SVD per se

Sometimes the SVD is presented as an end in itself. The most famous example is latent semantic indexing (LSI) which represents the term document matrix via $X = \sum_k \sigma_k u_k v_k'$. The k'th term in this sum can be interpreted as the k'th topic in the corpus.

A query is represented by a vector q shaped like a column of X. In the vector space model, the documents get ranked by the value of $X'q$. If we approximate X by $\hat{X}$, truncating the SVD, we then rank documents by $\hat{X}'q$. In the approximation $\hat{X}$ some of the u_k will pool together terms that tend to co-occur frequently. When things go well, synonyms get pooled. This is helpful for a sparse query. To take an example of Kolda and O'Leary, a query with "Mark Twain" could then pull up documents containing "Samuel Clemens" (Twain's real name) even if those documents did not contain the term (or terms) in "Mark Twain".

Notice that the topics are represented by orthogonal vectors. That is not necessarily natural. If we have two different topics, then they might plausibly overlap somehow. It could be hard to draw the line between two overlapping topics, and orthogonality is one way to force a choice. We'll see others later.

LSI typically gets used with modestly large numbers of singular vectors, say 100 to 150. Then $\hat{X}$ can be stored in much less space than X. Note that $\hat{X}$ is typically dense.

In bake-offs LSI has mixed results. Manning and Shutze say it does well in high recall searches. It can have poor precision, attributed to noise. The meaning of the word 'noise' in information retrieval may be slightly different than what statisticians have in mind. There one sometimes sees $\hat{X}$ described as adding noise to X (filling in 0s for example). The statistical view is usually that X might be a noisy version of some latent quantity close to $\hat{X}$.

A big problem with LSI is that new documents keep arriving. These may bring new terms, and may affect the inverse document frequencies too. Even if the only change is adding more columns to X, it is expensive to keep redoing the SVD. There are approximate 'folding in' methods to keep $\hat{X}$ up to date.

Crop Science

Models of the SVD type have long been used in crop science. Back in 1923, Fisher and MacKenzie had employed a model for potato yield of the form

$$Y_{ij} \doteq \mu + \alpha_i + \beta_j + \lambda \gamma_i \delta_j$$

which is a two way anova plus one term of an SVD. Here i represents a plant variety and j represents a manurial treatment. Subsequent authors modeled genome by environment interactions in general.

This aspect of Fisher and MacKenzie's work seems to have been overlooked. In the intervening decades log or square root transformations have been more popular.

Another famous example from the ANOVA literature is Tukey's one degree of freedom for non-additivity. Tukey fits a model of the form

$$Y_{ij} \doteq \mu + \alpha_i + \beta_j + \lambda\alpha_i\beta_j.$$

By reusing the main effects, only one more degree of freedom must be fit.

More modern models are in the linear-bilinear framework of Gabriel (1978). Let Y be an $n \times m$ matrix of responses, representing for example, one row per variety and one column per treatment. The linear-bilinear model has the form

$$Y \doteq X\beta + H$$

where X is an $n \times k$ matrix of predictors and H has rank r. We may write this model as

$$Y \doteq X\beta + Z\gamma$$

where Z is $n \times r$ and γ is $r \times m$. The factors Z and γ in $Z\gamma$ are not identifiable because $Z\gamma = ZAA^{-1}\gamma$ for any invertible $r \times r$ matrix A.