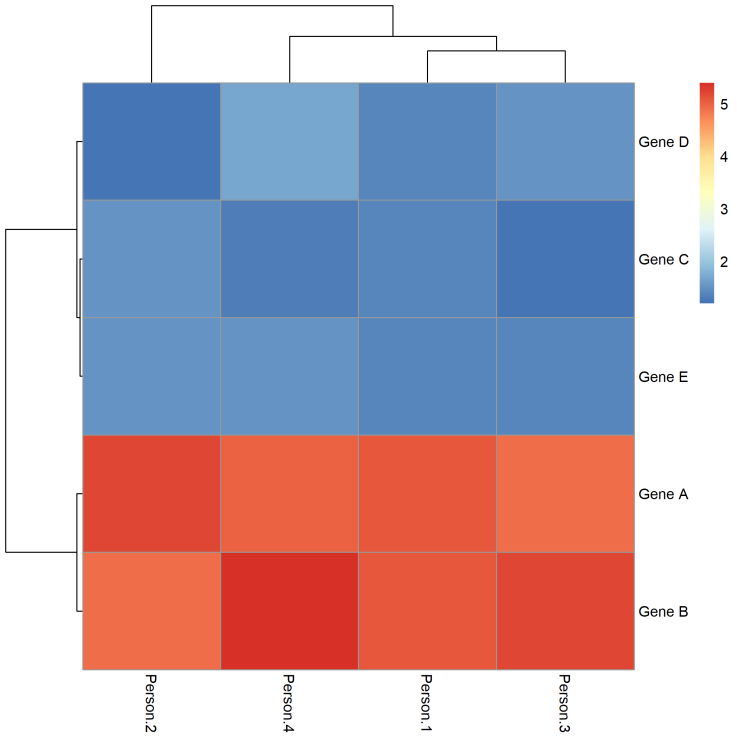

Figure 1: Heatmap of expression data in table 1 showing the clusters based upon magnitude of expression.

this case genes A and C are one cluster, genes B and D another with gene E in a cluster alone, as can be seen in figure 2. As this is the type of data we wish to cluster across, we therefore most standardise our expression data before clustering can be implemented.

Genes	Person 1	Person 2	Person 3	Person 4
A	0.39	1.16	-1.16	-0.39
B	-0.24	-1.20	0.24	1.20
C	0.39	1.16	-1.16	-0.39
D	-0.24	-1.20	0.24	1.20
E	-0.87	0.87	-0.87	0.87

Table 2: Example standardised gene expression data.

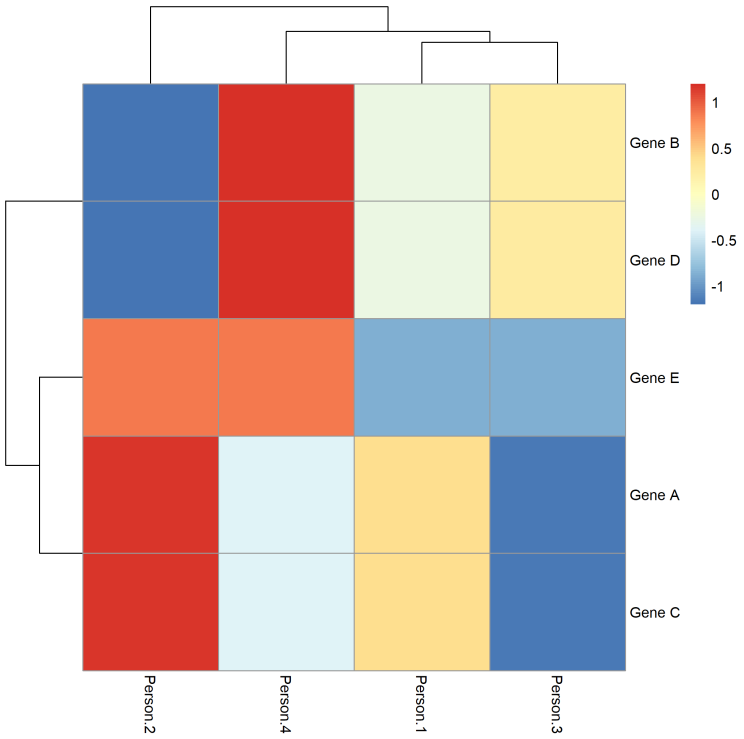

Figure 2: Heatmap of expression data in table 2 showing the clusters based upon variation of expression across people.