Finding Needles in Haystacks:

Tools for Finding Structure in Large Datasets

Brian D. Ripley

Visualization

Challenge is to explore data in more than two or perhaps three dimensions.

via projections

Principal components is the most obvious technique: kD projection of data with largest variance matrix (in several senses). Usually ‘shear’ the view to give uncorrelated axes.

Lots of other projections looking for ‘interesting’ views, for example groupings, outliers, clumping. Known as (exploratory) *projection pursuit*.

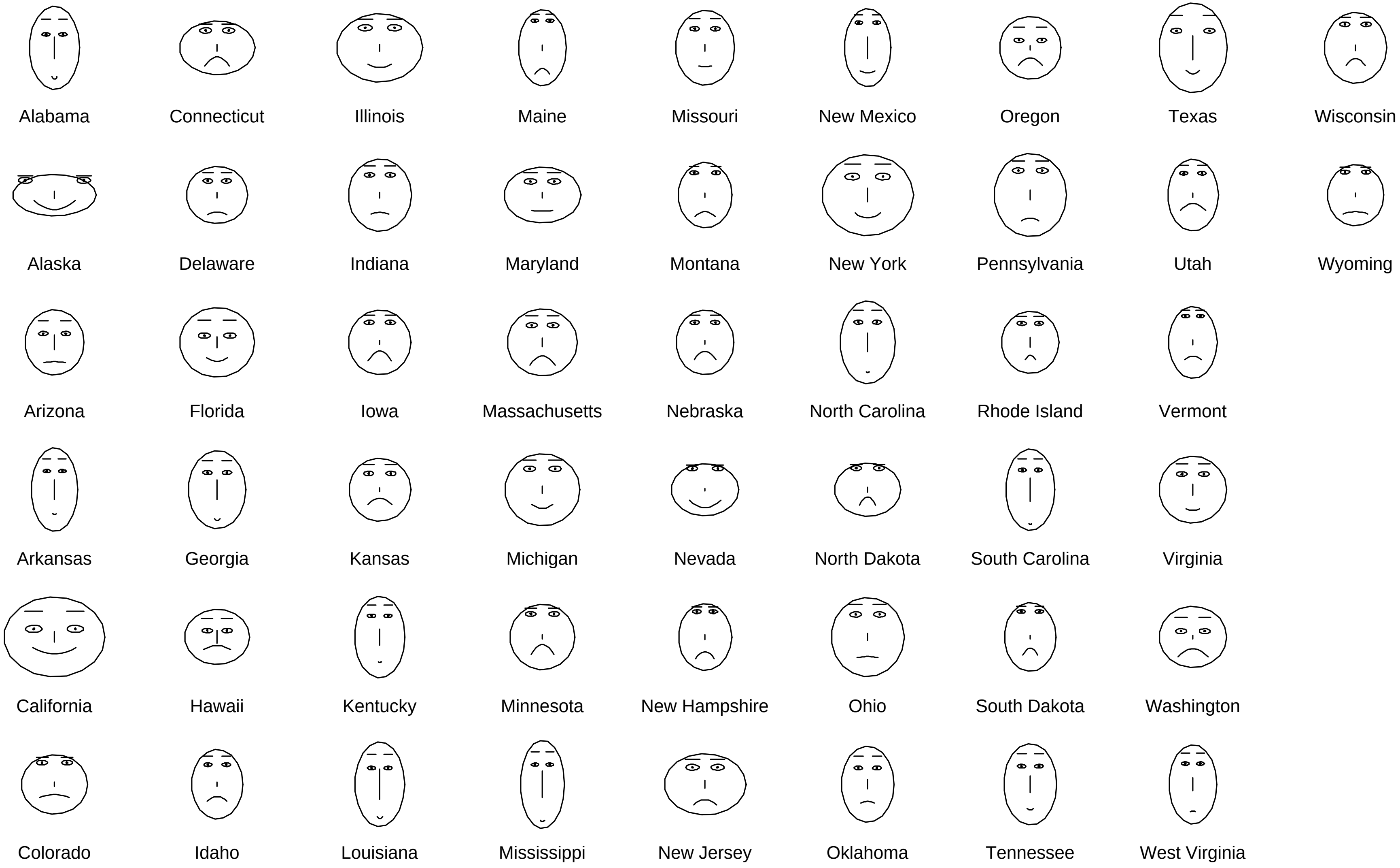
Implementation via numerical brute-force: freely available in GGobi.
‘Random’ searching (so-called *grand tours*) are not viable even in 5D.

Glyph representations

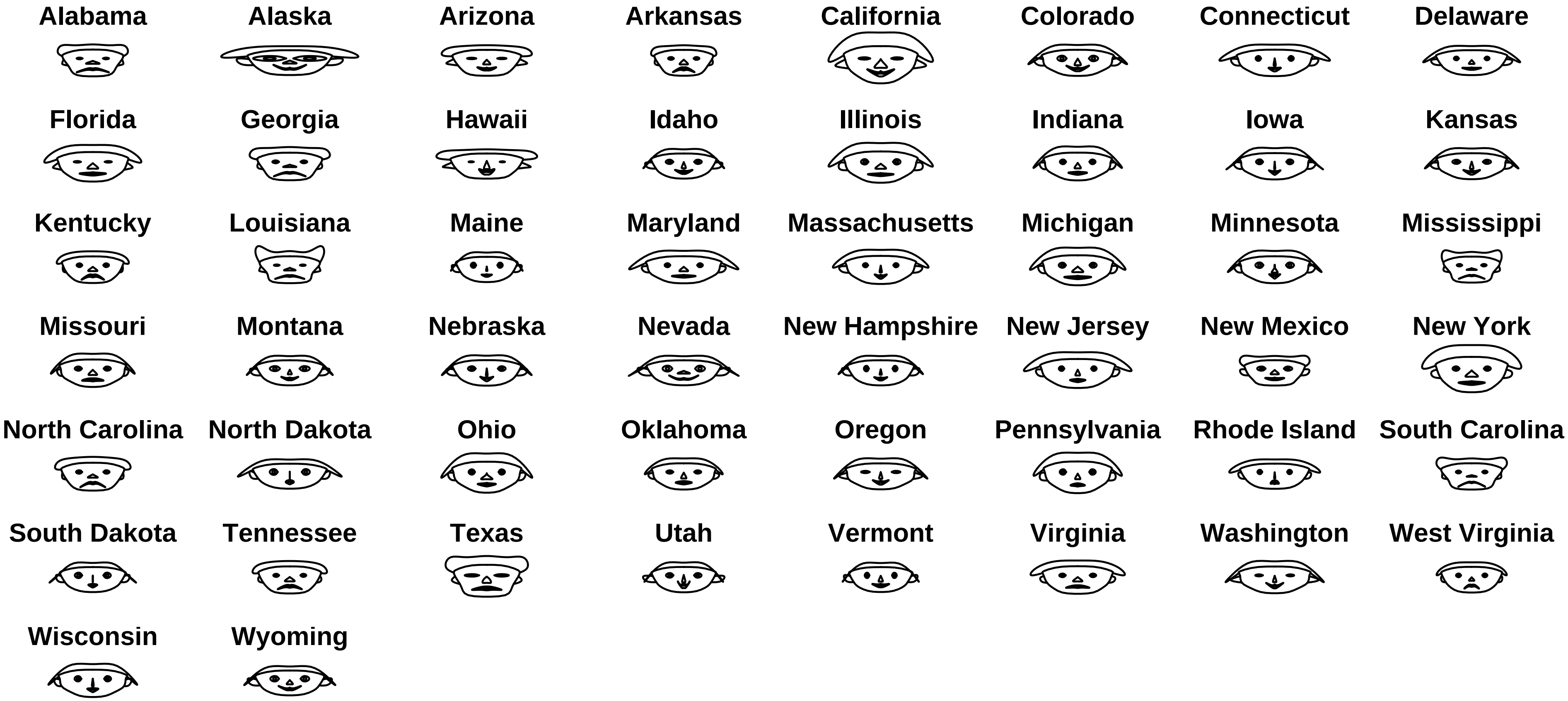
There are many ways to represent each case by a small diagram, of which Chernoff's faces are the most (in)famous.

Wilkinson, L. (2005) *The Grammar of Graphics.* Second ed. Springer.

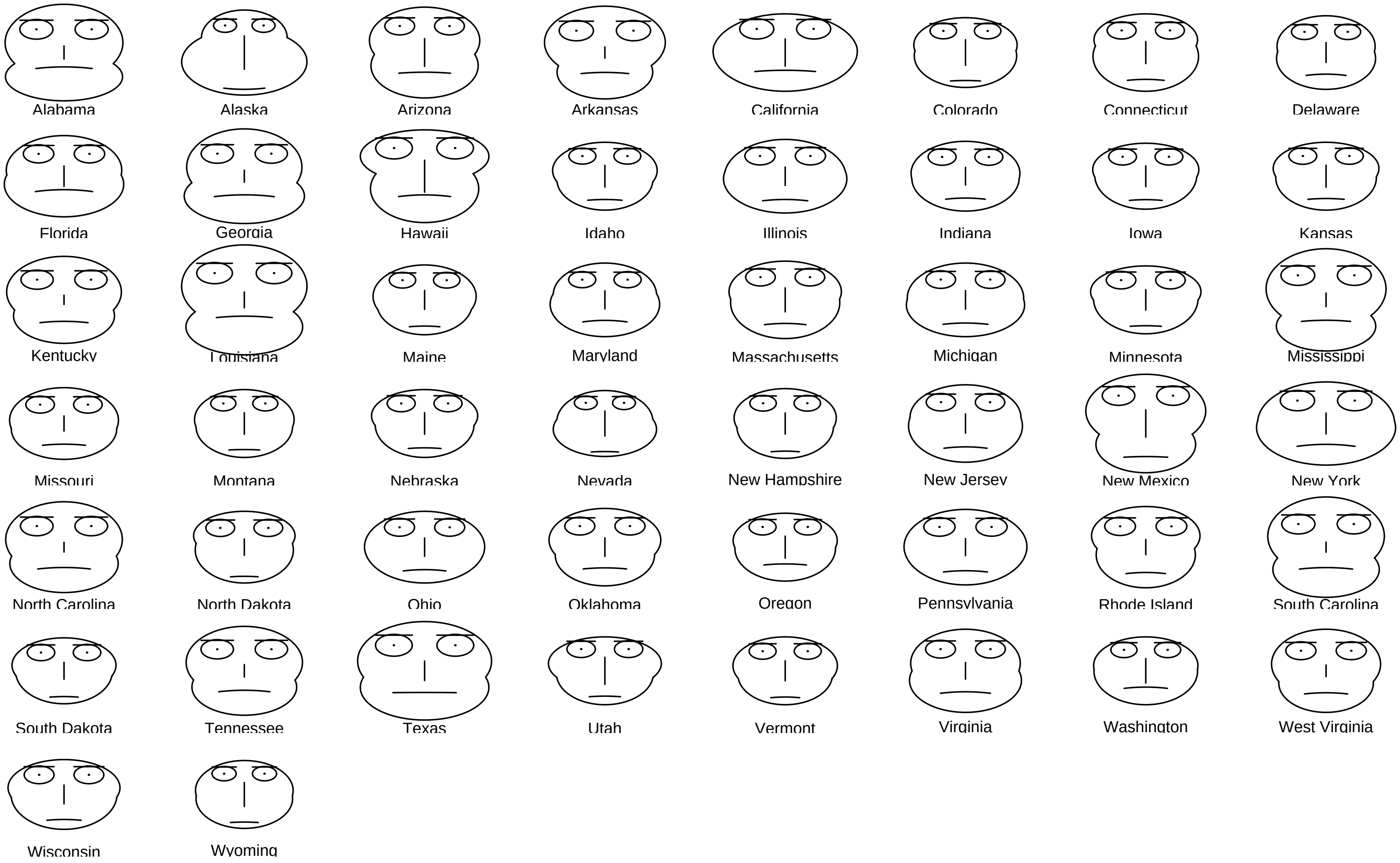
These glyph plots do depend on the ordering of variables and perhaps also their scaling, and they do rely on properties of human visual perception. So they have rightly been criticised as subject to manipulation, and one should be aware of the possibility that the effect may differ by viewer. (Especially if colour is involved; it is amazingly common to overlook the prevalence of red–green colour blindness.)



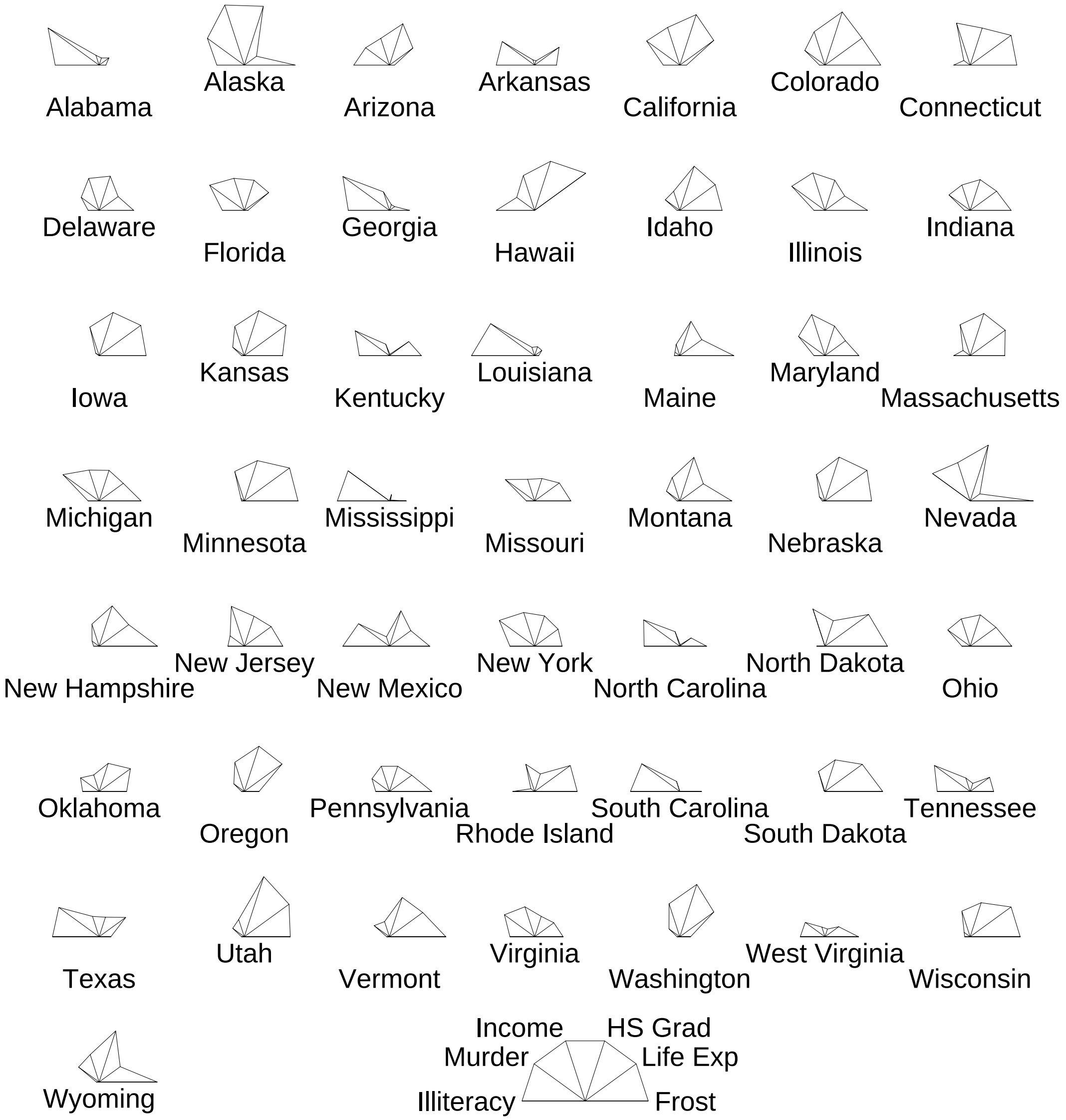
Chernoff faces plot of the `state.x77` dataset, from S-PLUS.



Chernoff faces plot of the `state.x77` dataset, from `TeachingDemos` package (`faces`).



Chernoff faces plot of the `state.x77` dataset, from `TeachingDemos` package (`faces2`).



R version of stars plot of the `state.x77` dataset.

Leptograpsus variegatus Crabs

200 crabs from Western Australia. Two colour forms, blue and orange; collected 50 of each form of each sex. Are the colour forms species?

Measurements of carapace (shell) length `CL` and width `CW`, the size of the frontal lobe `FL`, rear width `RW` and body depth `BD`.

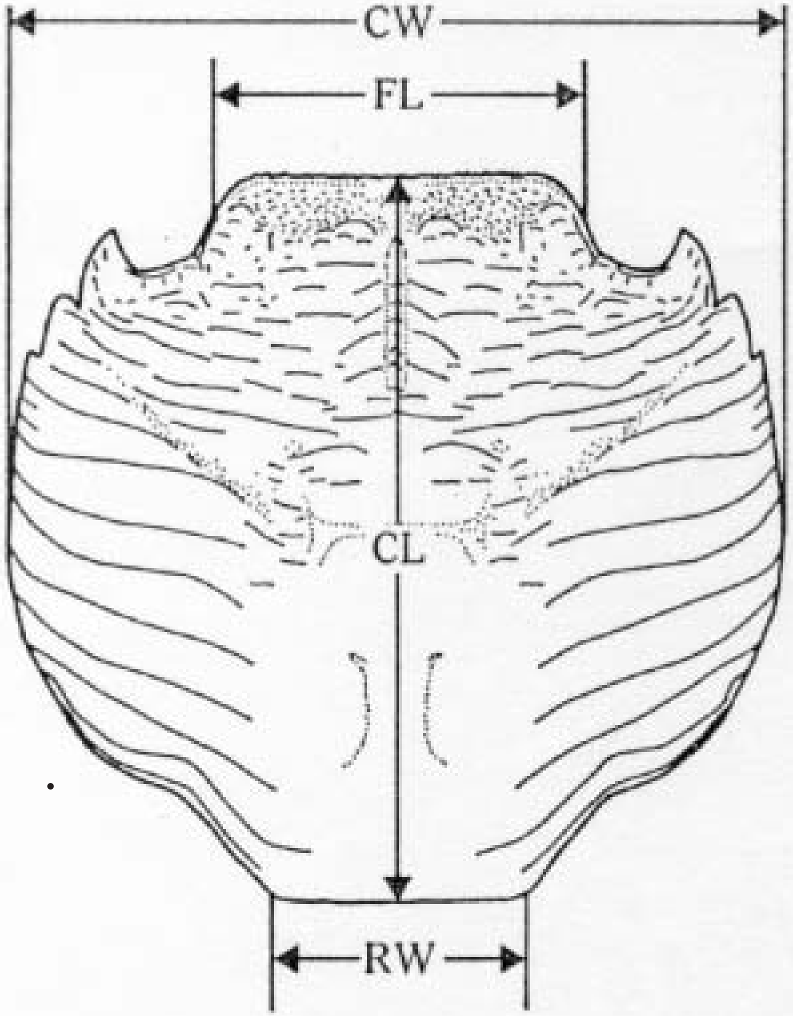
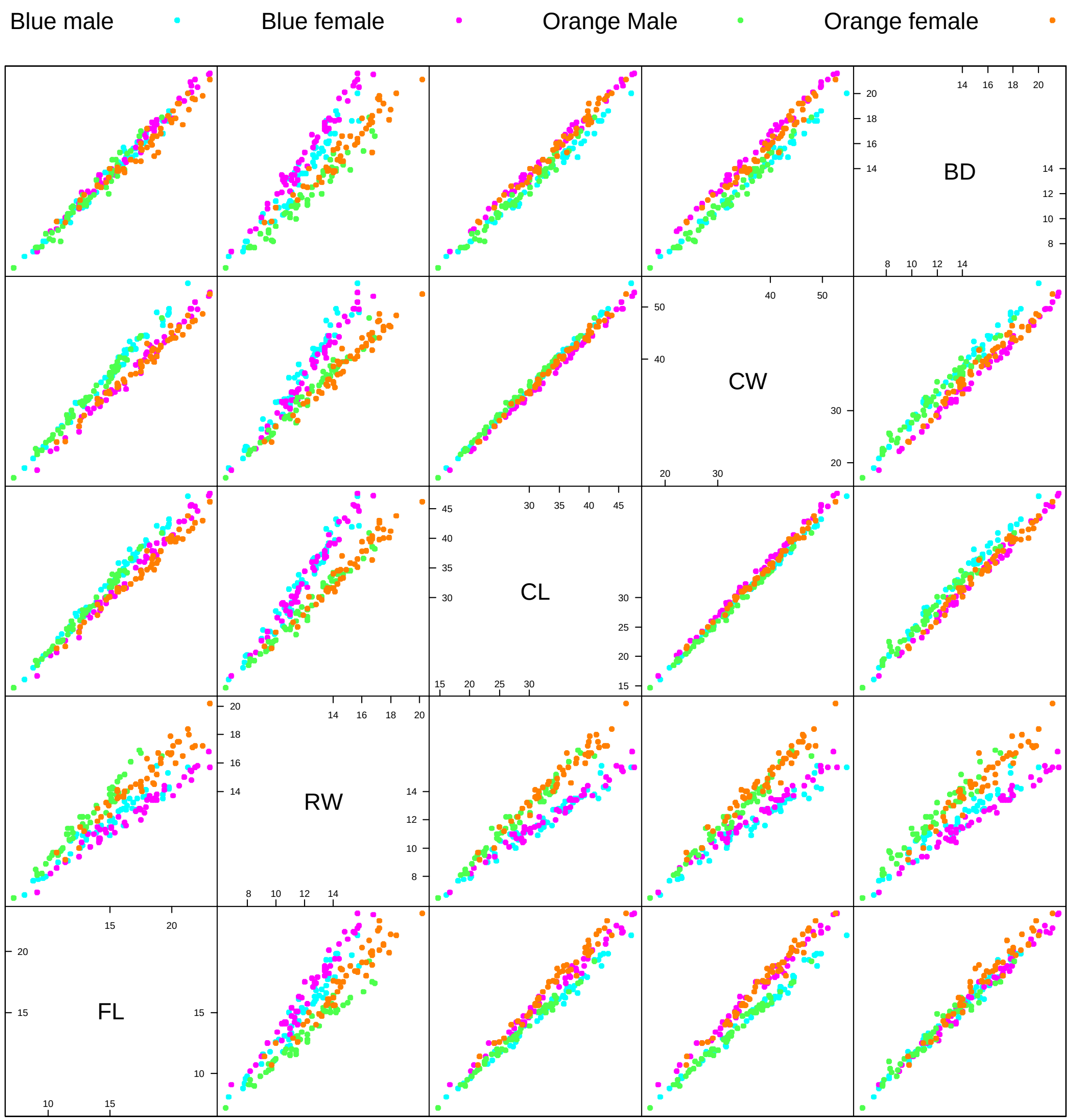
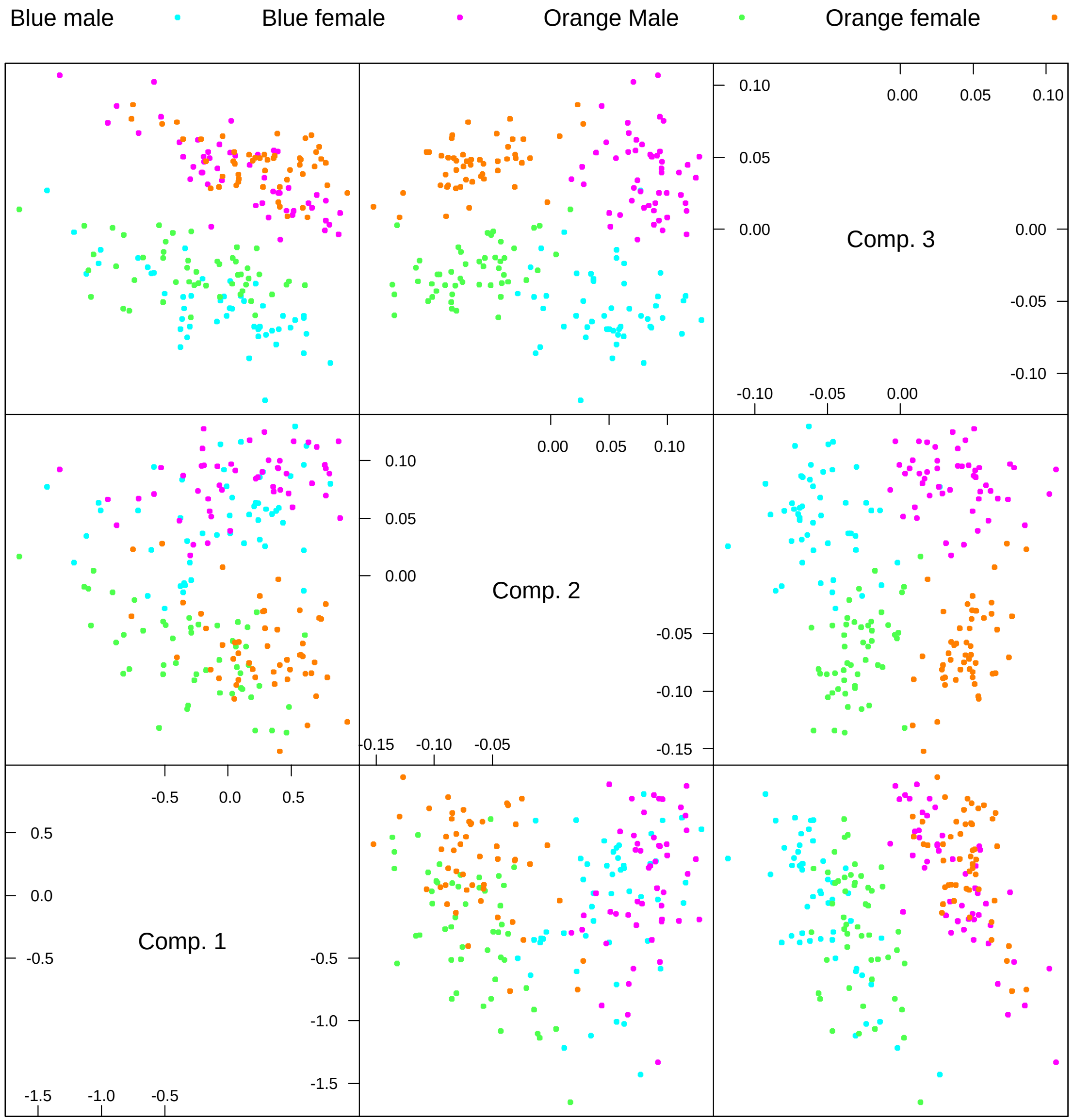
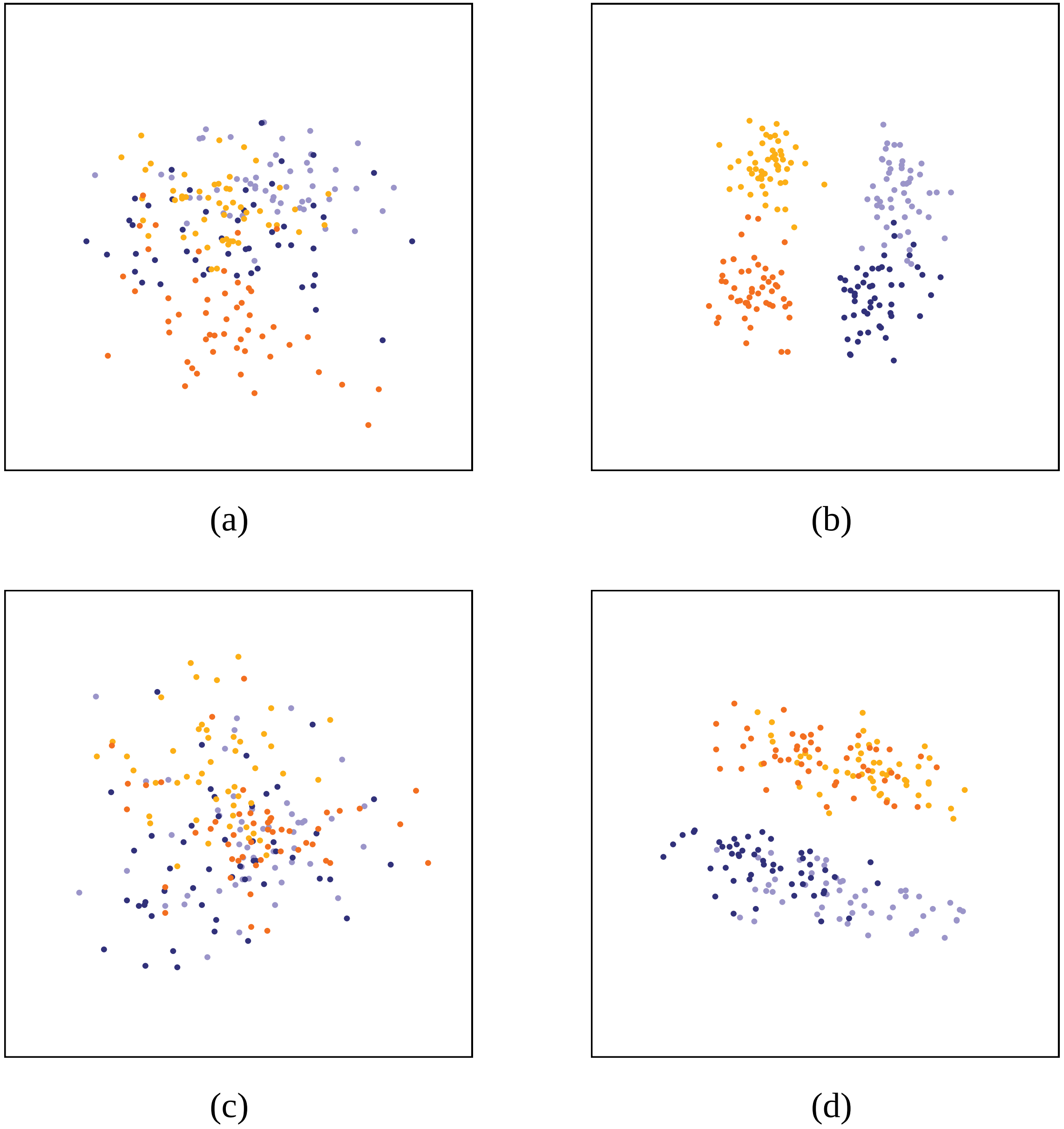


Fig. 1. Dorsal view of carapace of *Leptograpsus*, showing measurements taken. *FL*, width of frontal region just anterior to frontal tubercles. *RW*, width of posterior region. *CL*, length along midline. *CW*, maximum width. The body depth was also measured; in females but not in males the abdomen was first displaced.


Blue male ●
Blue female ●
Orange Male ●
Orange female ●
BD
14 16 18 20
20
18
16
14
14
12
10
8
8 10 12 14
CW
40 50
50
40
30
20
20 30
CL
30 35 40 45
45
40
35
30
30
25
20
15
15 20 25 30
RW
14 16 18 20
20
18
16
14
14
12
10
8
8 10 12 14
FL
15 20
20
15
15
10
10 15



First three principal components on log scale.


(a)

(b)

(c)

(d)

Projections of the *Leptograpsus* crabs data found by projection pursuit. View (a) is a random projection. View (b) was found using the natural Hermite index, view (c) by the Friedman–Tukey index and view (d) by Friedman's (1987) index.

Multidimensional Scaling

Aim is to represent distances between points well.

Suppose we have distances (d_{ij}) between all pairs of n points, or a *dissimilarity* matrix. Classical MDS plots the first k principal components, and minimizes

$$\sum_{i \neq j} d_{ij}^2 - \widetilde{d}_{ij}^2$$

where $(\widetilde{d}_{ij})$ are the Euclidean distances in the kD space.

More interested in getting small distances right. Sammon (1969) proposed

$$\min E(d, \widetilde{d}) = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \widetilde{d}_{ij})^2}{d_{ij}}$$

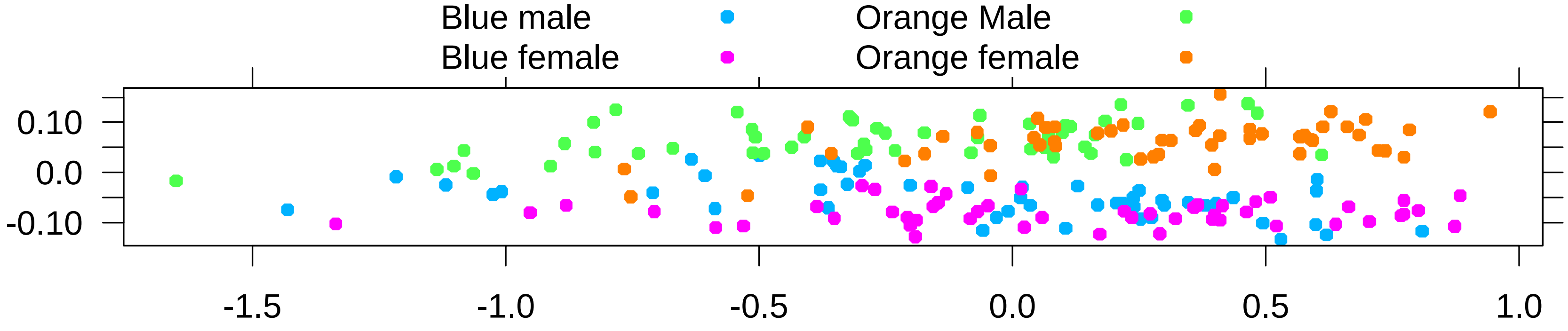
Shepard and Kruskal (1962–4) proposed only to preserve the ordering of distances, minimizing

$$STRESS^2 = \frac{\sum_{i\neq j}\left[\theta(d_{ij}) - \widetilde{d}_{ij}\right]^2}{\sum_{i\neq j}\widetilde{d}_{ij}^2}$$

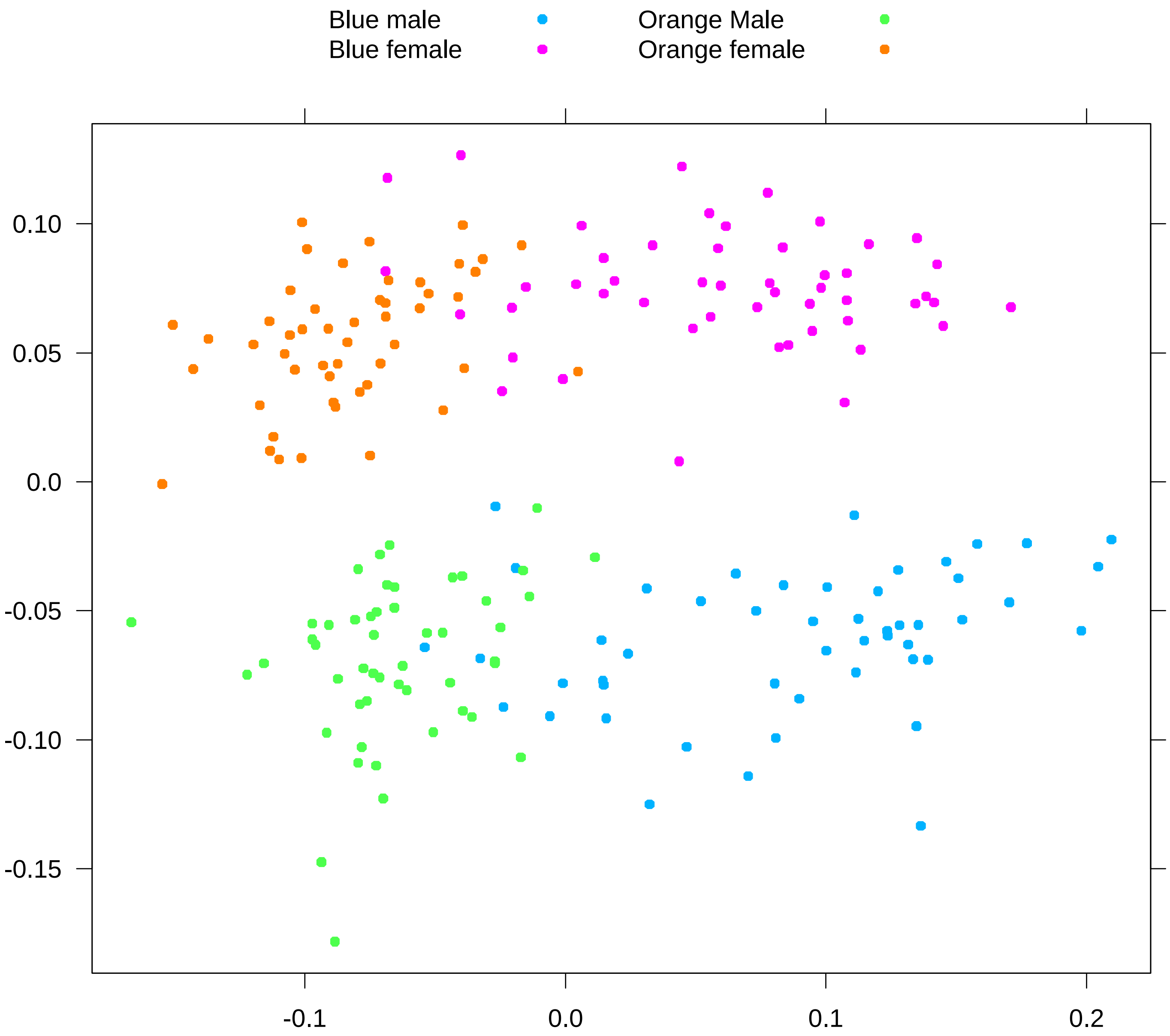
over both the configuration of points and an increasing function θ.

The optimization task is quite difficult and this can be slow.

Multidimensional scaling



An order-preserving MDS plot of the (raw) crabs data.

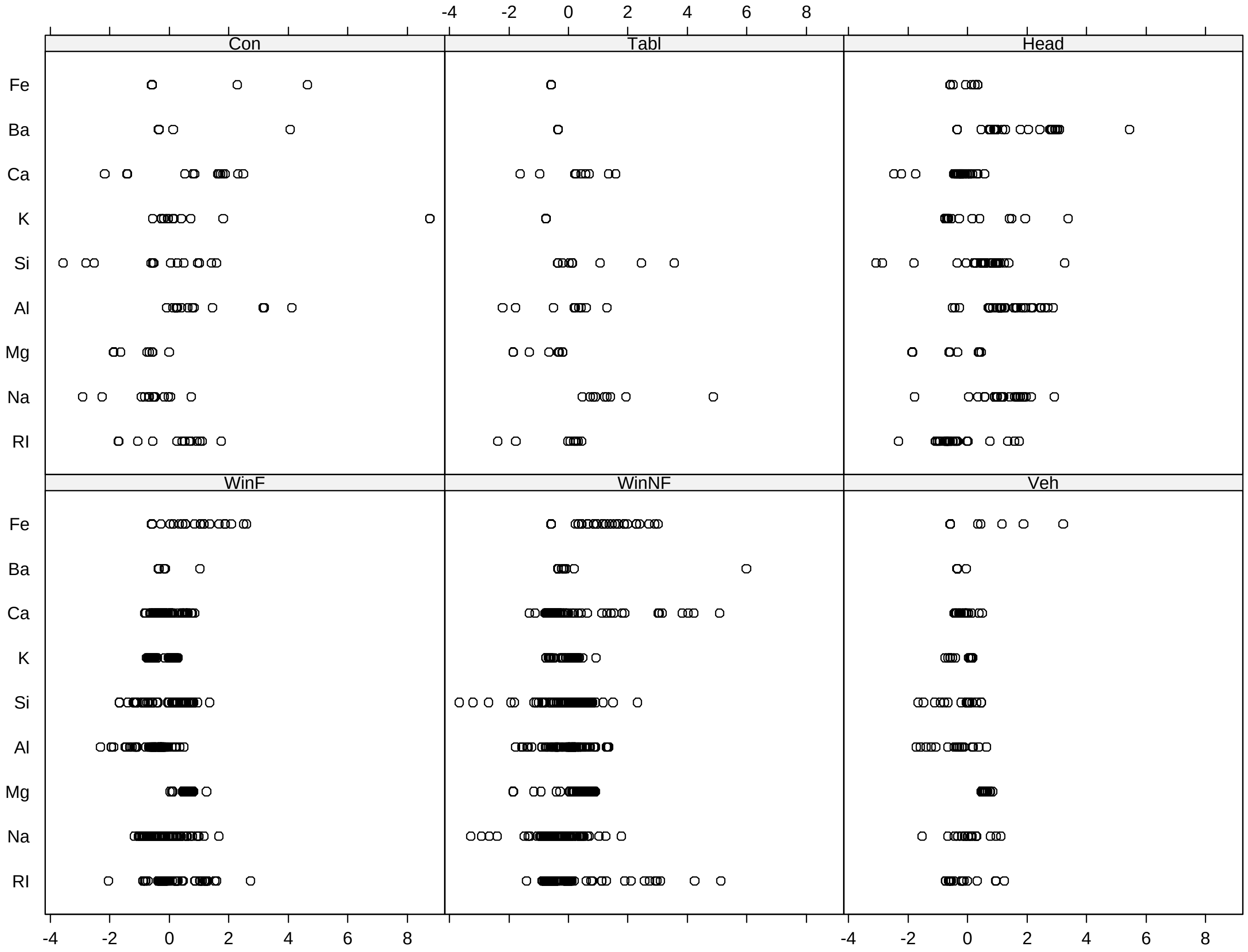


After re-scaling to (approximately) constant carapace area.

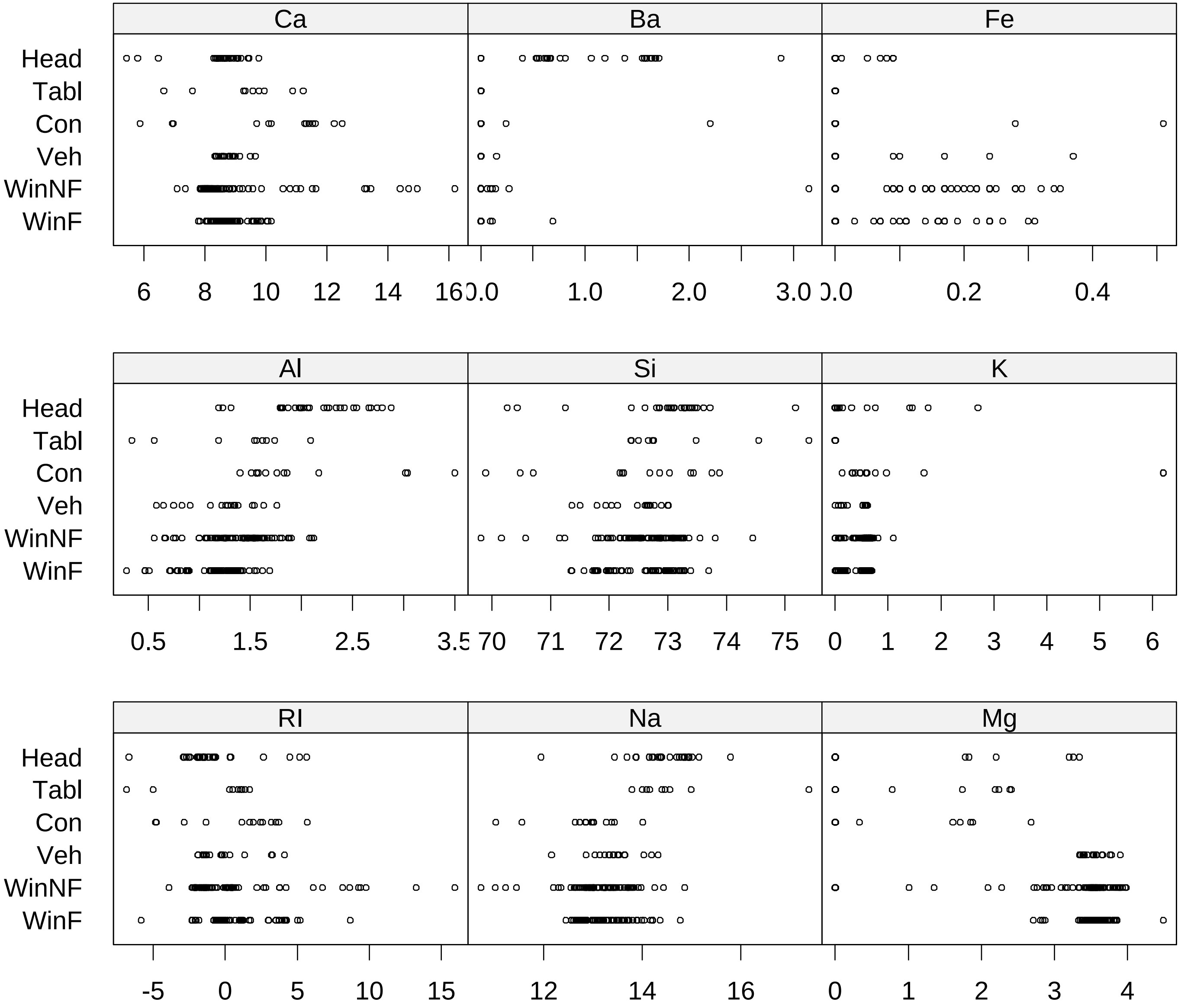
A Forensic Example

Data on 214 fragments of glass collected at scenes of crimes. Each has a measured refractive index and composition (weight percent of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe).

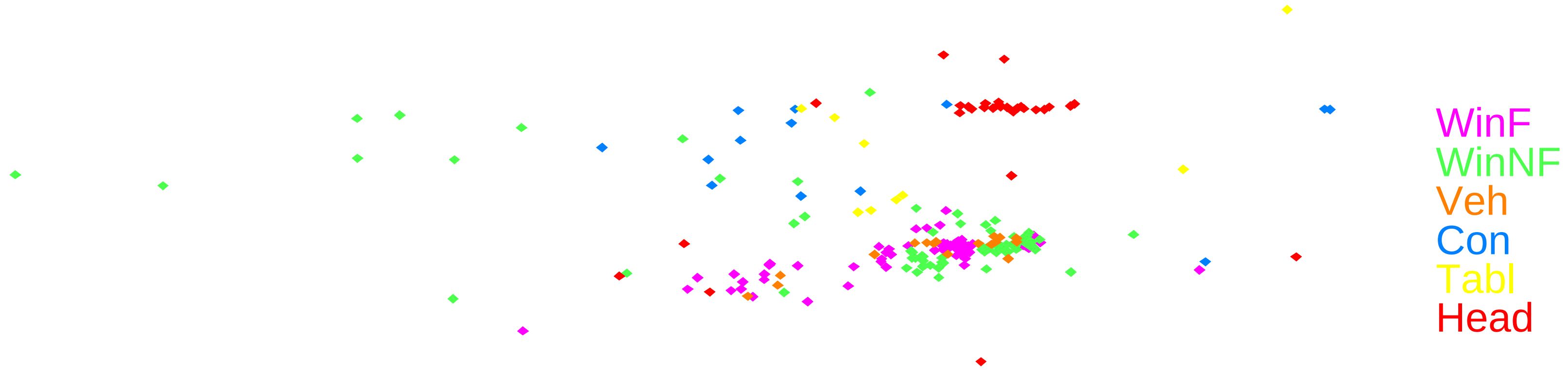
Grouped as window float glass (70), window non-float glass (76), vehicle window glass (17) and other (containers, tableware, headlamps) (22).



Strip plot by type of glass.



Strip plot by type of analyte.



Isotonic multidimensional scaling representation.