1368 CD
DOI: 10.1017/S1431927607078063
Microsc Microanal 13(Suppl 2), 2007

Multivariate Statistical Analysis of Compositional Data within the Simplex

John Henry J. Scott

National Institute of Standards and Technology, Gaithersburg, MD 20899-8370

Many multivariate statistical tools assume the data is drawn from a continuous and unbounded sample space, such as the *d*-dimensional set of real numbers $\Re^d$. In these spaces, variables can take on values ranging from $-\infty$ to $+\infty$, including zero. In contrast, many of the data that appear in quantitative microanalysis experiments are drawn from more restricted spaces. The counts in an XEDS detector channel, for example, are constrained to be non-negative. It does not make sense to speak of a 10 eV wide X-ray channel with negative counts, and when commonly-applied algorithms such as Principal Components Analysis (PCA) return results with negative values, many analysts are confused. Compositional data are even more constrained than count data. Consider the microanalysis of a Ni-Al-Fe alloy in an SEM using XEDS. The measured composition at each pixel on the sample is expressed in terms of the mass fraction of each component of the mixture, such as $Ni_{0.25}Al_{0.5}Fe_{0.25}$. The mass fraction data for each component is constrained to vary from 0 to 1, an extremely narrow range. To complicate matters further, the three component values are jointly constrained to sum to one, i.e. $x_{Ni} + x_{Al} + x_{Fe} = 1$. Mathematically, the components are said to fall within a simplex, a severely restricted subspace of $\Re^d$.

The practical implications of these constraints, and the resulting pitfalls in interpretation of unconstrained statistical analyses, were described by Karl Pearson as early as 1897 in a paper on spurious correlations [1]. The effect of the unit-sum constraint on bulk compositional data was first investigated by geologists in the early 1960s [2,3]. Since then many descriptions of the consequences of these effects have appeared in the statistical literature, such as the negative-bias difficulty, the basis difficulty, the null-correlation difficulty, and the absence of interpretable covariance structure when using crude PCA and crude multivariate curve resolution (MCR).Only recently has a mathematically-mature approach for "stay-in-the-simplex" analysis appeared, based on an alternative form of linear algebra where the arithmetic operations of addition and multiplication are replaced with the binary operations of *perturbation* and *powering* [4].

Here, these techniques are applied to microanalysis data using custom scripts for the R open source statistical environment [5]. Experimental k-ratios, compositions from ZAF and φ(ρz) corrections, and synthetic spectrum images from NISTMonte and DTSA can be processed using the new approach. Figure 1 shows the effect on compositional error ellipses when Ni-Al-Fe data are analyzed using conventional ("crude") covariance structure vice "stay-in-the-simplex" methods. Figure 2 shows the same for PCA loadings.

References

[1] K. Pearson, "Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs", *Proc. R. Soc.* **60** (1897) 489-498.
[2] O.V. Sarmanov and A.B. Vistelius, "On the correlation of percentage values", *Dokl. Akad. Nauk. SSSR,* **126** (1959) 22-25.
[3] F. Chayes, "On correlation between variables of constant sum", *J. Geophys. Res.* **65** (1960) 4185-4193.
[4] J. Aitchison, *The Statistical Analysis of Compositional Data*, Blackburn Press, 1986, 42.
[5] http://www.r-project.org

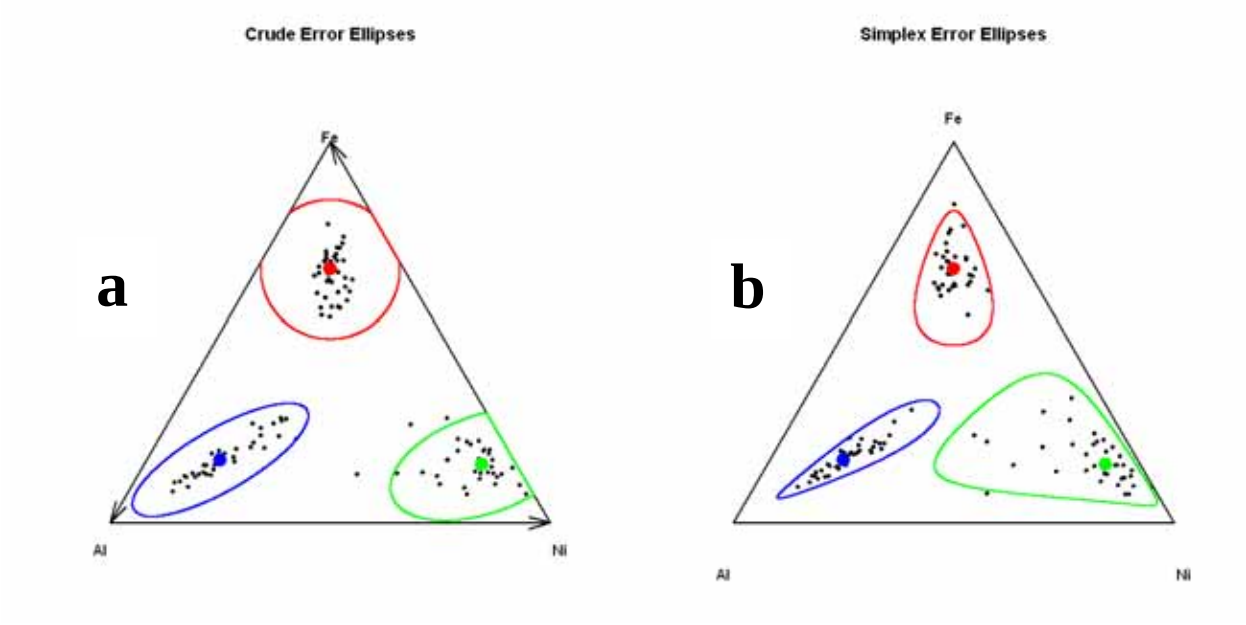

Figure 1. Differences between compositional error ellipses using **(a)** crude covariance structure, and **(b)** simplex-based analyses. The Ni-Al-Fe compositional data points are clustered around three phases with different covariance matrices. The ellipses in (a) are one sigma in radius (r=0.5 for green) and are ignorant of the ternary diagram boundaries. Those in (b) are three sigma ellipses and obey the simplex constraints.

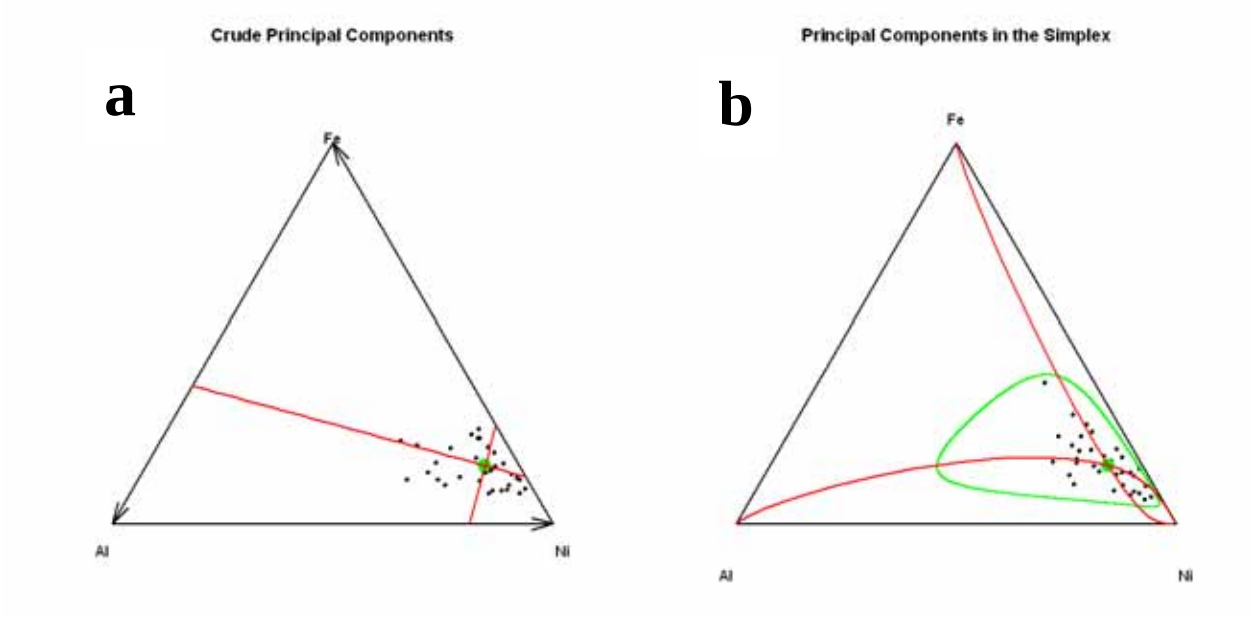

Figure 2. Differences between the first two principal component directions for the green phase only, using **(a)** conventional PCA, and **(b)** Aitchison simplex compositional PCA (with three sigma error ellipse drawn for reference). These differences apply to all multivariate statistical analyses performed using the crude covariance (including pixel classification schemes) and are not limited to PCA and MCR.