Football and the dark side of cluster analysis

(and of exploratory multivariate analysis in general, really)

Christian Hennig and Serhat Akhanli

Department of Statistical Science, UCL
email: c.hennig@ucl.ac.uk,serhat.akhanli.14@ucl.ac.uk

1 A principle for data preprocessing

"The dark side of cluster analysis": clustering and mapping multivariate data are strongly affected by preprocessing decisions such as variable transformations ("data cleaning" belongs to preprocessing but is not treated here). The variety of options is huge and guidance is scant.
The framework here is the design of a dissimilarity measure, used for multidimensional scaling and dissimilarity-based clustering.
Clustering and mapping are unsupervised; decisions cannot be made by optimising cross-validated prediction quality. Neither is it a convincing rationale to transform data to standard distributional shapes such as the Gaussian.

General principle: Data should be preprocessed in such a way that the resulting effective distance between observations matches how distance is interpreted in the application of interest.

Corollary: Different ways of data preprocessing are not objectively "right" or "wrong"; they implicitly construct different interpretations of the data.

Data driven principles such as optimising stability or "clusterability" are suspicious: can the data decide on their own how they should be interpreted?

2 Overview of decisions

Representation: decisions about how to represent the relevant information in the variables properly; this may involve excluding variables, defining new variables summarising or framing information in better ways, and certain kinds of "interpretation-based" (as opposed to data-based) standardisation.

Transformation: variables should be transformed in such a way that the resulting differences match appropriate "interpretative distances" adapted to the meaning of the variables and the specific application.

Standardisation: variables should be standardised in such a way that a difference in one variable can be traded off against the same difference in another variable when aggregating variables for computing dissimilarities.

Weighting: some variables may be more important/relevant than others - weighting is about appropriately matching the importance of variables.
Mathematically, both standardisation and weighting are multiplications by a constant, but the rationales are quite different.

Variable selection and **dimension reduction** are special cases of representation and weighting.
Defining indexes summarising information guided by interpretation is an alternative to data-based dimension reduction.

3 Basic ingredients

The framework here is the **construction of a dissimilarity measure** by aggregating variable-wise distances, e.g.,

Gower aggregation (Gower (1971))
$d(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{p} w_i d_i(x_{1i}, x_{2i})$

Euclidean aggregation
$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{p} w_i d_i(x_{1i}, x_{2i})^2}$.

Alternatives exist but are not treated here.

Mapping is then done by Kruskal's nonparametric multidimensional scaling (Cox and Cox (1994)) and clustering by "partitioning around medoids" (Kaufman and Rousseeuw (1990)).

The general principle above also applies to the choice of mapping and clustering method, but not treated here.

4 Football players dataset

Football players characterised by 125 variables taken from whoscored.com (have > 2000 players but use only 75 prominent ones for illustration).
Variables:

12 position variables (binary) - indicating where a player can play.

Age, height, weight (ratio scale numbers)

Subjective data: Man of the match, media ratings

Appearance data of player and team, number of appearances, minutes played

Count variables (top level): goals, tackles, shots, passes, fouls, clearances etc.

Count variables (lower level): subdivisions such as shots by body parts, type of pass (long, corner, freekick etc.), successful/failed etc.

Aim: provide mapping and classification of players that can be used by managers and clubs looking for players.

5 Representation

- Standardise count variables by number of minutes played.
- Top level/lower level count variables:

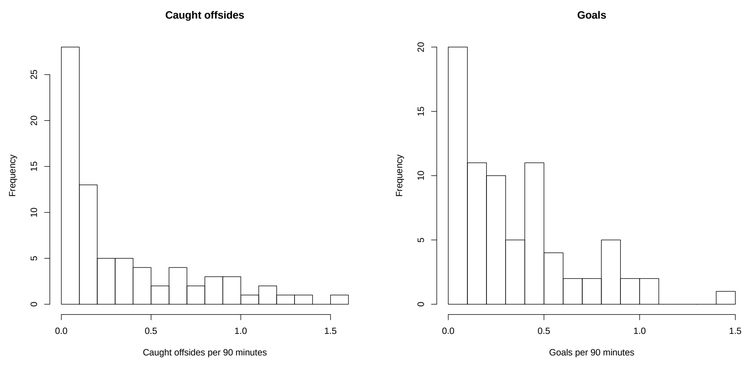
use total count (per 90 minutes),
use percentages on lower level, e.g.,
shots 5.5, shots right foot 3.8, left foot 0.8, header 0.9,
use shots 5.5,
percentages right foot 69, left foot 14, header 16
percentage profiles are complementary information to total.

- For binary position data use "geco coefficient" (Hennig and Hausdorf (2006)) based on aggregating "geographical distance" for every position to closest position of other player.
- We decided to not use subjective variables.

It'd be legitimate to use them - this is a decision about what meaning the results should have.

6 Transformation

Some variables are very skewly distributed.



More variation at upper end, suggests that "interpretative distance" between large values should be transformed down.
But goals count (approximately) linearly in football.
Use concave transformation for "Caught offsides"
but none for "Goals".
Decide $\log(x + c)$ or $\sqrt{x + c}$, value of c by looking at what it does to the values and what seems appropriate (subjective football expertise). Explore by *sensitivity analysis* what difference it makes. (Use plain square root here.)

Transformations: data dependent?
Unfortunately, researchers have no clear formal idea about "interpretative distance". Rationale of transformation is matching "interpretative distance" independently of the data. But researchers may need to look at data for having clearer idea about "interpretative distance".

7 Standardisation

Percentage variables, player age, goals, passes per 90 minutes don't have compatible variation. Standardisation is needed.
But different percentages at same level (shots left, right, header) should be standardised by pooled variance, because variations are compatible and relative sizes should be preserved. Bigger variation should have bigger implicit weight.

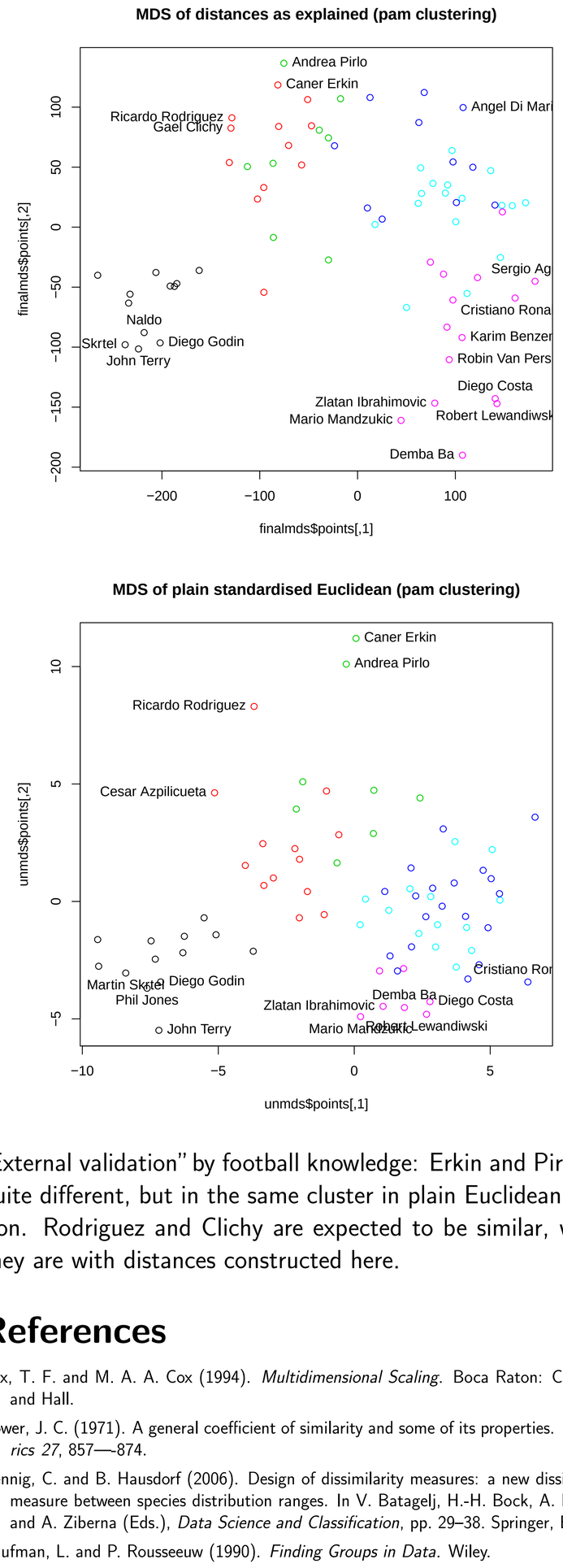
Standardisation should not destroy implicit weighting by variance, where appropriate.

8 Weighting

Variable weights are useful if some variables seem more important than others. This influences the meaning of the results.
Weighted percentage distributions as "one variable", e.g., left foot, right foot, header shots each weighted 1/3.

Correlation, shared information: if variables are correlated because of *redundant* information (e.g., percentages adding to 100), weight them down.
If variables with *complementary meanings* are correlated, no reason not to give them full weight.

9 Results



"External validation" by football knowledge: Erkin and Pirlo are quite different, but in the same cluster in plain Euclidean solution. Rodriguez and Clichy are expected to be similar, which they are with distances constructed here.

References

Cox, T. F. and M. A. A. Cox (1994). *Multidimensional Scaling*. Boca Raton: Chapman and Hall.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 857—874.

Hennig, C. and B. Hausdorf (2006). Design of dissimilarity measures: a new dissimilarity measure between species distribution ranges. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna (Eds.), *Data Science and Classification*, pp. 29–38. Springer, Berlin.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data*. Wiley.