HEATMAPS

Leland Wilkinson
SPSS Inc.
Northwestern University

MGA Workshop III: Multiscale Structures in the Analysis of High-Dimensional Data
October 25-29, 2004
UCLA

A heatmap is a tiling of a planar region, with each tile colored according to a color scale.

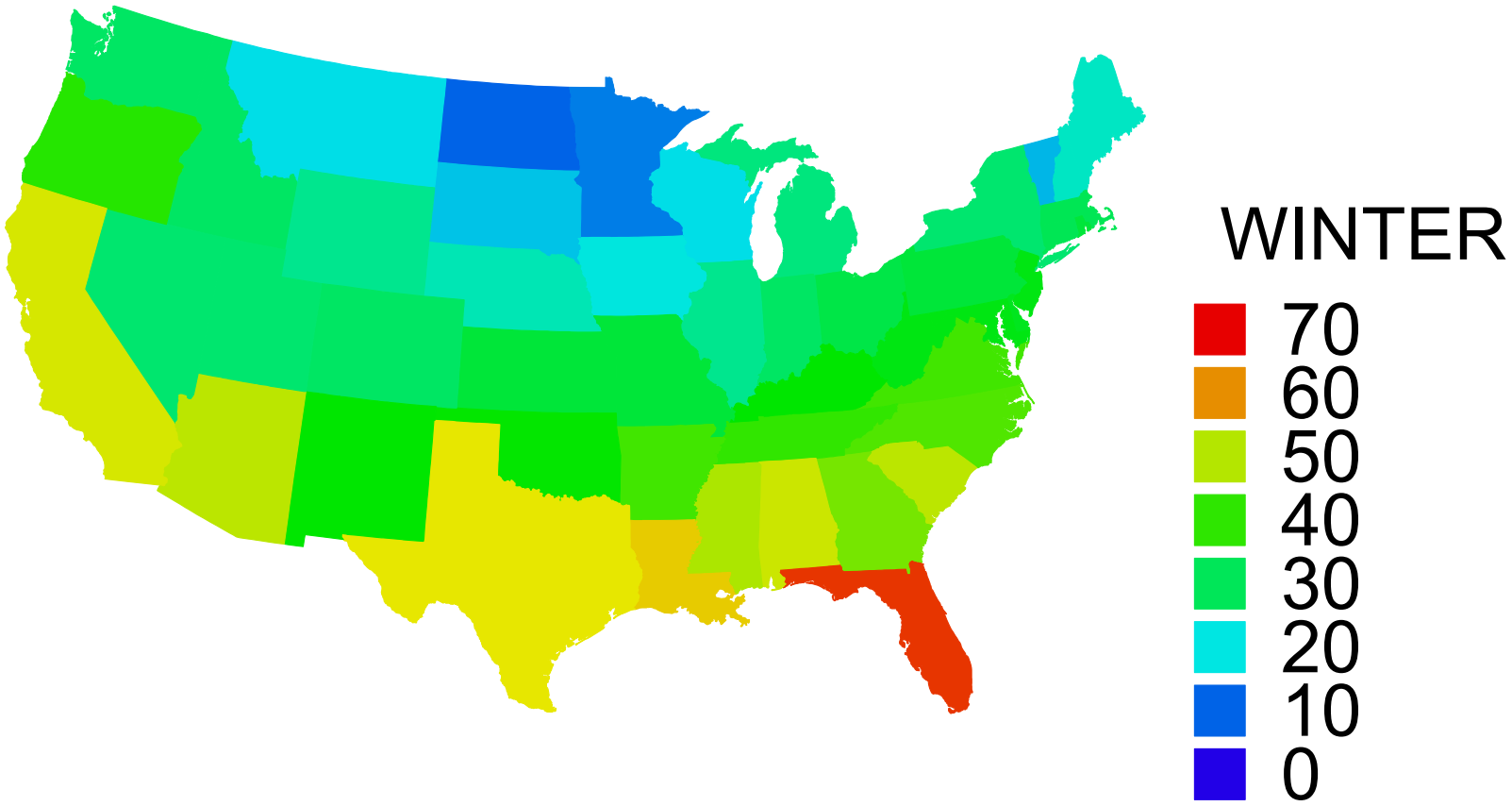
Tilings are usually rectangular.

Color scales are usually chosen to suggest temperature or other perceptual dimensions.

A simple idea. How useful is it?

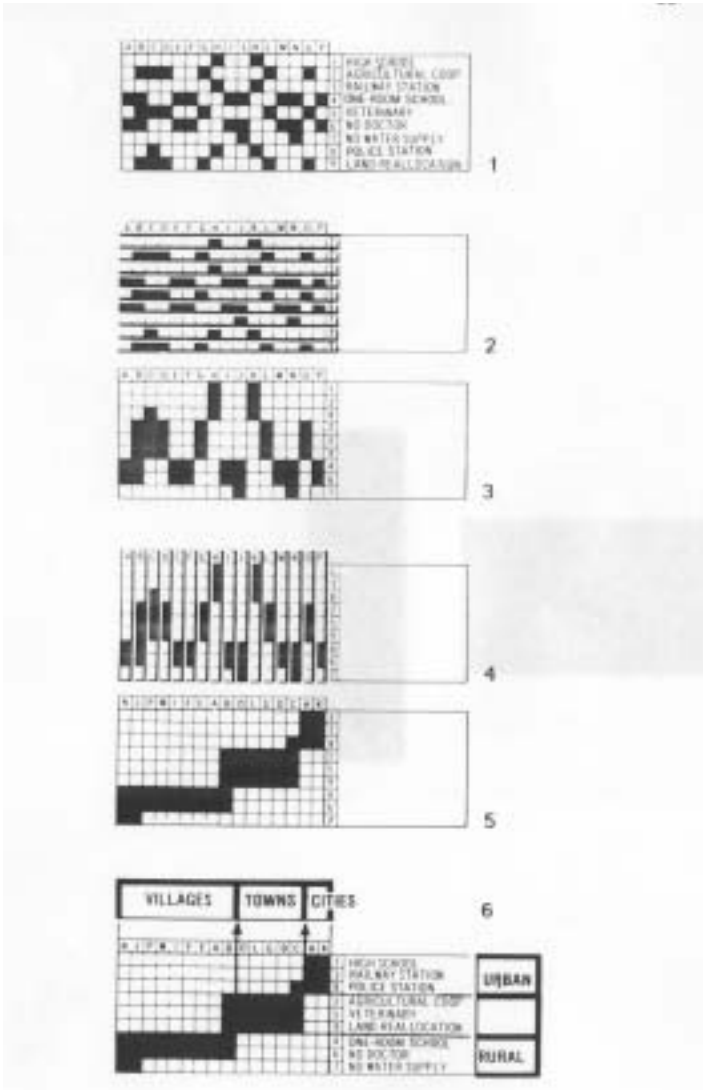
It all depends on how you order the map.

Heat Map (Old, very old)



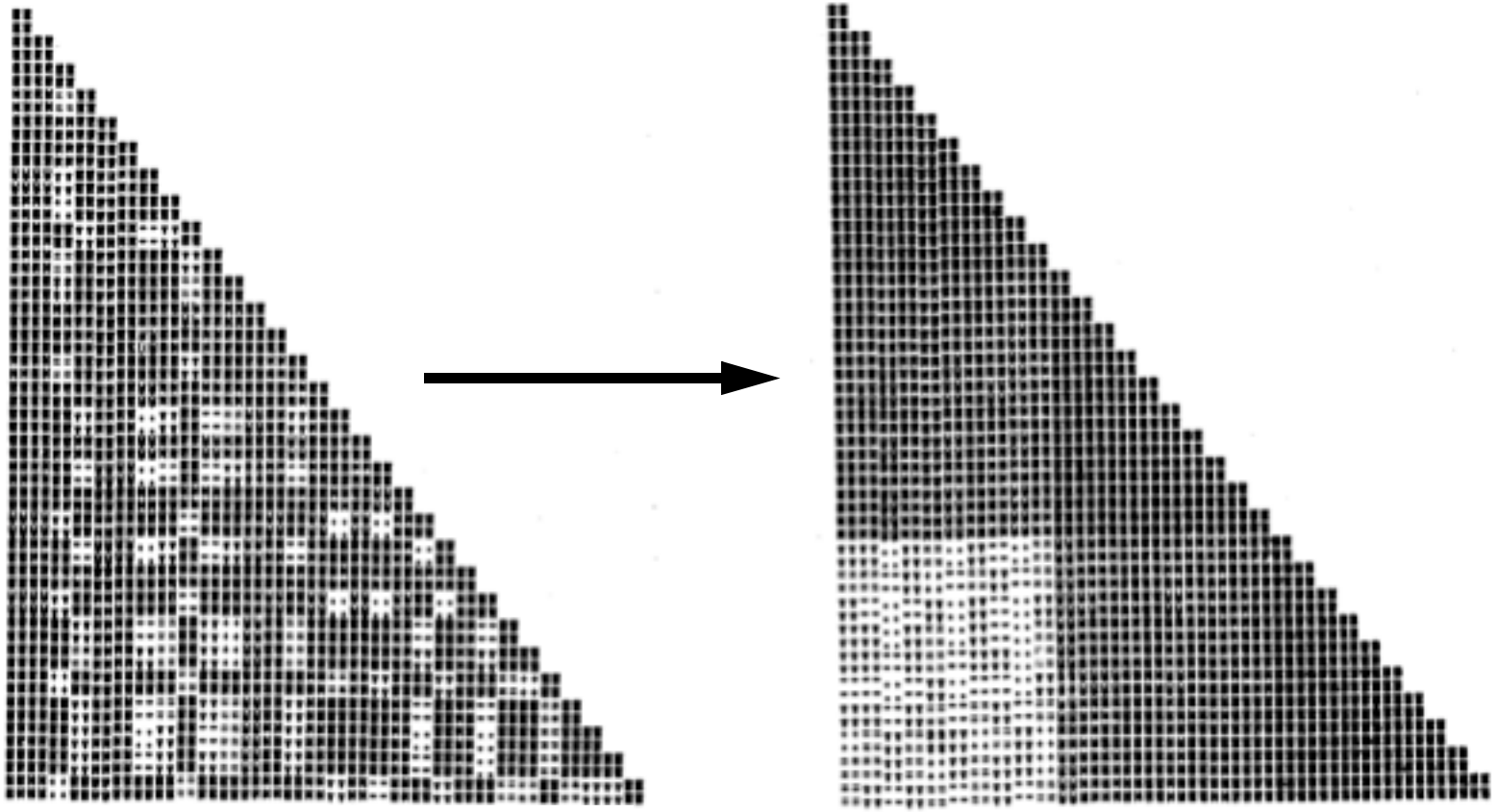
Geography's not a bad way to order the world. Aggregating over states, however, makes less sense for temperature than for electoral college data!

Permuted Matrix (Bertin, 1967)



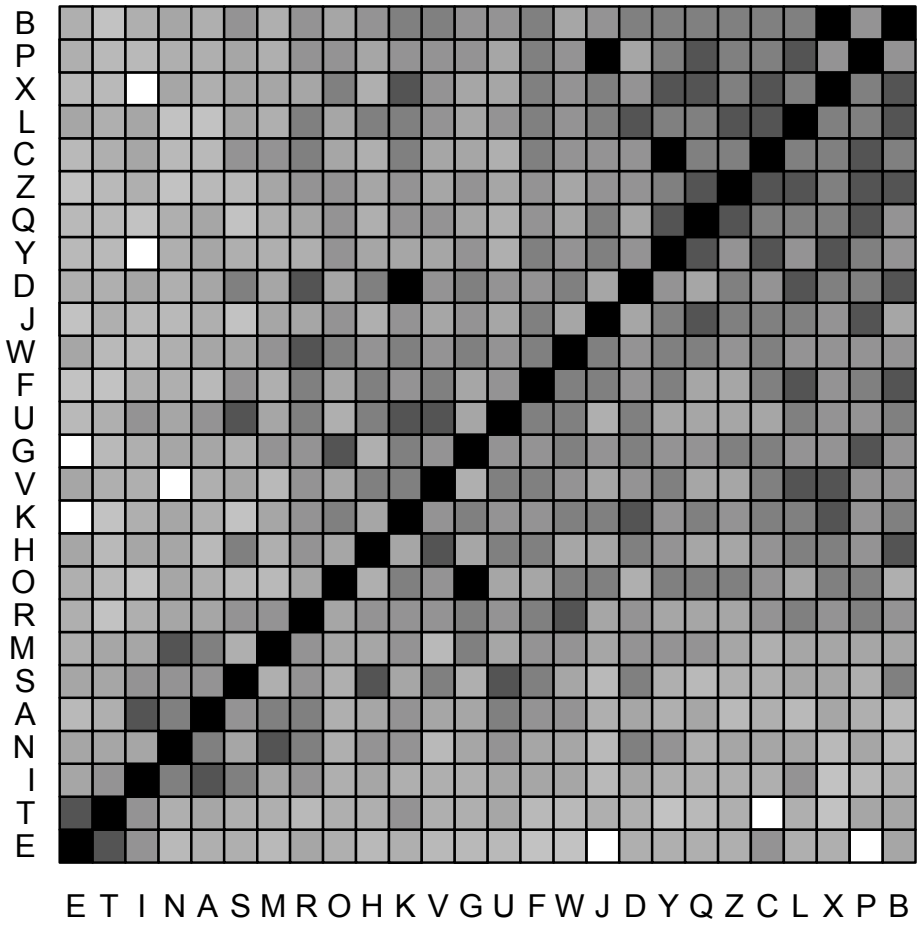
Bertin knew where he was going. He ordered on rural/urban vs. available public services.

Shaded Correlation Matrix (Ling, 1973)



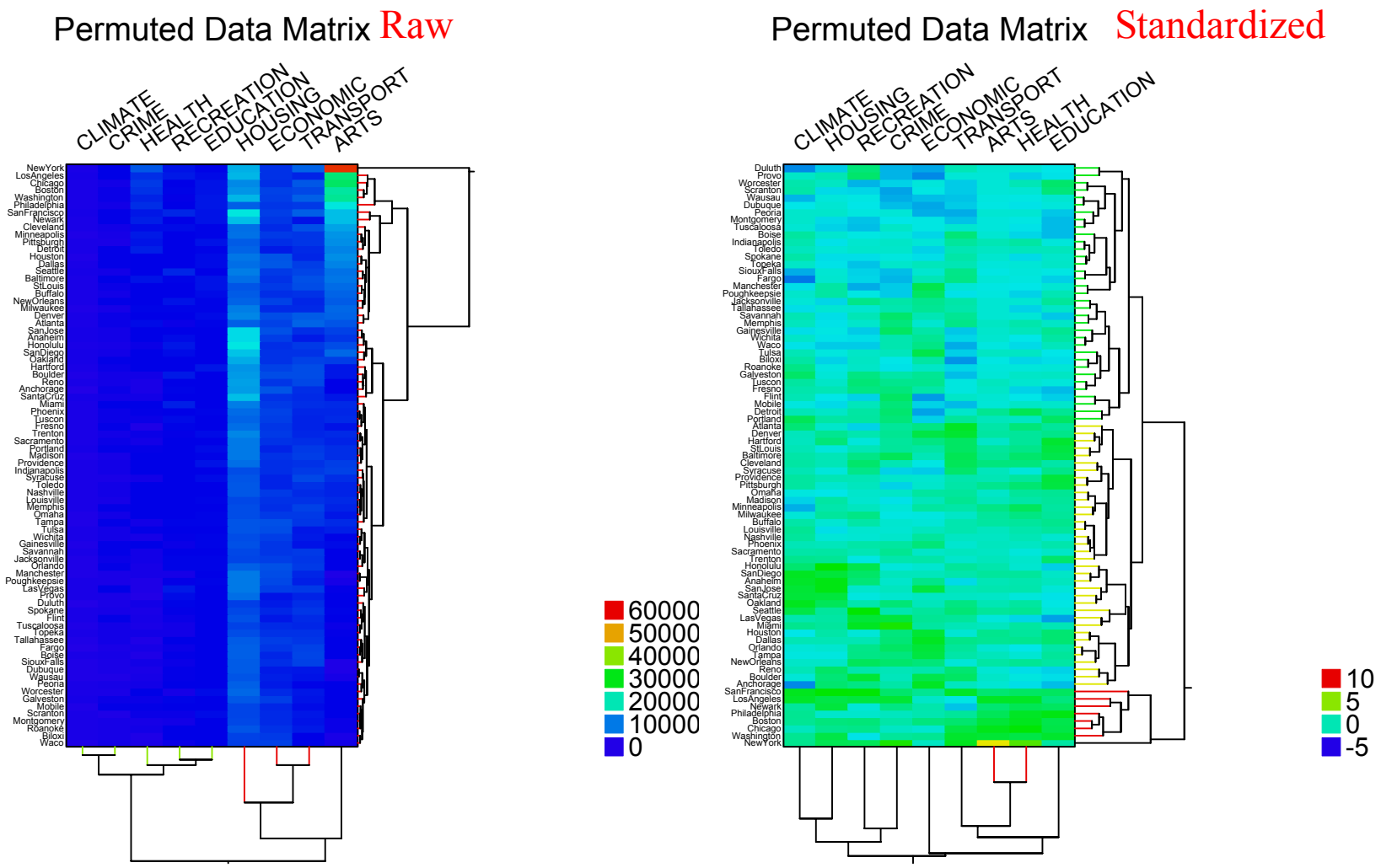
Ling was looking for factors (this figure adapted from Everitt).

Shaded Confusion Matrix (Rothkopf data)



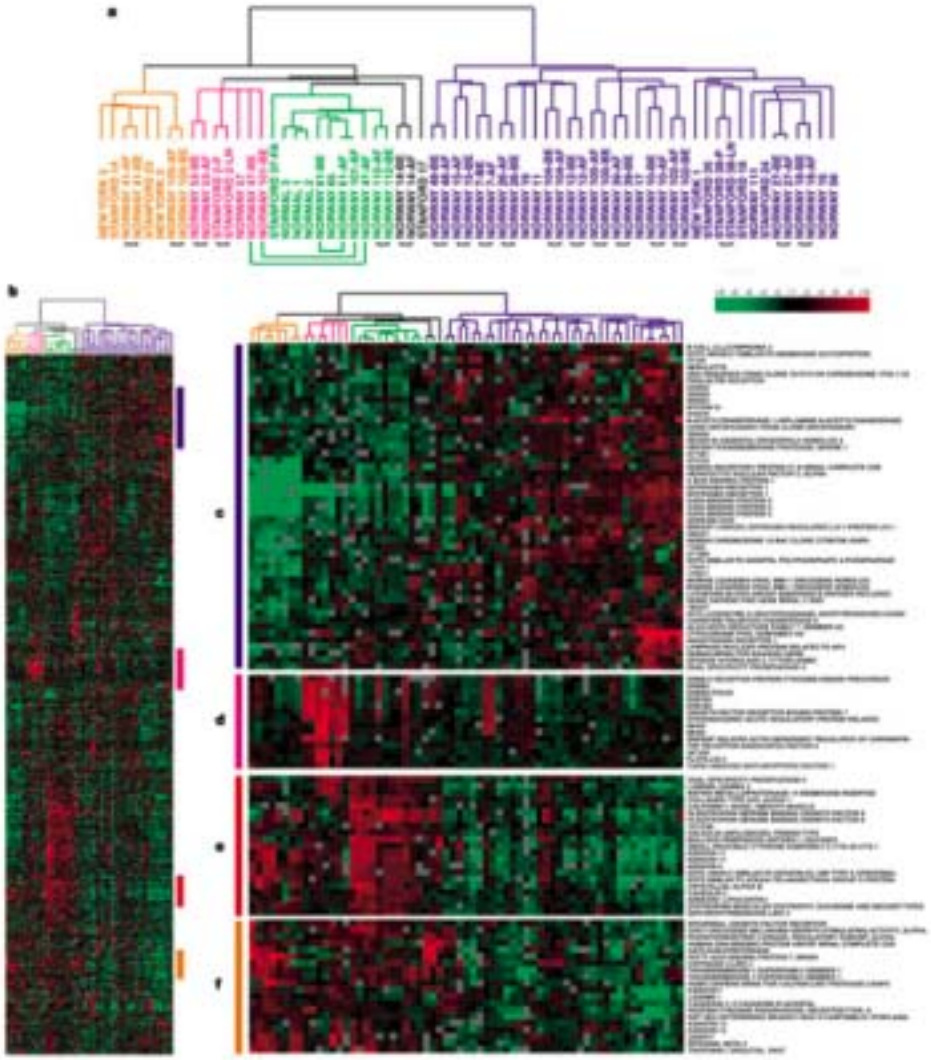
This matrix is asymmetric because there are order effects in the confusion of symbols.

Direct Clustering of a Data Matrix (Hartigan, 1972; SYSTAT 1987)



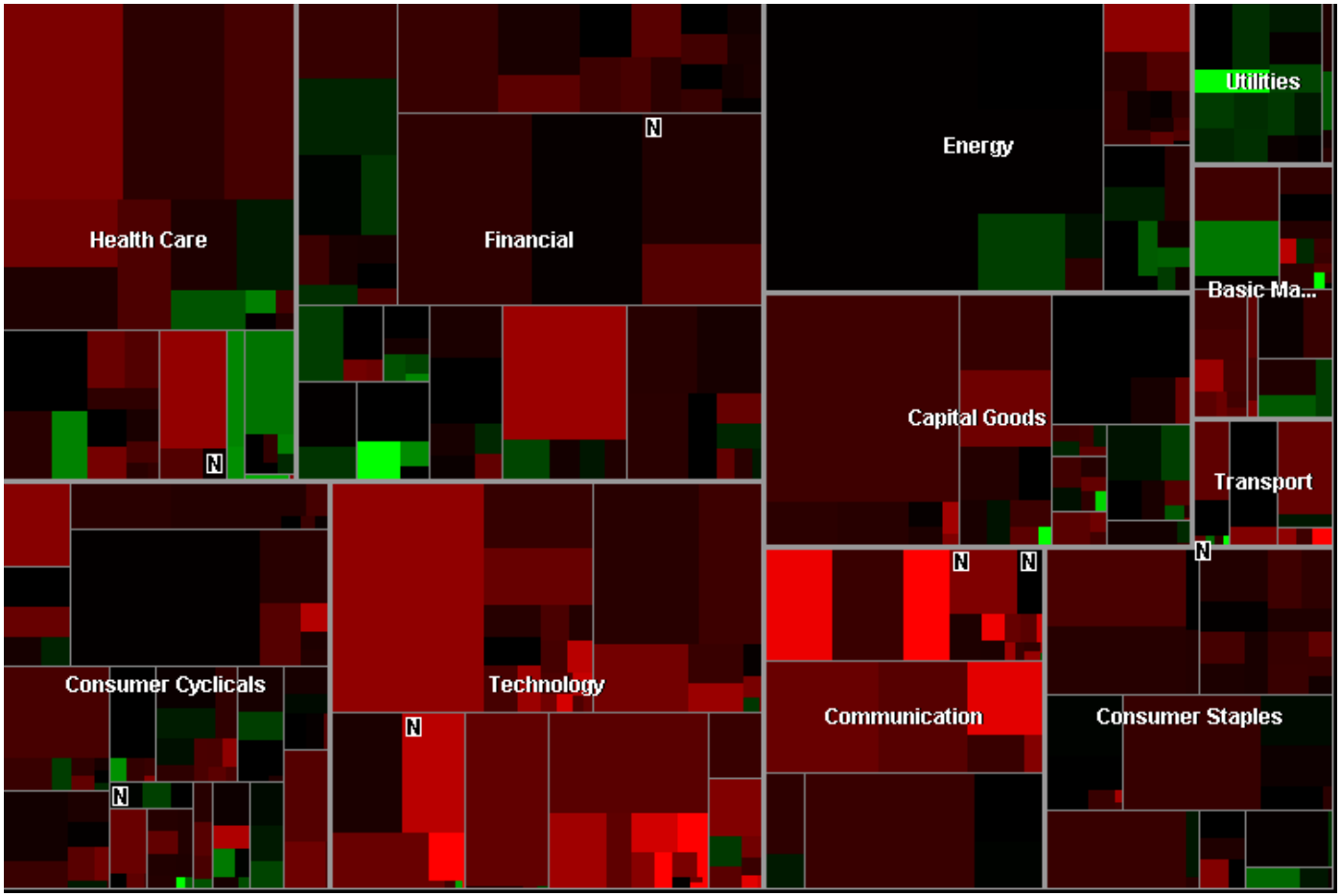
Hartigan had ANOVA in mind. SYSTAT clustered rows and columns independently.

Microarray Plot (Perou *et al*., 2000)



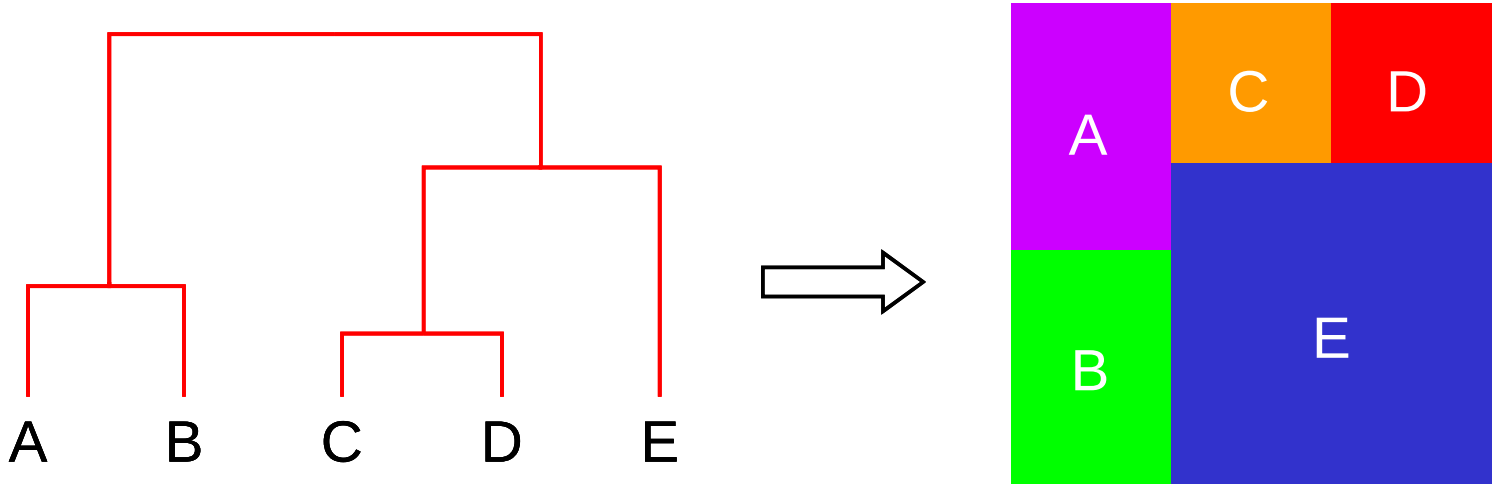
Two-way model (following SYSTAT model, not Hartigan's).

Treemap (Schneiderman, 1992)



www.smartmoney.com

Treemap Scheme



$Node \rightarrow Block$: Sequence = <horizontal, vertical, horizontal, vertical, ...>

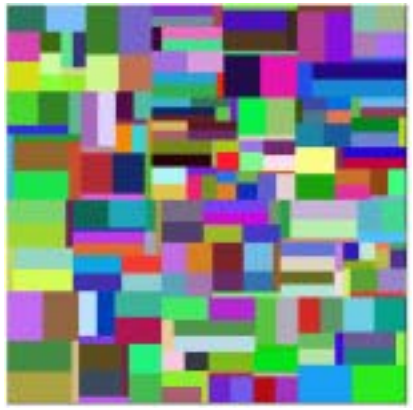
Area($Block_i$) $\propto X_i$, where X is extrinsic variable

Hierarchical Clustering Treemaps (data from Kooijman, 1979)

Single

Complete

Average



For single linkage, the correlation of the inter-rectangle distances with the original data distances is only .14. For complete linkage, it is .33. And for average linkage, it is .39.

Matrix Permutation

$$\mathbf{X} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix}$$

$$\mathbf{P}_L = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{P}_R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

P is square and binary with rowsum = 1 and colsum = 1 for all rows and columns.

$$\mathbf{P}_L\mathbf{X}\mathbf{P}_R = \begin{bmatrix} j & l & k \\ d & f & e \\ a & c & b \\ g & i & h \end{bmatrix}$$

$\mathbf{P}_{L1}\mathbf{P}_{L2}\mathbf{P}_{L3}\cdots\mathbf{X}\,\mathbf{P}_{R1}\mathbf{P}_{R2}\mathbf{P}_{R3}\cdots$

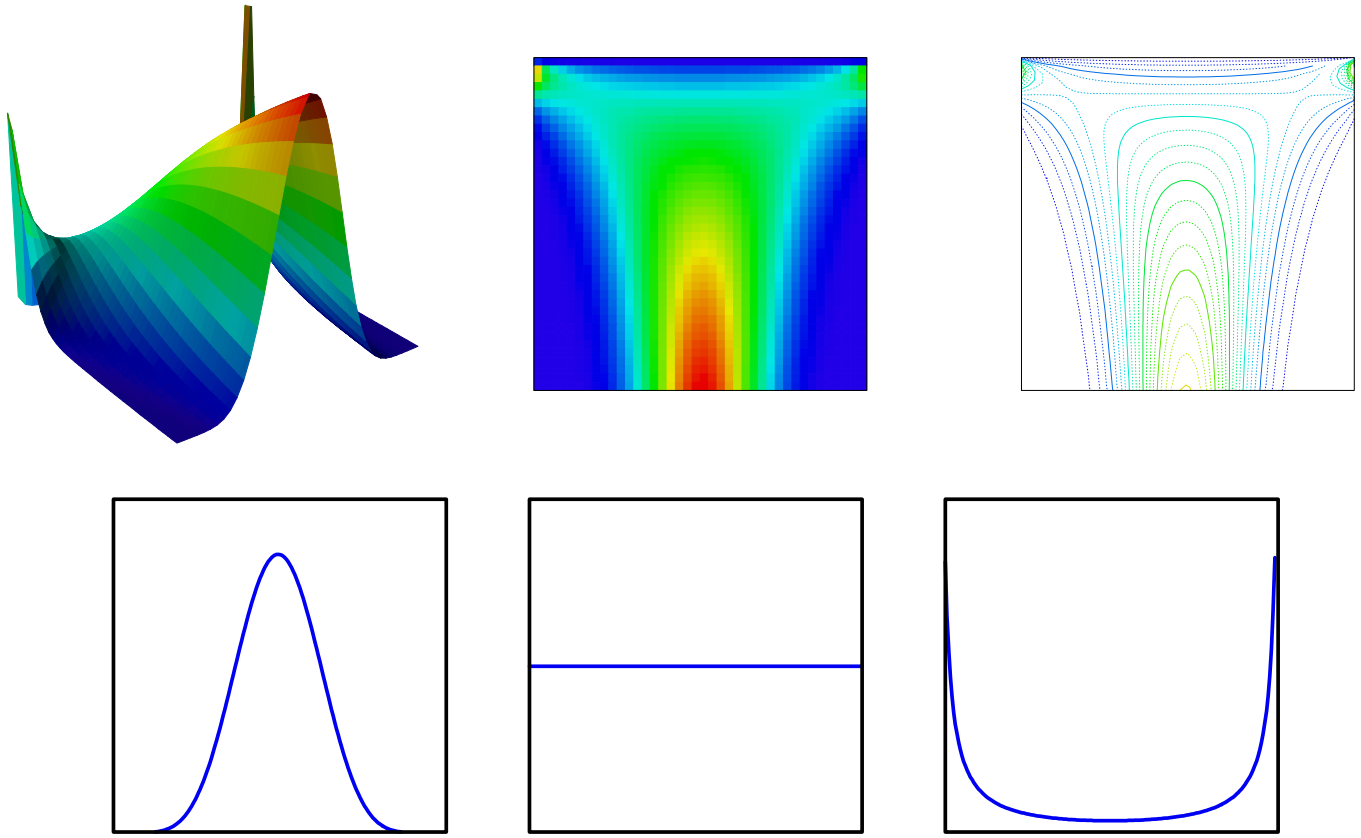
P is composable

For heatmaps, we seek $\mathbf{P}_L$ and $\mathbf{P}_R$ that yield low[$Loss(\mathbf{P}_L\mathbf{X}\mathbf{P}_R)$]. Heatmap users seem to define $Loss()$ as lack of relative color homogeneity within a local region of the heatmap PLUS more than one connected region having the same color.

In other words, users want to permute successively the rows and columns of a real matrix until similar-valued entries (pixels) are near each other. However, their class of loss does not include knowledge of, or inference concerning, the function that generated the values in the matrix. Is there an algorithm that gives them what they want for general data?

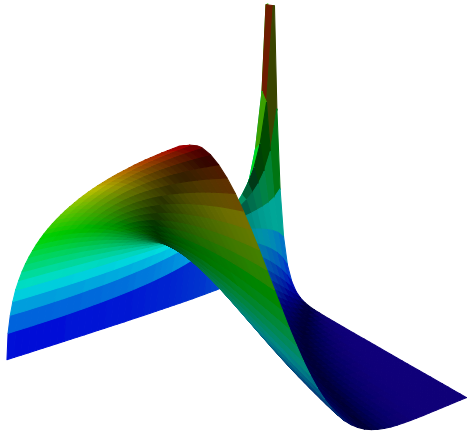
Let the column and row indices ($j = 1, ..., p, i = 1, ... n$) of $\mathbf{M}_{n\times p} = \mathbf{P}_L\mathbf{X}\mathbf{P}_R$ map one-to-one to the integer mesh $L = L_x \times L_y = \{1, \ldots, p\} \times \{1, \ldots, n\}$, and let $f(x, y)$ represent the corresponding real-valued entries of $\mathbf{M}$. This definition allows us to view the heatmap as a functional surface for a given permutation. Now we can view the implications of a local loss function (as opposed to a global model).

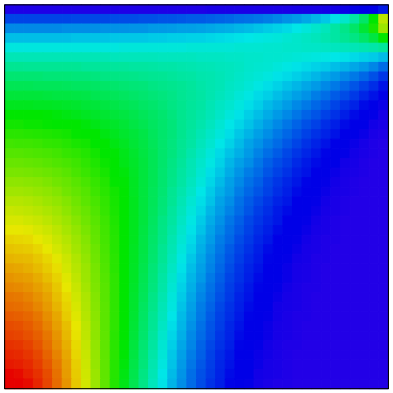
Not Cool (higher loss because of symmetry):

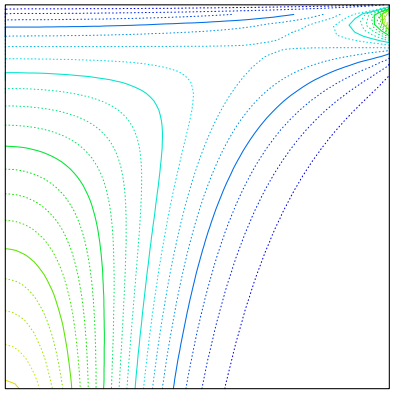


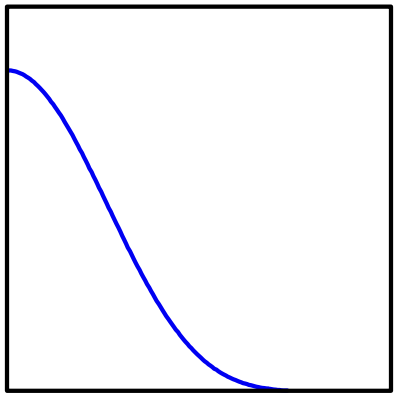
If $f(x, y_i)$ is symmetric or anti-symmetric about some x_j for all $i = 1, \ldots n$, then our permutation has multiple regions with one color. The same applies to $f(x_j, y)$. This example folds at $j = (p+1)/2$. The heatmap is shown at top center. The (smoothed) function is graphed at top left and contoured at top right. The bottom sequence shows $f(x, y_i)$ at $i = 1$, $i = n/2$, and $i = n$.

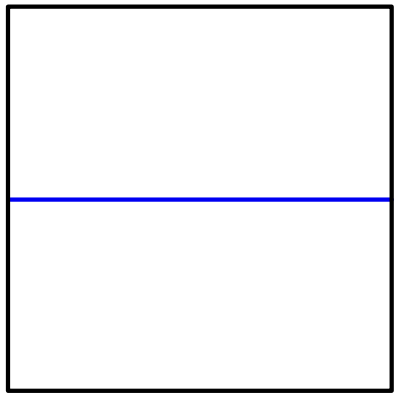
Fold symmetric functions:

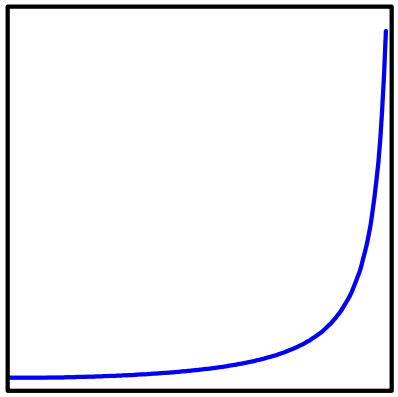






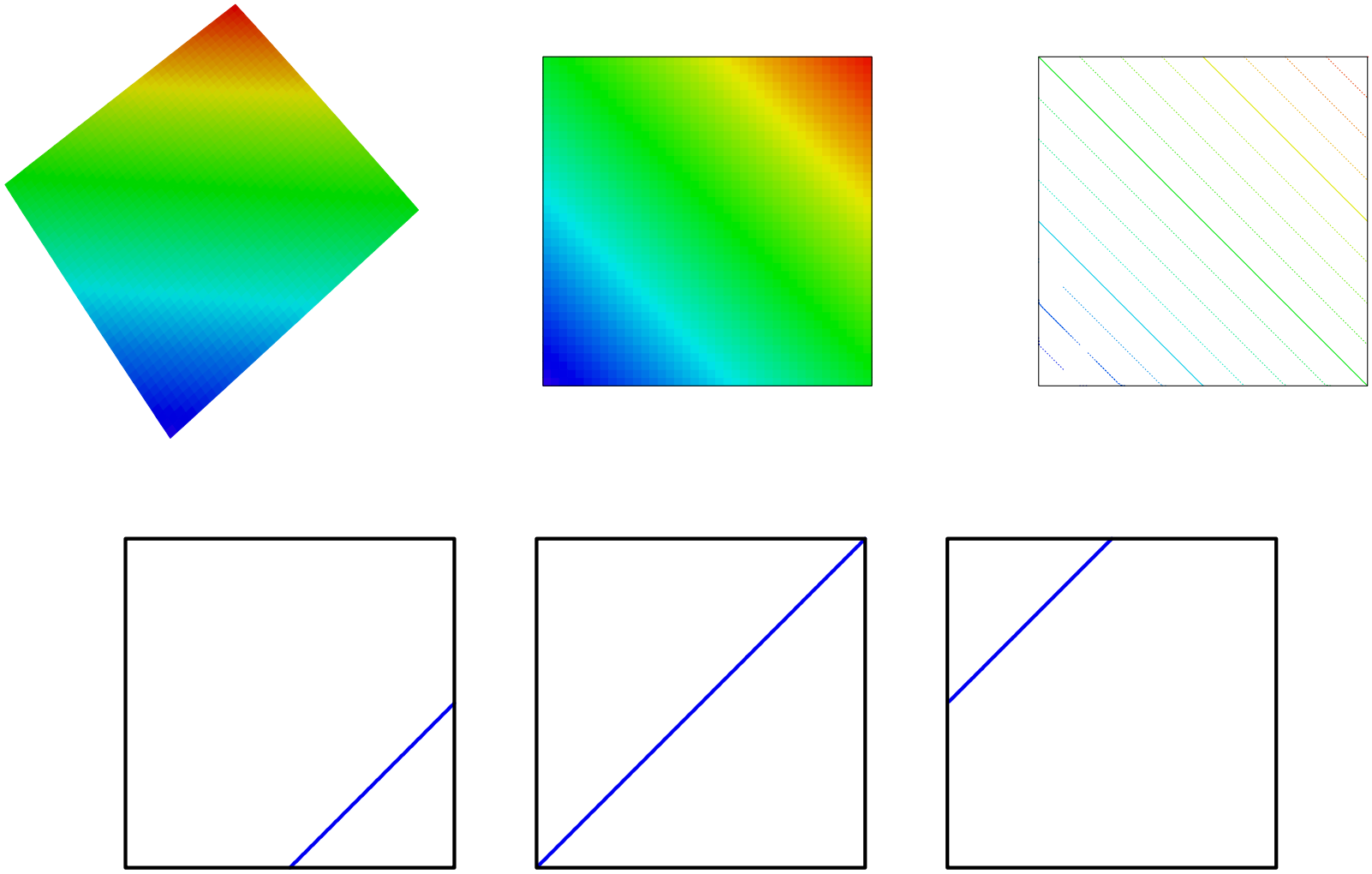






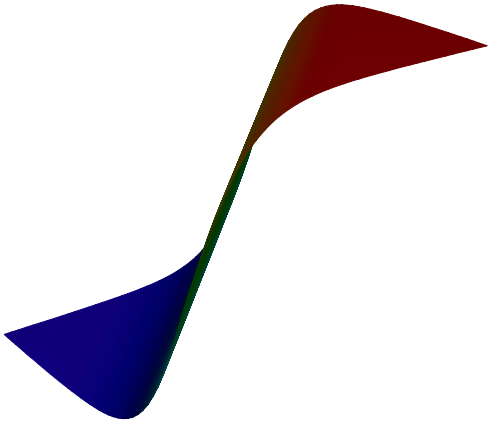
We folded it. Now it’s simpler (lower loss).

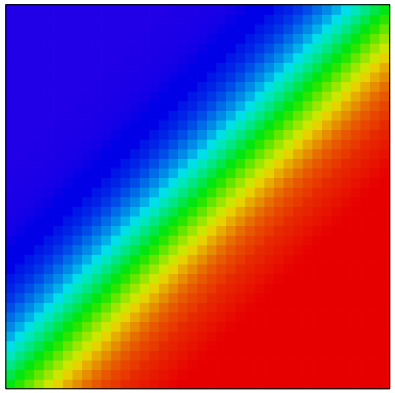
This has lower loss.

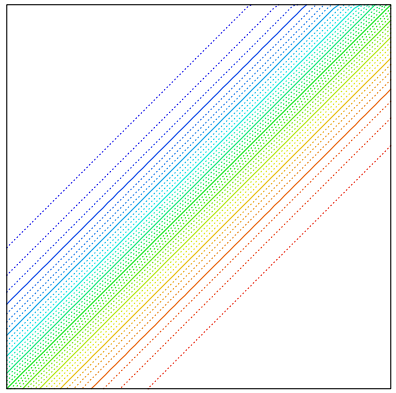


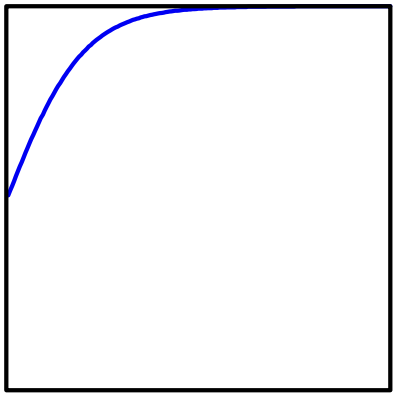
This is a main-effects model (no interaction between rows and columns). Think of partial derivatives or parallelism of level curves.

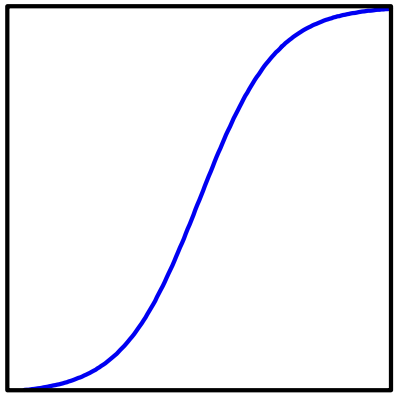
So does this.

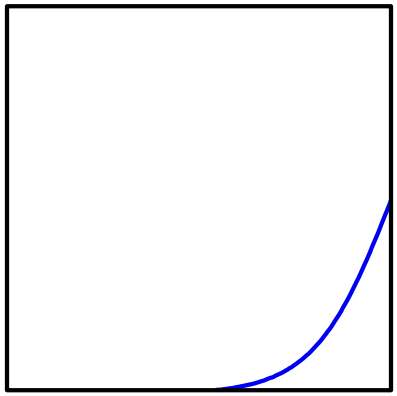






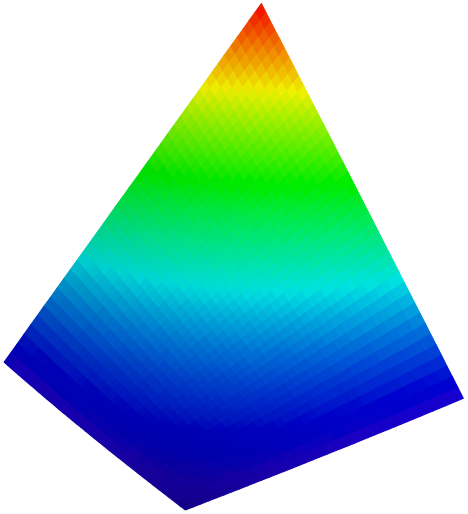


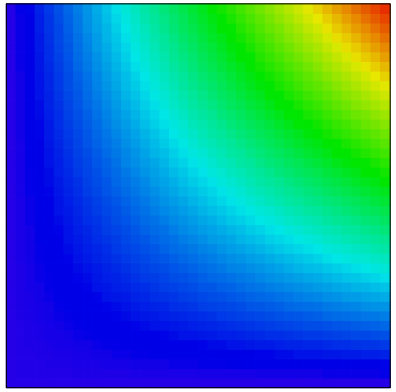


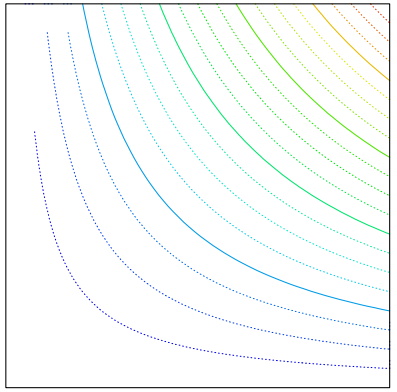


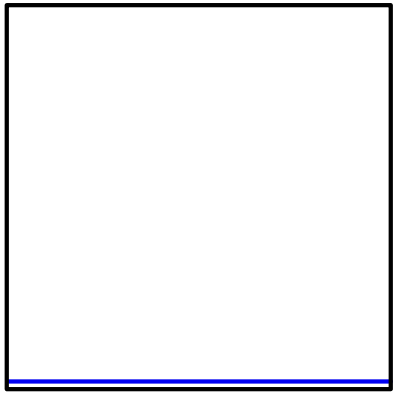
This is another main effects model under logistic transformation.

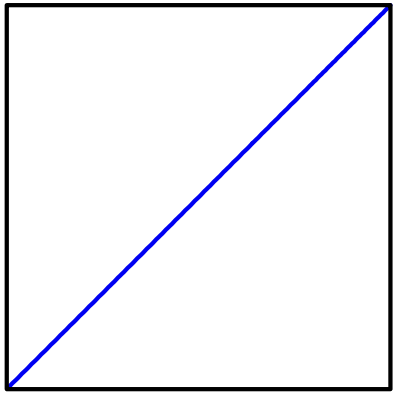
So does this.

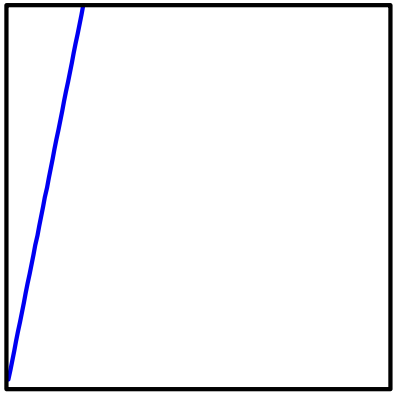






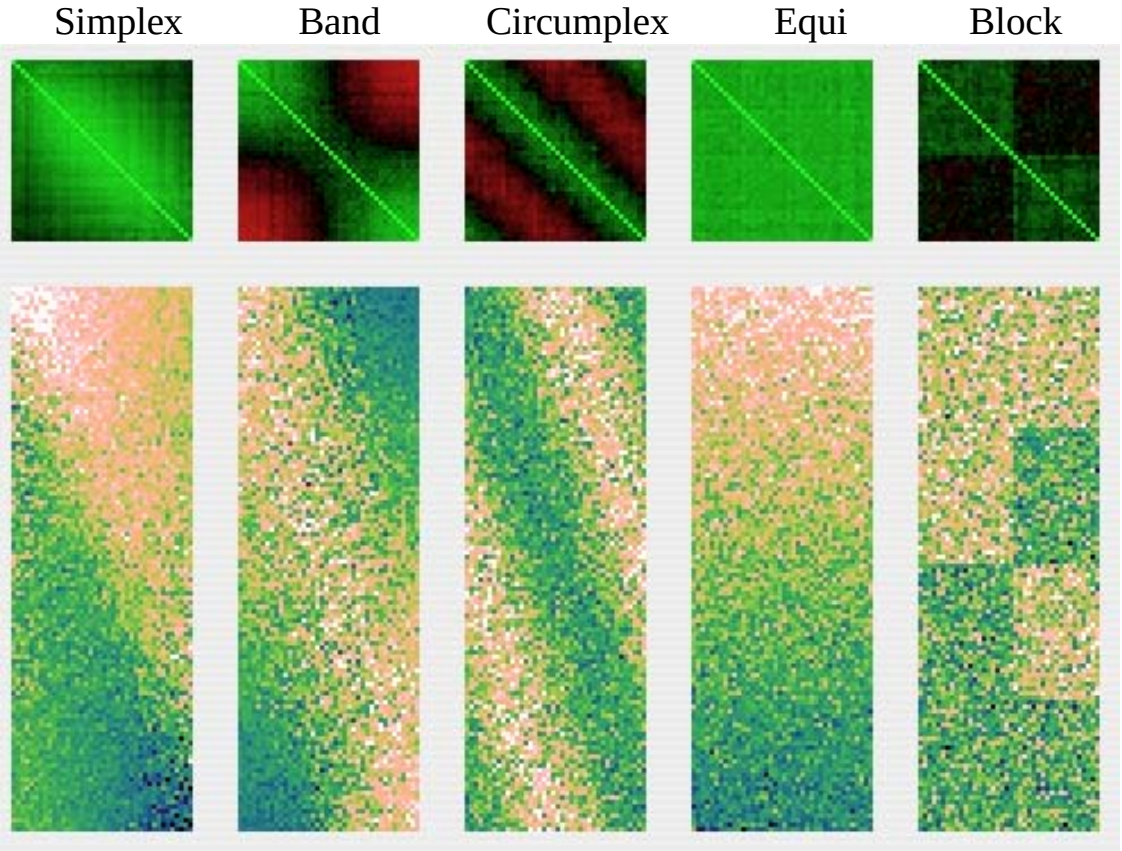






Big-time interaction going on here. It's a product model.

What kind of data have simple permutations?



These five patterns occur frequently in scientific datasets. Above each matrix is a heatmap of the covariance matrix on the columns (red for negative correlations, black for zero, green for positive). All except Block are (within error) topologically one-dimensonal. Simplex maps to a spiral, band maps to a line, Circumplex maps to a circle, and Equi maps to a line. Block is two-dimensional.

Simplex

The data values in the first dataset (Simplex) were generated from the following formula (ignoring rescaling the columns to unit standard deviations):

$$x_{ij} = e^t/(1 + e^t) + u_{ij},$$

where

$i = 1, \ldots, n, \quad j = 1, \ldots, p,$

$t = (s_j - r_i)/b,$

$s_j = j/p$, and

$r_i = i/n.$

The positive, nonzero parameter b determines the slope of the logistic function generated by the exponentials, and thus the sharpness of the boundary between the blue and pink regions. The random error u_{ij} is based on a weighted standard normal random variable Z:

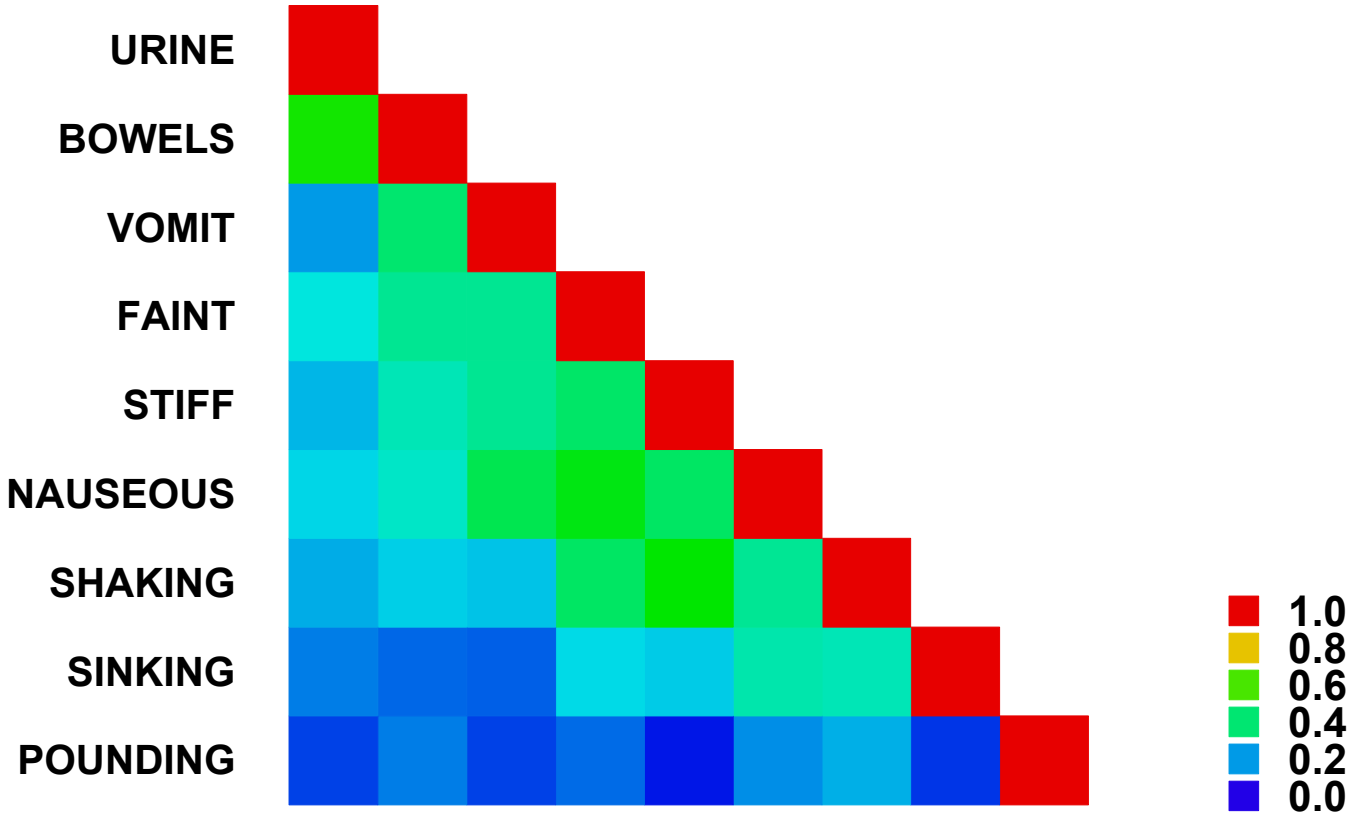
$$u_{ij} = wZ,$$

where

$$w = ke^{-t^2}$$

This dataset is named after the term coined by Louis Guttman (Guttman, 1954). If b is near zero and k is zero, we have the data structure Guttman called *simplex*, As Guttman noted for simplex-type data structures, the correlations are all positive; correlations between near columns are high and the correlations between distant columns are low. Guttman's simplex correlation pattern is a specific case of a **Toeplitz matrix**. If the columns of X can be assumed to consist of measurements ordered in time, then a first-order autoregressive model based on the indices of these columns produces a **Markov process** whose correlation matrix has a simplex structure (Morrison, 1976). In social measurements, the **latent structure model** (Lazarsfeld and Henry, 1968) and **ordered multinomial model** (Goodman, 1978) involve a simplex correlation structure. In educational measurement, the **latent trait model** (Rasch, 1960; Lord and Novick, 1968; Bock, 1975) is a generalization of the simplex. In the latent trait model, the columns of X represent an item difficulty dimension (one column per item) and the rows of X represent subjects (*e.g.*, students). Each row of data contains a set of item scores (usually binary) for a given subject. The estimate of a subject's ability is based on a logistic or normal cumulative distribution fit to the profile of test scores. In physics, Brownian motion follows a **Wiener stochastic process**. A Wiener process is a continuous-time stochastic process $W(t)$ for $t \geq 0$, with $W(0) = 0$ and $W(s) - W(r) \sim N(0, s - r)$. The differences are normally distributed and the covariance matrix is a simplex. In archaeology, seriation problems are defined in terms of a Robinson matrix (Robinson, 1951), which is a Toeplitz dissimilarity matrix.

Here's a simplex example (Stouffer *et al.*, 1950):



Correlation matrix of symptoms of combat stress in WWII

Band

The data values in the second dataset (Band) were generated from the formula

$$x_{ij} = e^{-t^2} + u_{ij},$$

where the parameters are defined similarly to the simplex model. Unlike the simplex, correlations are both positive and negative; correlations between near columns are positive and correlations between distant columns are negative.

Thurstone (1927) proposed a **law of comparative judgment** that involved a one-dimensional probabilistic, nonmonotonic magnitude item scale. Coombs and Avrunin (1977) presented a similar model for preferences. If the columns of **X** index a set of ordered objects and the rows represent a set of subjects, then a subject's preference for a set of objects can be represented by a (usually single-peaked, symmetric) probability density function centered on or near the most preferred object. Coombs, Dawes, and Tversky (1970) describe several other varieties of this model. In some physical systems (acoustics, optics), single-peaked multivariate spectra exhibit a banded correlation structure when ordered along a common frequency dimension according to spectral sensitivity. In information retrieval, citation patterns sometimes follow a band structure (Packer, 1989). Related texts cite each other but unrelated texts do not. More general citation and hyperlink data often involve a band structure.

Circumplex

The data values in the third dataset (Circumplex) were generated from the same formula used for Band, except the variate has been circularized:

$$x_{ij} = e^{-t^2} + u_{ij},$$

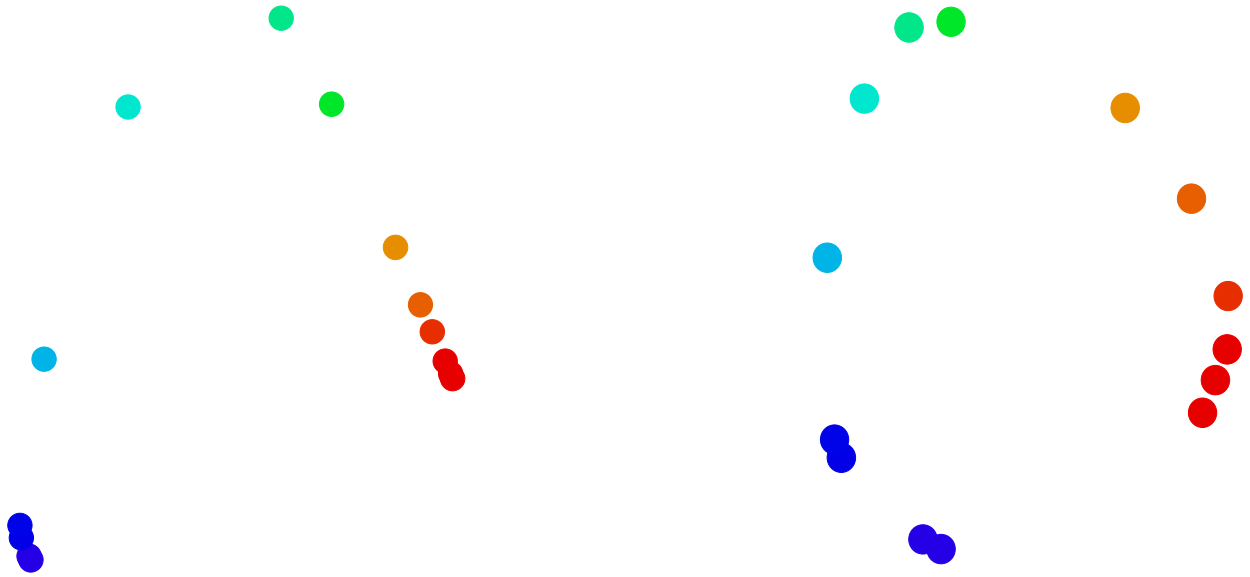
where

$$t = \cos(\pi(s_j - r_i)),$$

This dataset is named after a term coined by Louis Guttman (Guttman, 1954) to describe a circular correlation structure. In a circumplex correlation matrix, correlations between near columns are positive and correlations between distant columns are negative. As with Guttman's simplex, the circumplex correlation matrix follows a Toeplitz pattern. With the circumplex, however distance is measured on the circumference of a circle. That is, the series $j = 1, \ldots, p$ is mapped to points on a circle between $-\pi$ and π.

The circumplex structure arises in a variety of contexts. In psychology, examples occur in personality, social psychology, and perception (Plutchik and Conte, 1997). LaForge and Suczek (1955) found this structure in student's ratings of other's personalties. The psychopharmic (pharmopsychic?) Timothy Leary popularized a variant of this scale as the Leary Interpersonal Circle (Leary, 1957).

Other psychological phenomena have been found to be best fit by a circumplex model: emotions (Ekman and Friesen, 1978; Benjamin, 1979; Russell, 1980), interpersonal traits (Schaefer, 1959; Wiggins, 1979, 1982, 1996), color perception (Ekman, 1954), and pitch perception (Shepard, 1964). Browne (1977, 1992) and Shepard (1978) describe other circumplex models in psychology. In time and spatial models, the **circular serial correlation coefficient** measures circular dependence. Olkin and Press (1969) discuss a circular moving average model.



On left is actual CIE chromaticity diagram. On right, MDS of Ekman (1954) data.

Equicorrelation

The data values in the fourth dataset (Equi) were generated from the formula

$x_{ij} = t + u_{ij}$,

where

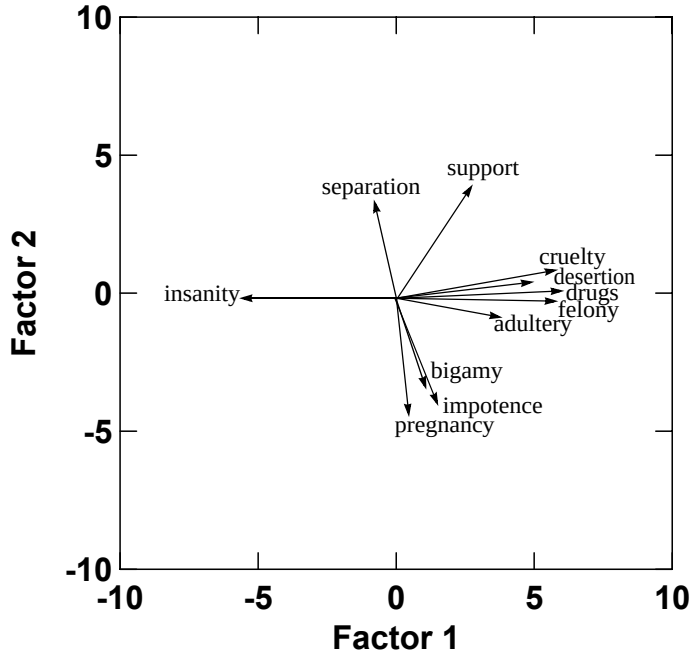
$t = bi/n$,

and u_{ij} is sampled from a standard normal random variable. The correlations in this dataset are relatively large and positive ($\bar{r} = 0.65, \quad sd_r = 0.05$).

The equi-correlation pattern is found in single-factor datasets. That is, data depending on a single factor have an equi-correlation matrix. Single-factor datasets occur when there are multiple measurements of a single trait, all with common reliability. Perhaps the most famous single-factor model is the **general intelligence** factor called *g*, originally proposed by Spearman (1904). Spearman analyzed tests involving cognitive abilities and found that all the correlations among these tests were positive. Spearman mistakenly concluded that this "positive manifold" of correlations was evidence of a single underlying intelligence factor. It was a mistaken conclusion because there are ways to generate equi-correlations other than a single factor. Even intercorrelations above .9 do not prove that a set of variates measure (or are caused by) the same thing.

Block

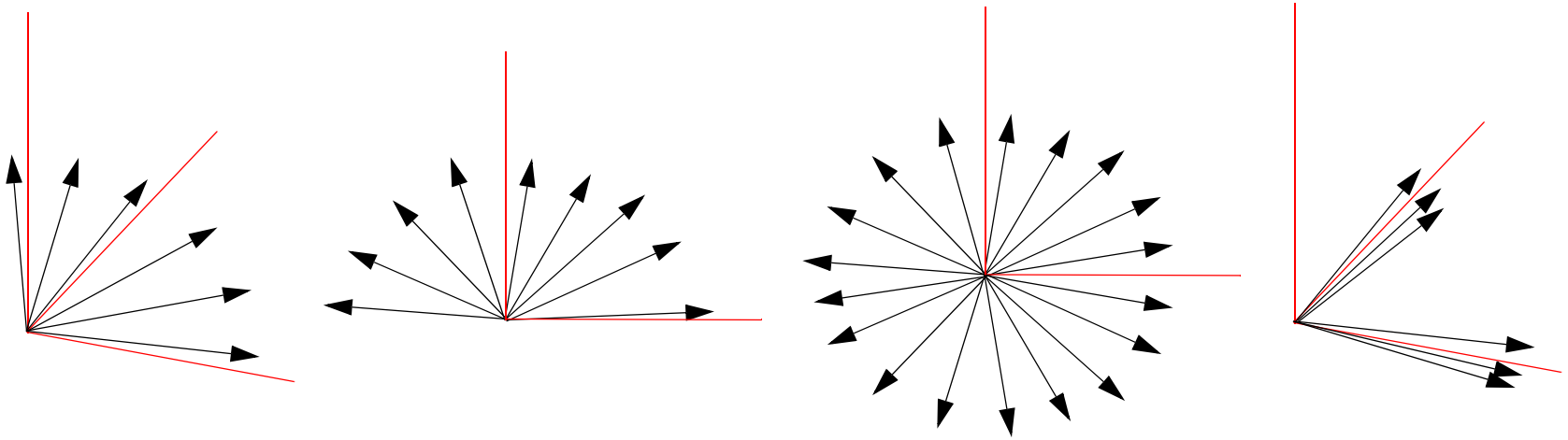
The data values in the fifth dataset (Block) were generated by a formula that constructed row-wise blocks in a bitwise pattern (00, 01, 10, 11). If we reverse the signs of Separation, Support, and Insanity in the following grounds-for-divorce-in-US dataset, then we have a two-factor pattern similar to Block.



Sections of microarray data appear to fit this model.

A look at the column vectors in row space

Simplex Band Circumplex Block



Because **X** is centered and standardized in these examples, cosines of angles between vectors are correlations. These vectors are all near one-dimensional manifolds. Simplex spirals into *n-1* dimensions, Band and Circumplex lie near a semicircle and circle, respectively. Equi lies near a cone and Block lies near two cones.

How do we permute?

Popular methods are:

- Hierarchical Cluster analysis
- SVD (Principal Components)
- Multidimensional scaling
- Sorting on margins

Possible other methods include local image-processing approaches:

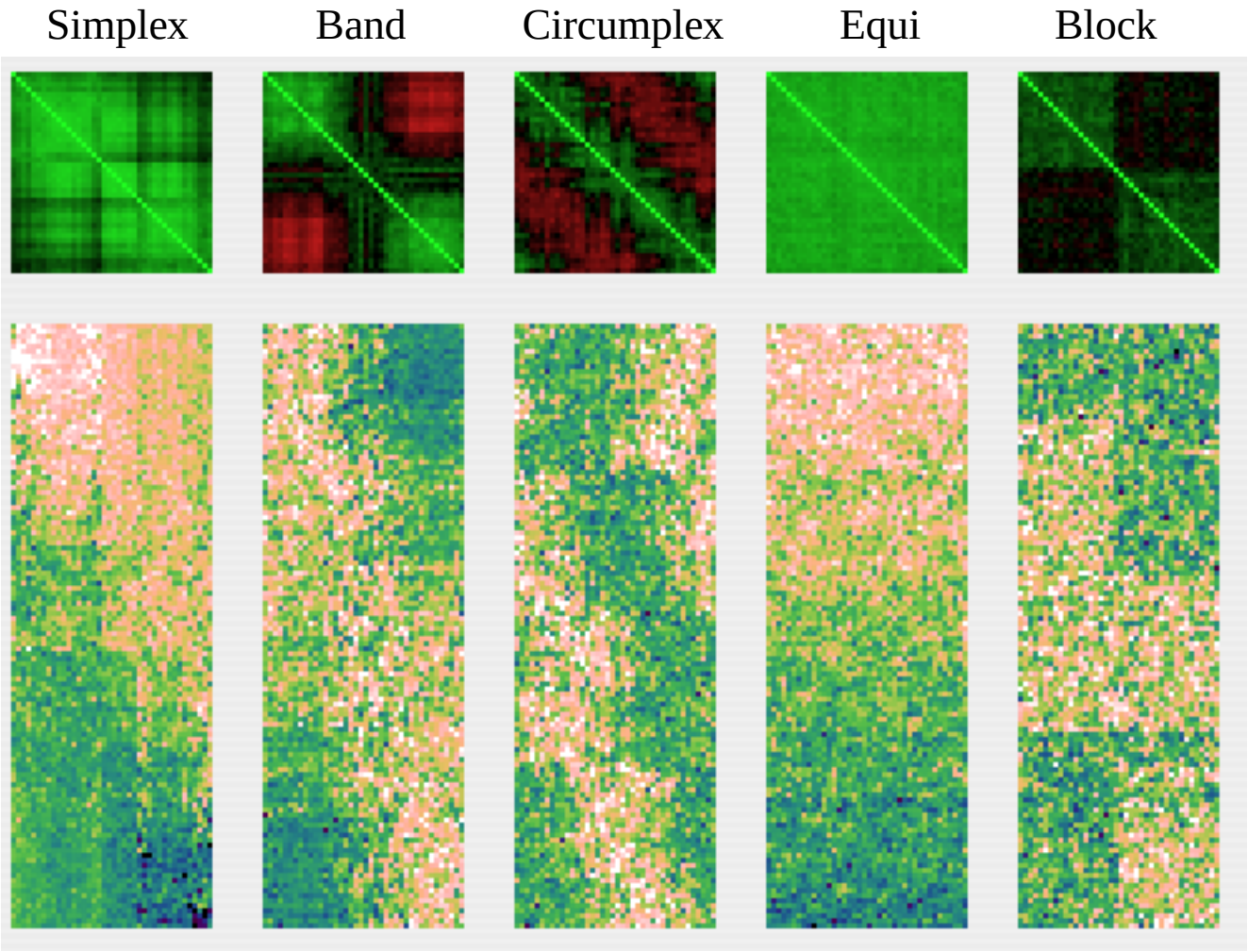
- Loss function based on spectrum from FFT (large ratio of low frequency to high).
- Low-pass filter and compare smoothed pixel value to raw value.

These are NP-hard, but can be ameliorated with initial approximation from popular methods followed by annealing.

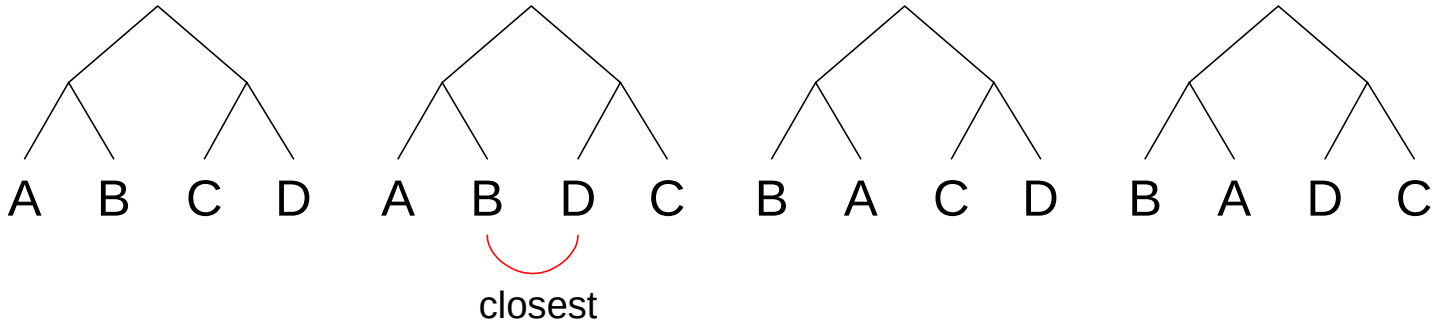
Finally (surprise, surprise),

- Decomposition of normalized Laplacian matrix derived from embedded graph.

Average Linkage Cluster result



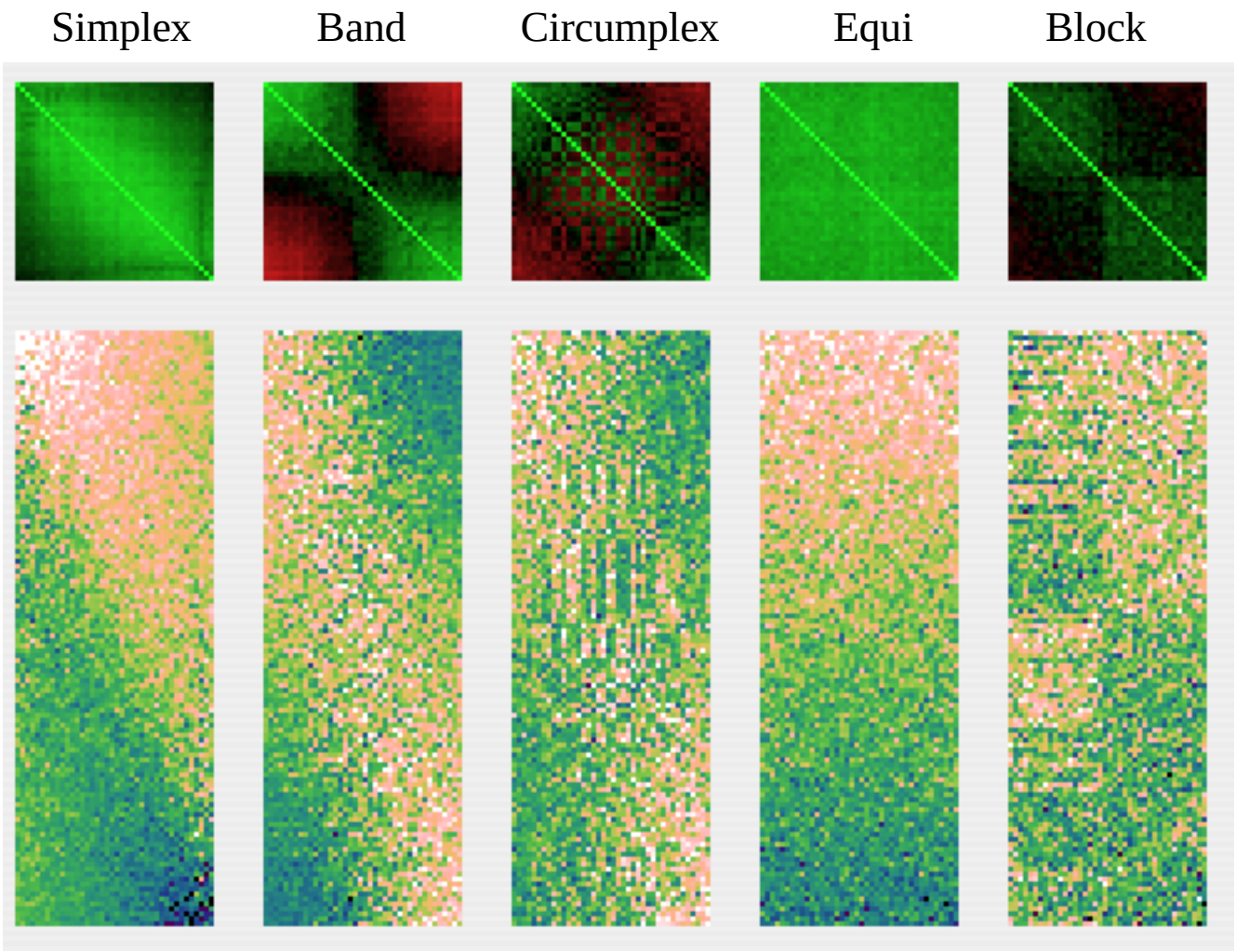
How do we order leaves of tree?



Compute four possible reflections of children of binary node. Pick ordering with smallest interior edge distance based on data. Reorder leaves after every join. Result is equivalent to a topological sort for a DAG (Gruvaeus and Wainer, 1972).

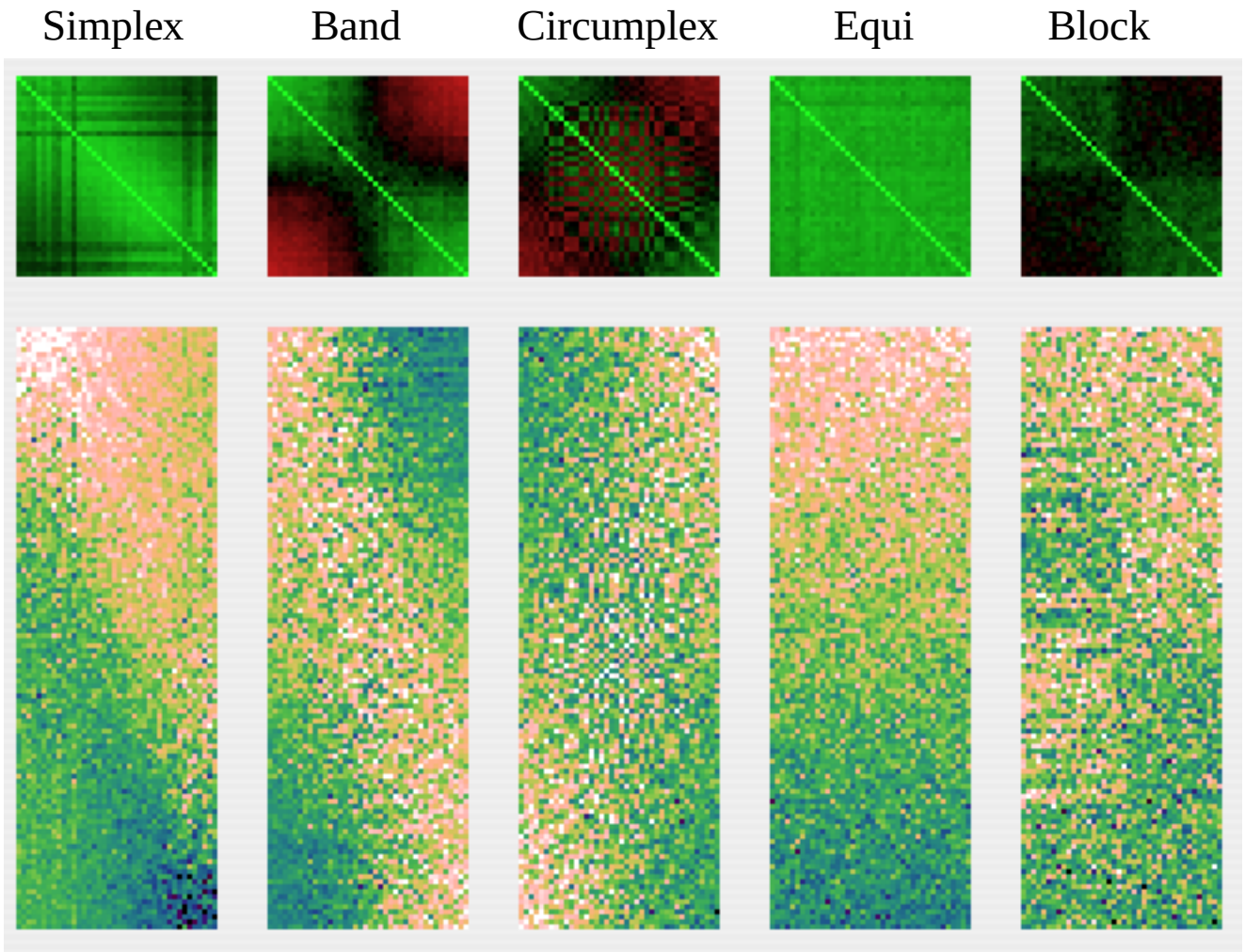
I suspect most heatmap cluster programs do not sort, so row/col dimensions are not globally interpretable.

MDS result



MDS does quite well. It folds circumplex into semicircle, however.

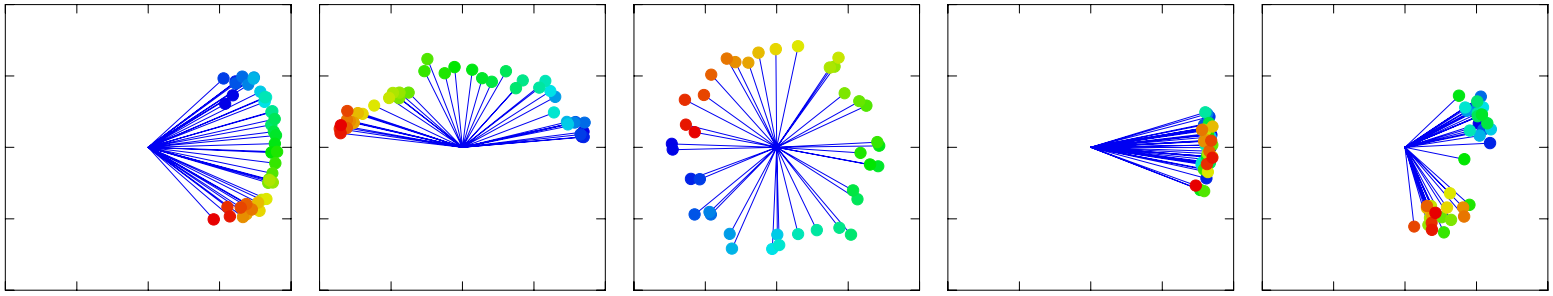
SVD result



SVD does well for band. Folds circle into semicircle, like MDS. Fairly good for block.

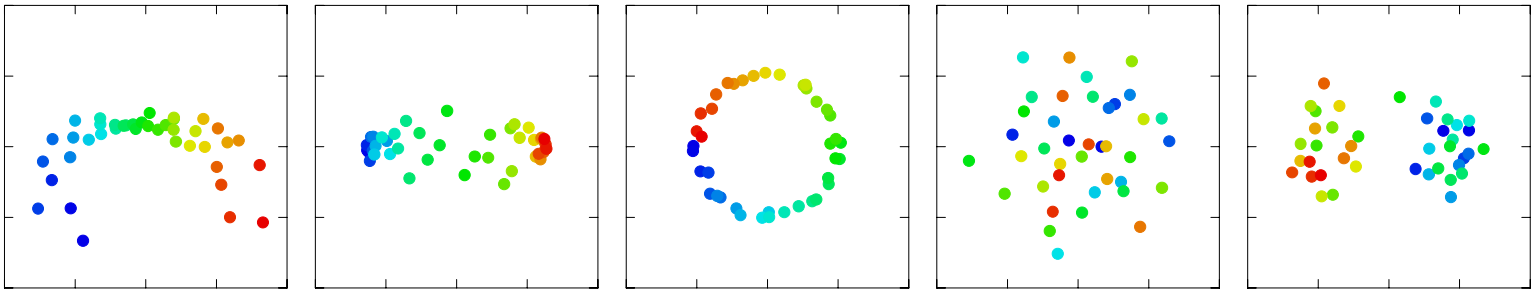
First two dimensions of SVD and MDS show how they perform:

Simplex Band Circumplex Equi Block



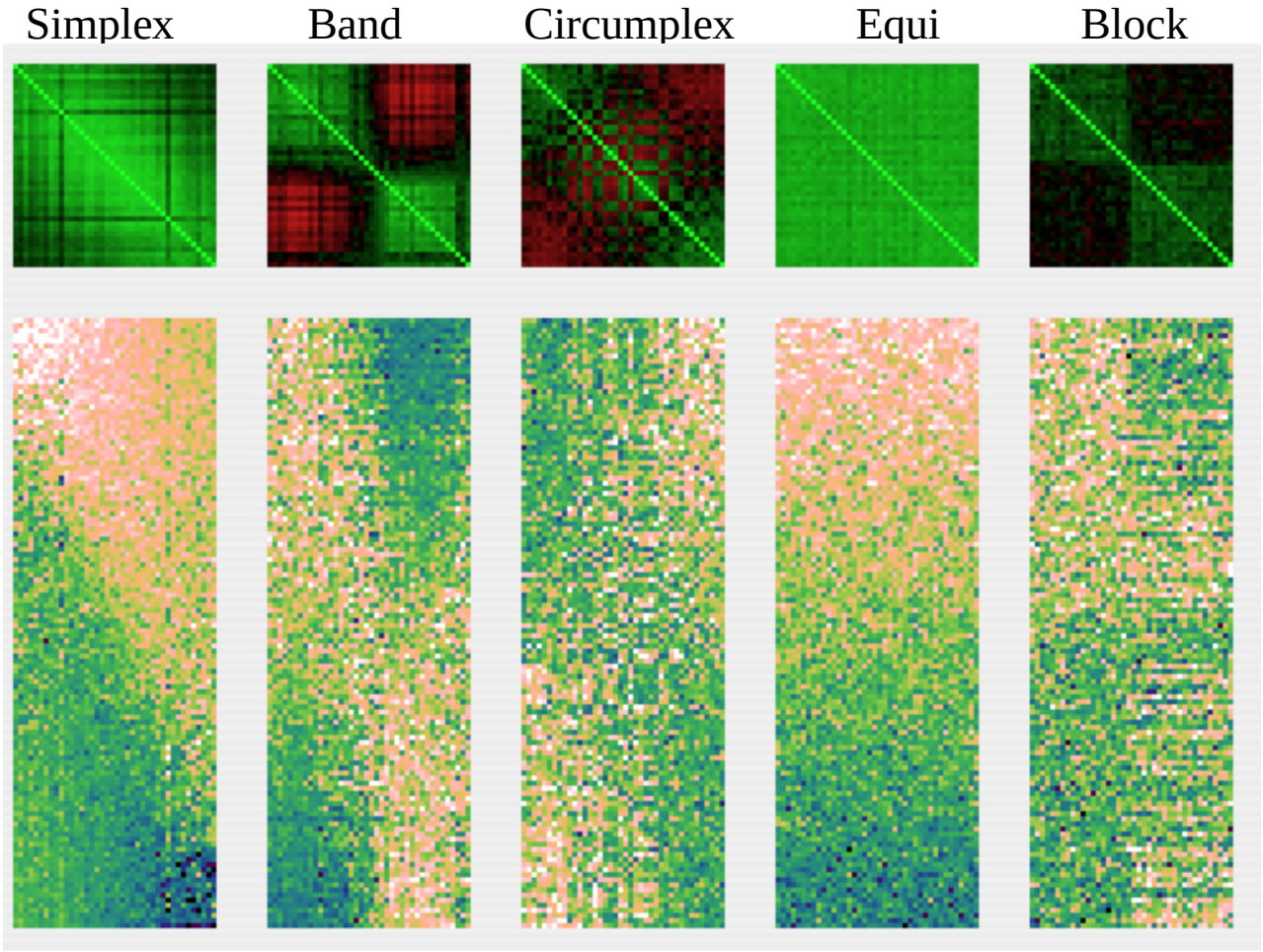
SVD

Simplex Band Circumplex Equi Block



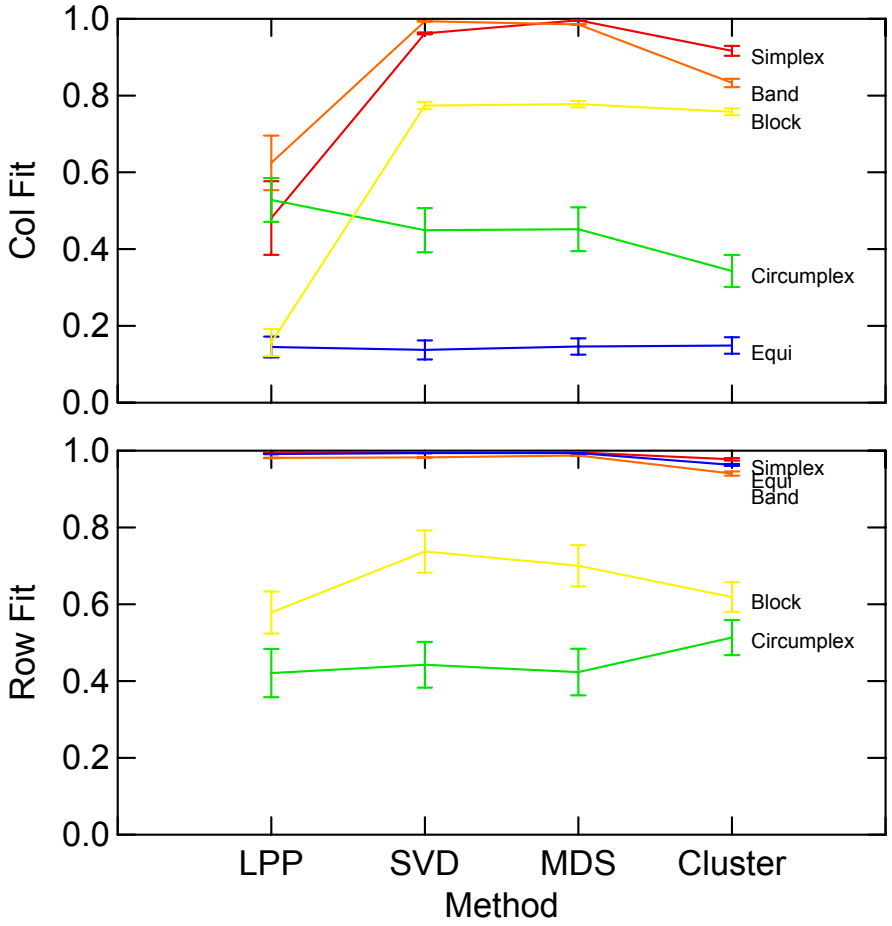
MDS

LPP (He and Niyogi, 2002) result



Locally linear mappings are relatively effective if data lie on manifolds, such as Simplex and Band. This implementation used k=3 nearest neighbor graph with “simple-minded” weights.

Overall performance of methods based on 20 runs



Circumplex is difficult to fit because it needs to be unwrapped. To do this, we would have to know we have a closed circle. Equi column order is random. LPP does poorly with Block because points do not lie on a manifold. Otherwise, SVD and MDS do best.

Conclusions

Pictures are not worth 1,000 words.

There is no substitute for prior knowledge of ordering.

There is no magic permuter for all data because the heatmap space works well only for 1D manifolds (unless we partition). And we need to handle closed circles).

Scaling is thorny.

High-dimensional parallel coordinate plots, scatterplot matrices, Andrews Fourier plots, and similar displays are affected by similar problems. Without sorting, they are difficult to interpret.