PCA eigen-vector plot

Though the vector overlay has a tendency to clutter the plot, the changing contaminant load along this E-W transect of sampling sites (Fig. 1.5 in CiMC) is clear. The end points S1 and S12 lie close together and there is a strong trend from S1 to the dump centre at S6 (left to right on axis PC1), and a reversal of that trend for S6 to S12, moving away from the dump centre. The trajectory differs on the PC2 axis, however, for the two arms of the transect. The results window (heading *Eigenvalues*) shows that a 2-d PCA is a very good description of structure in the higher (11-d) space, the first axis (PC1) accounting for much of the variability (62%) and PC2 most of the remainder (a further 27%), i.e. 89% between them. The *Eigenvectors* are the linear combinations which define the axes:

$$ \text{PC1} = 0.378(\ln \mathit{Cu})^* – 0.213(\ln \mathit{Mn})^* – 0.075(\ln \mathit{Co})^* + 0.149(\ln \mathit{Ni})^* + …; $$

$$ \text{PC2} = –0.035(\ln \mathit{Cu})^* + 0.418(\ln \mathit{Mn})^* + 0.539(\ln \mathit{Co})^* + 0.466(\ln \mathit{Ni})^* + …, $$

the asterisks being a reminder that the transformed variables are normalised. It is the coefficients in these equations (eigenvectors) that the vector plot shows graphically: $(\ln \mathit{Cu})^*$ has coefficients 0.378 and –0.035, so its main contribution is to the first axis, increasing from left to right because the coefficient is large and positive, with only a slight decrease in the PC2 direction because of the small negative sign; $(\ln \mathit{Ni})^*$ has coefficients 0.149 and 0.466 so points slightly right (positive but small PC1 coefficient) and strongly upwards (large and positive on PC2), etc. The vector length reflects the importance of that variable’s contribution to these particular two PC axes, in relation to all possible PC axes – if the line reaches the circle then none of that variable’s other coefficients in the *Eigenvectors* table will differ from 0. The vector plot (or more clearly the eigenvector results table) show that PC1 is a roughly equally weighted combination of most of the heavy metals, Cu, Zn, Cd, Pb, Cr and organics, but not Co, Mn, Ni and Depth. The situation is reversed on the PC2 axis, with the first batch scarcely contributing at all, but the second set all increasing strongly in the positive PC2 direction. So, the first PC gives a natural way of combining the different contaminant levels into a single summary variable that characterises the main contaminant gradient.

Chapters 4 and 11 of CiMC give more on this particular example, but the principle of using a Principal Component axis as a natural, objective combination of a suite of variables is one that applies equally strongly to biomarkers, morphometric measures, water-quality metrics etc. The only difference in the latter case is that the metrics may already be standardised to a common impact scale (0 to 10, perhaps) so no prior transformation or normalisation is needed before PCA is carried out. For morphometric measurements too, transformation is often not needed and lengths, widths etc. may be in common units, but normalisation may still be needed if widely different measurement ranges are involved (overall body length, setae width), to stop the larger readings completely dominating the PC’s. For typical biomarker suites, transformation would need to be considered and normalisation would be essential, since entirely different scales are often involved.

Revision #29
Created 27 September 2024 21:54:04 by Arden
Updated 12 February 2025 00:35:57 by Abby Miller