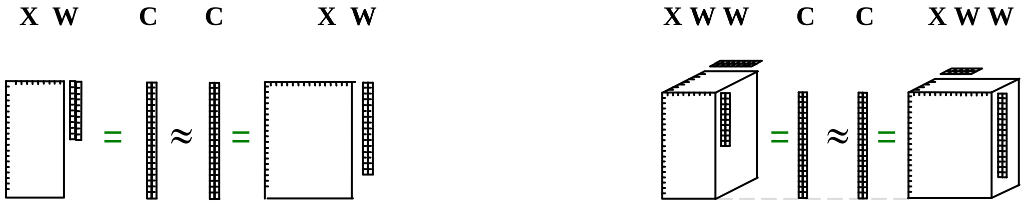
Introducing new ‘ways’ into data analysis:

Making linear models *multi*linear gives them important new properties

Richard A. Harshman, Psychology Dept., University of Western Ontario
London, Canada, harshman@uwo.ca http://publish.uwo.ca/~harshman

In 1970, I developed PARAFAC (PARAllel FACtor analysis), a generalization of factor/component analysis from matrices to three-way arrays of data (e.g., to measurements of n cases on m variables on each of p occasions, or to correlations of n variables with the same n variables in each of p different circumstances). The motivation was to enhance *validity*: by parallel factoring of multiple non-identical mixtures of the same patterns, the three-way model could often overcome the rotational ambiguity of standard factor/component analysis and uniquely recover the source patterns that originally generated the mixtures, without any restrictions on the loadings.. In the last 10 years there has been a rapid growth of important PARAFAC applications in diverse fields, ranging from chemistry and physics (e.g., E-E flourescence and XES x-ray spectroscopy), to signal engineering (e.g., cell-phone signals, noisy radar), to neuroscience (EEG and fMRI brain signals), etc. A Google search for “parafac” now returns over 50,000 hits. After explaining the PARAFAC generalization, I will focus on how similar generalization of other methods gives them similarly improved properties.

Extending almost any of the standard statistical models from linear to multilinear can, in appropriate applications, give it substantially increased power and important new properties. The extensions can be briefly explained as follows: while traditional methods find an optimal linear combination across one index of a two-way data array (combining columns of data), the generalized methods find jointly-optimal linear combinations across two (or more) indices of a three- (or higher)-way array. The figure below shows how a standard canonical correlation for the General Linear Model (GLM) is modified for a “level 1” multilinear generalization. The canonical weight vectors (columns of **W** on both sides) are chosen so that the correlation between the left and right canonical variates (columns of **C**) is maximal. Note that the data sources on the two sides do not need to have the same number of ‘ways’, so either side can be a matrix or a four-way array, etc.



By introducing multilinear generalizations into the General Linear Model, this approach implicitly also generalizes its many special cases, such as Discriminant Analysis, (M)ANOVA /(M)ANCOVA, etc. In many of these applications, one side of the canonical relation would be a ‘design matrix’ or ‘design array’. Statistical tests could be based on distribution free compute-intensive methods such as randomization tests or bootstrapping.

A further kind of generalization will also be described, called “level 2 multilinearity”. Here, the *patterns themselves* are multilinear, and take the form of matrices or arrays with low-rank outer-product structure. For example, in the level 2 GLM, the canonical variates become tensors of order 2 or higher. Patterns with such added structure can convey “deeper” or “higher order” information about the data generating processes, including how specific latent properties in one ‘way’ of the array ‘interact’ or act jointly with specific latent properties in another.