Detecting homogenous subsets of a population.

Bruce Oddson
School of Human Kinetics
Laurentian University

Goal

So what I would really like to do is to present a method that can detect subpopulations that might be interesting, even though they might have means on observable variables that don't distinguish them from the rest of the population

Group differences


Some groups are very easy to tell apart

Group differences

Other times it can be very difficult to decide if our data represent a continuum or if there are separate groups to be observed.

Homogenous groups

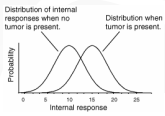
Sometimes we are interesting in really specific groups that can hide within a larger population:

-people with special risk factors for disease
-people with underlying genetic differences
-people who might respond to different treatments

Homogenous subsets

But these groups can be impossible to detect inside a variable population.

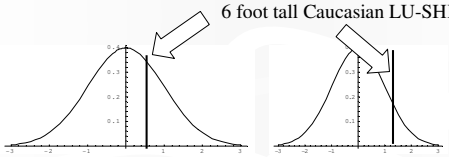
Variance based methods



The only thing that can define a group under classical statistics is a difference of means.

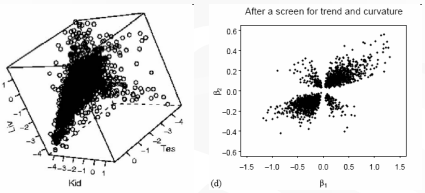
Variance based methods

6 foot tall Caucasian LU-SHK faculty



The more variable the population, the less it will be possible to detect a subgroup that fits with the "average".

Variance based methods



Here are some samples of things that happen with cluster analysis/ discriminant analysis when groups overlap (plots from Bryan, 2004).

Variance based methods

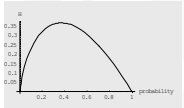
One problem is that the presence of groups may or may not create usefully detectable changes in variation.

Another is that we simply cannot use variance when we wish to link to categorical or ordinal data.

Entropy (a brief introduction)

$$H(x) = \int p \, \mathrm{Log}\, p \, dx$$

Entropy is similar to variance; the less predictable a set of scores, the higher the entropy.



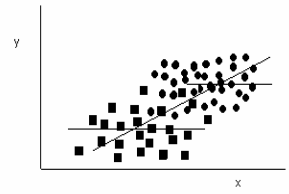
Entropy

The advantages of working with information/entropy:

a) it applies to anything we can count
– e.g. categorical, boolean genetic models

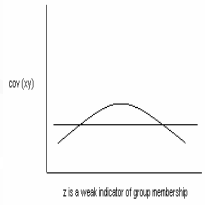
b) it can be used to create equivalent statistical tests to most GLM estimations
(e.g. McGill, 1958; KullBack, 1959).

Max-Cov



$$cov(xy) = p\ cov_1(xy) + q\ cov_2(xy) + pq\ (x_1-x_2)(y_1-y_2)$$

Max-Cov



If S_k (S_4) of cov(xy) > 0 then z indicates group membership

* Not necessary condition
* Not sufficient!

But potentially very useful!

Extension

Not all patterns of interest will have such a clear picture - it will depend on the underlying distributions of each variable and the extent to which group membership can be identified with a particular statistics.

Extension

If the observed entropy of any joint distribution of variables partitioned on a variable not in the set changes as a function of any other variable; then there is information about group structure in that joint pattern of data.

$$H(partition) \neq f(H(overall))$$

If we choose $\sqrt{N}$ partitions on V variables, with C (V) joint statistics on each partition, then this is still much less work then traditional methods.

Extension

Having detected all of the possible variations in joint distribution density, we can then rank them and evaluate each according to specific hypotheses.

Decisions about which variations are of interest will depend largely on the ability to create specific distributional models for each set of observations.

Alternative

A different direction is to use H distance measures with PAM or PAMSIL where:

$$f(\mathrm{M}) = -\sum_j d_1(\mathbf{x}_j\ , \mathrm{M}).$$

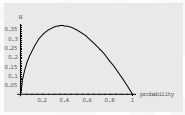
$$S_j(\mathrm{M}) = \frac{b_j - a_j}{\max(a_j, b_j)}.$$

$$a_j = \operatorname{avg} d(\mathbf{x}_j\ , \mathbf{x}_{j'}), j' \in \{i : l_1(\mathbf{x}_i, M) = l_1(\mathbf{x}_j\ , M)\}.$$

$$b_{jk} = \operatorname{avg} d(\mathbf{x}_j\ , \mathbf{x}_{j'}), j' \in \{i : l_1(\mathbf{x}_i, M) = k\}.$$

Alternative

Clustering methods depend on a separation between groups on the distance measure you select.



1. They do not distinguish between continuum & groups
2. H is not a true distance measure and will not always apply.
3. H will also tend to value certain parts of a distribution more than others.

Conclusion

Homogenous subgroups by definition will explain very little variance. This makes them difficult to find using ordinary GLM methods.

Using H as a distance measure can potentially be adapted to group detection methods, but will not be helpful for CA, DA, or related methods.

Next Step

If the H of some partitioned joint distribution is different from the non-partitioned joint distribution

1. We need to be able to assign a probability
2. We need to be able to test directly against models of population distribution to decide if this might reflect a homogenous subset.
3. We would like to link this type of selection model directly to genetic data (PBNs, for example).

Thank you

References

Bryan, J. (2004) Problems in gene clustering based on gene expression data. Journal of Multivariate Analysis, 90, 44-66.

Kullback, S. (1959) Information theory and statistics.

McGill, W. (1954) Multivariate information transmission. Psychometrika, 19, 97-116.