Index of Lecture 11: Classification and Cluster detection

Practical Information

Today's lecture — wrap-up of multivariate methods with classical statistical material on linear discriminant analysis (LDA) and classification, and a brief introduction to cluster detection,

* LDA; Manly 3/4, Chapter 8,
* classification with logistic regression; Manly 3/4, Chapter 8 (but not much new),
* ***K*-nearest neighbor** classification; SL (James et al. (2013) text), Section 2.2,[1]
* introduction to spatial cluster analysis: Sections 26.3.3-26.6.2 (parts) of VER2,

— this will be the LAST MULTIVARIATE LECTURE !

Schedule:

- next week (if we have classes):
  - lab 11−P on Monday,
  - shared lecture 12a−L with VHM 812 on Friday,
- project outline due today (I will return these before the weekend),
- second multivariate home assignment (#5) is due next Thursday.

[1] The SL text also contains nice introductions to more radical novel regression techiques, e.g. support vector machines.

Classification/Discrimination: Overview

Assume p-dimensional observations of the form $\mathbf{X} = (X_1\ X_2\ \ldots X_p)^t$ on observations for which a (true and perfect) classification into g groups (or classes) exist;

Interest is in developing a classification rule based on $\mathbf{X}$ to "predict" group membership, in order to

– classify (probabilistically[2]) future observations,

– obtain insight into any structure involved in group membership (e.g., associations with $\mathbf{X}$-components).

Many approaches exist for such problems; one possible classification. . . :

* parametric, i.e. based on a specific (statistical) model assumptions, e.g. normality,
* partly parametric, i.e. based on assumed forms of probability of group membership, e.g. logistic or polytomous/multinomial regression models,
* distribution-free techniques, e.g. based on specific distance measures or training of a complex classifier.

Classical methods, such as linear discriminant analysis, have a long history, but are increasingly overtaken by computer-intensive methods with a more intuitive and less technical mathematical/statistical foundation.

[2] Prior probabilities for the groups will play into this, via Bayes' formula (Lecture 3, VHM 801).

Logistic Regression as Classification

Example: sparrow data, with 5 quantitative measurements on 49 sparrows after a storm, to predict survival (0/1).

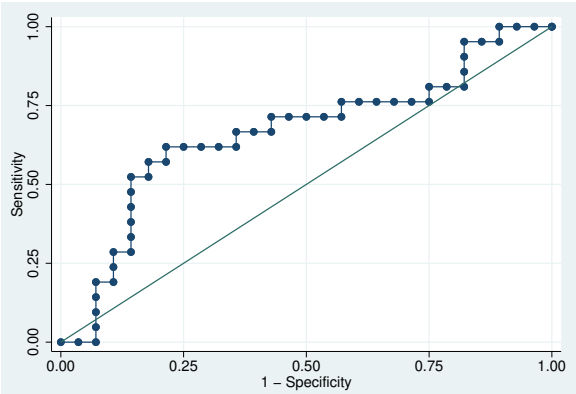
Logistic regression for survival:

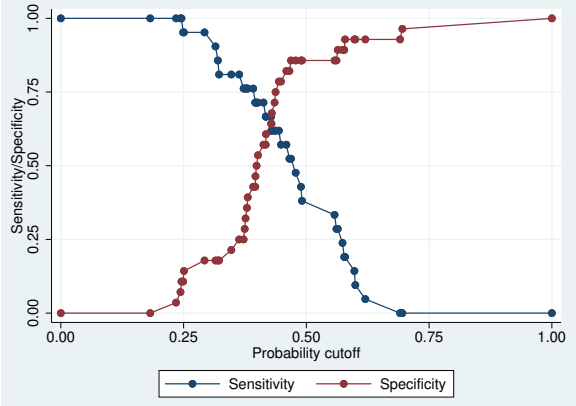
Statistic	Effect/Predictor					
	intercept	length	alar	beak_head	humerous	keel_stern
estimate	13.58	-0.16	-0.03	-0.08	1.06	0.07
SE	15.86	0.14	0.11	0.63	1.02	0.42
P-value	–	0.24	0.79	0.89	0.30	0.86

- overall, non-significant model ($\chi^2(5) = 2.85$),
- classification table (at prob. cut-off 0.5) (later also termed confusion matrix):

true surv	predicted survival 0	1	total
0	24	4	28
1	14	7	21
total	38	11	49

- ROC curve: (ROC area = 0.66 — very low!),
- Se = 0.33, Sp = 0.86 (at cut-off 0.5), 63% correct,
- Se = 0.62, Sp = 0.64 (at cut-off 21/49 ≈ 0.43).





Performance of Discrimination

We should not measure the performance of a classification rule on the same data as it was developed,

- naturally an unrealistically good agreement between observed and predicted,
- difficult to assess degree of "overfitting": a too close adaptation of classifier to the data $\sim$ good data fit but potentially poor predictive performance for other data.

Alternatives for performance assessment ("reliability" in VER2, Section 15.9):

- split of the data into "training (or learning) data", and "validation data" on which the performance is measured,
  — guidelines for split not very specific, e.g. about proportions of data in the two parts and whether one or multiple randomized procedure(s) are required,
- (leave-one-out) cross-validation: training with full data minus one observation left out, which in turn becomes the validation data point; results are then summarized after iteratively leaving observations out in turn.

Typical summaries include:

* confusion matrix: a cross-tabulation of true and predicted class memberships,
* proportion correctly classified, possibly sensitivity and specificity, or involving misclassification costs (all computed from the confusion matrix).

Linear Discriminant Analysis (LDA)

Classical two-part multivariate methods[3] for data with known division into g groups:

(1) classification by Mahalanobis distance ("predictive LDA"),

(2) discrimination by linear function(s) of X-variables ("descriptive LDA").

(1): Classification from group means $\hat{\mu}^{(j)} = \bar{\mathrm{X}}^{(j)}$ and the pooled variance matrix S:

* assign a new observation X to the group (j), to which it has the smallest (squared) Mahalanobis distance $d_M(\mathrm{X}, \hat{\mu}^{(j)}, \mathrm{S})$,
* the j^{th} group classification probability[4]: $\hat{p}_j = \exp\left(-\frac{1}{2}\, d_M(\mathrm{X}, \hat{\mu}^{(j)}\right) / \sum_{l=0}^{g-1} \exp\left(-\frac{1}{2}\, d_M(\mathrm{X}, \hat{\mu}^{(l)})\right)$, is valid under an i.i.d. MVN$(\mu^{(k)}, \Sigma)$ assumption for all groups $k = 1, \ldots, g$.

(2): Linear (also canonical) discriminant function(s) of the form: $Z = a_1 X_1 + \ldots + a_p X_p$ aim to separate the groups as well as possible, in the following sense:

- Z_1 has largest possible $F = \mathrm{MS}_G/\mathrm{MS}_E$ in a 1-way ANOVA for Z_1,
- same for Z_2 subject to Z_2 being uncorrelated with Z_1 within groups; same for Z_3 subject to ...

Explicit solution using matrix algebra expressions:

- at most $\min(p, g-1)$ such variables may be found, but they may not all be significant,[5]
- the Z's are determined as eigenvectors for $\mathrm{W}^{-1}\mathrm{B}$, where W and B are within-group and between-group "variances"[6], and the eigenvectors and eigenvalues may be explored similarly to PCA.

[3] The methods involve different model/data assumptions; both assume equal within-group (co)variances across groups.
[4] A posterior probability for equal prior probabilities $p_{0k} = 1/g$; non-uniform prior probabilities are also possible.
[5] Formal tests for the discriminant functions require the MVN assumption.
[6] W is the MANOVA residual matrix, and B = T−W, with T the total SSCP (sum of squares/cross-products) matrix.

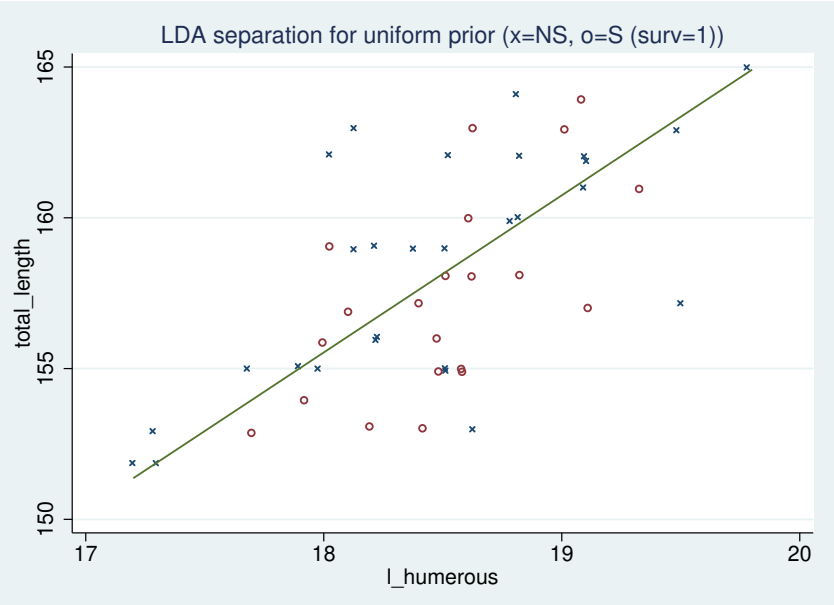
LDA Illustration with 2 Groups

Objective: illustrate LDA with 2 predictors and 2 groups:[7]

Sparrow data with predictors:
`total_length` and `l_humerous`;
estimated discriminant functions:[8]

$$\text{LDA}: \quad 0 = -0.357\,\texttt{tlen} + 1.859\,\texttt{lhum} + 22.033,$$
$$\text{logistic}: \quad 0 = -0.177\,\texttt{tlen} + 0.922\,\texttt{lhum} + 10.623.$$

note: the equations are almost multiples of each other!



Same(!) performance of LDA and logistic classifiers (results for uniform priors):

Method		LDA		LDA cv[a]		logistic		logistic cv[a]	
Group		0	1	0	1	0	1	0	1
surv	0	19	9	15	13	19	9	15	13
	1	7	14	7	14	7	14	7	14
% correct		67.3		59.2		67.3		59.2	

[a] evaluation by leave-one-out cross-validation

[7] With two groups LDA, may be considered as inferior to logistic classification (slide L11−10).
[8] Note that the LDA discriminant function matches the separation for uniform (equal) prior probabilities.

Multinomial Logistic Classification

The multinomial multiple-category logistic regression model[9] with g groups (denoted as $0, \ldots, g-1$) assumes

$$\log \frac{\Pr(Y=j)}{\Pr(Y=0)} = \beta_0^{(j)} + \sum_1^p \beta_k^{(j)} X_k, \quad j = 1, \ldots, g-1.$$

For $g=2$, there is only the single logit equation for $j=1$; otherwise there are equations with separate parameters for all ratios relative to baseline ($g=0$).

- ML estimation of the parameters $\beta_0^{(j)}, \beta_1^{(j)}, \ldots, \beta_p^{(j)}$, $j = 1, \ldots, g-1$,
- given a predictor value $\mathrm{x} = (x_1, \ldots, x_p)$, we can compute estimates for the ratios $\Pr(Y=j|\mathrm{x})/\Pr(Y=0|\mathrm{x})$ and hence also all $\Pr(Y=0|\mathrm{x}), \ldots, \Pr(Y=g|\mathrm{x})$,
- rule: assign to group $0, \ldots, g-1$ with highest probability.

Adjustment for prior probabilities:

- estimates for $\beta_0^{(j)}$ are based on the observed proportions of groups $0, \ldots, g-1$, but can also be adjusted by known prior probabilities (footnote 2).

Example: `beef_ultra` data with 8 measures (5 quantitative, 3 binary) on 487 beef cattle, to predict carcass quality (AAA,AA,A); AAA is best.

- prediction for first row (#43): $\Pr(\mathrm{A})=0.165$, $\Pr(\mathrm{AA})=0.73$, $\Pr(\mathrm{AAA})=0.11$: $\rightarrow$ AA.

[9] Multinomial logistic regression is covered in Chapter 17 of VER/MER.

LDA (3 Groups) Example

Full analysis for `beef_ultra` data with data priors:[10]

- first eigenvalue (for $\mathbf{W}^{-1}\mathbf{B}$) with 95% of the variance,
- misclassification rate $\approx$ 36% (cross-validated), slightly lower than logistic classifier (37%), and both classificers struggle to identify grade A subjects;
  — cross-validated confusion matrices:

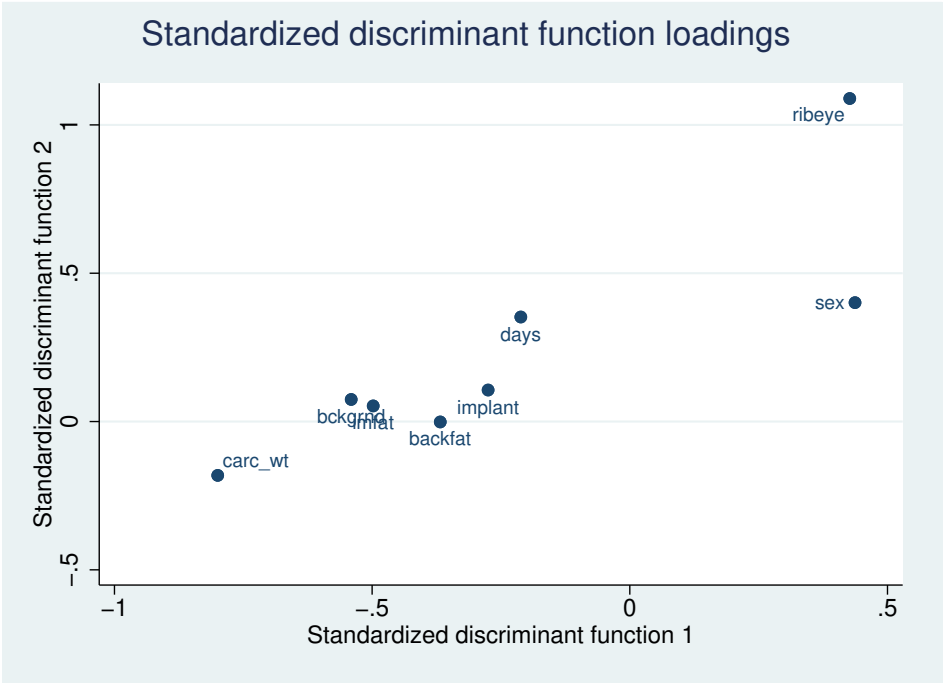
Method		LDA			logistic			
Group		AAA	AA	A	AAA	AA	A	total
`grade`	AAA	76	88	0	74	90	0	164
	AA	43	230	4	43	229	5	277
	A	3	39	4	3	40	3	46

- loading plot (next page): strongest influence on discriminant function by `ribeye`, `sex`, `carc_wt`,
- score plot (next page): no direct separation of groups, but visible association between first component and occurences of groups 1 and 2; third group seems not well captured at all.

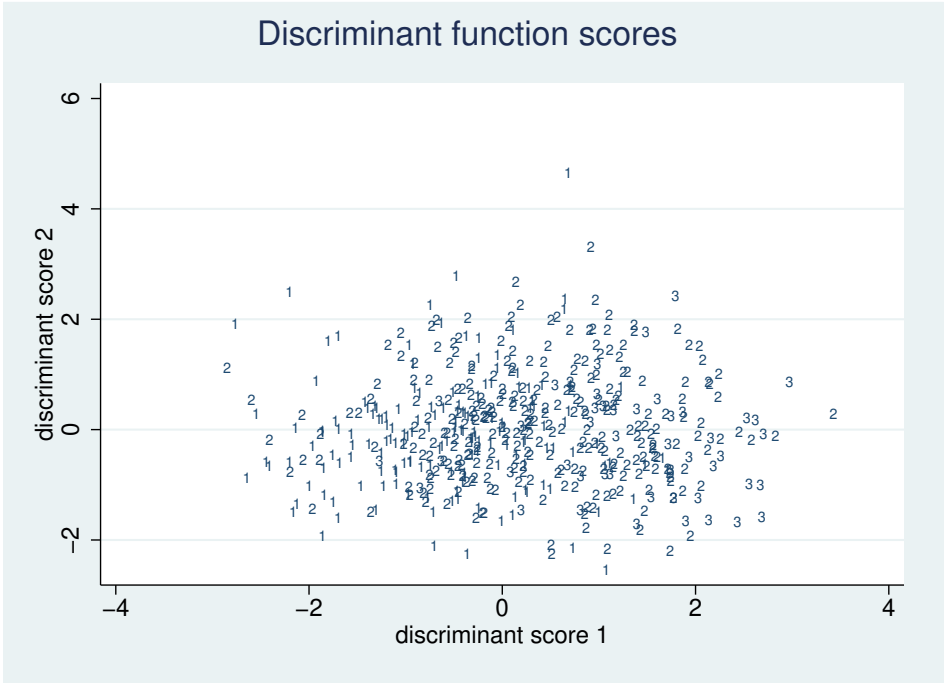
[10] All predictors treated as quantitative; there is really no other option with LDA.

Plots for 3-Group LDA Example

Loading plot:



Score plot:



Summary Remarks for Classical Methods

Main critical point for LDA is the reliance on normality assumptions,

- lack of ability to incorporate non-quantitative variables is a serious drawback,
- even for quantitative variables, the performance can be substantially affected by non-normality and outliers,
- additional assumption made (for predictive LDA): equal variances across groups — can be relaxed by so-called quadratic discriminant analysis (QDA),
  - did not perform well for `beef_ultra` data: — cross-validated confusion matrix:

Group		AAA	AA	A
grade	AAA	93	70	1
	AA	77	179	21
	A	4	30	12

  - extension seems of limited interest in this case,
- for two groups, LDA is similar to logistic classification (which does not have the drawbacks above); some potential advantages of LDA have also been noted: [11]
  - with well-separated classes, logistic classifiers may have unstable parameter estimates,
  - for small n and approximately normal distributions, the extra assumptions may stabilize the procedure.

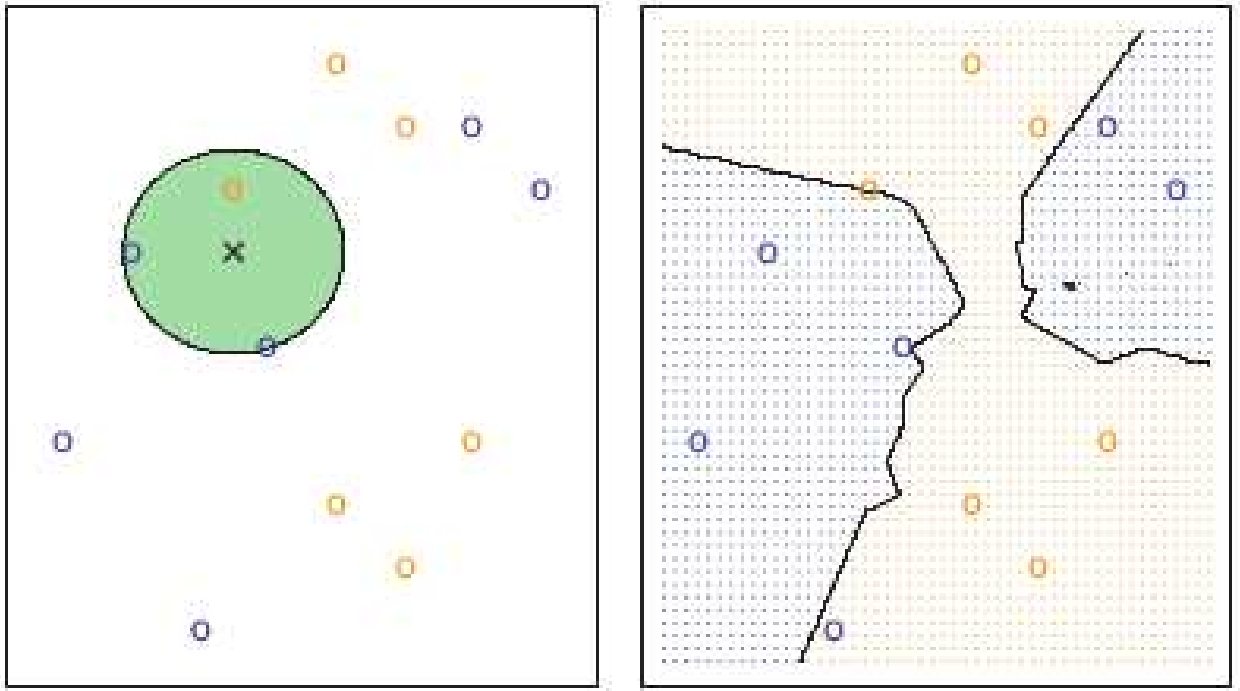
[11] For example, SL (James et al., 2013), Section 4.4.

Kth Nearest Neighbor (KNN) Classification

- a non-parametric discrimination algorithm dating back to the 1950s,
- its name comes from the idea of associating any new point (observation) x, for which a classification is desired, with its k nearest neighbours (say $\mathcal{N}_x$) in the training set,
- estimate group probabilities as the sample proportions in $\mathcal{N}_x$ — done!

Illustration
(Figure 2.14 of SL):

KNN with $k=3$,
6 blue and 6 orange
observations



left: prediction from new observation at x, right: full KNN decision boundary

KNN Settings and Examples

Settings/choices (suggestions from Stata manual):

- value of k[12]: with 2 groups in the data, choose k odd, in range $n^{.25} - n^{.375}$ for roughly equally-sized groups; or as $\sqrt{\cdot}$ of typical group size,
- distance measure: previous discussions of distance apply here as well.

Binary classification: `sparrow` data

- select $k = 3$ according to recommendations,
- misclassification rates with cross-validation $\approx$ same as with logistic/LDA classifiers,
- results clearly worse after standardization of variables.

Full 3-group classification: `beef_ultra` data with all predictors,

- at best, a slightly inferior performance to LDA and logistic classifiers; sample results (with $k = 7$, L_1-distance on standardized variables):

Group		AAA	AA	A
grade	AAA	69	94	1
	AA	56	219	2
	A	1	37	8

- quite commonly, some points are left unclassified by the algorithm (due to ties between groups).

[12] The role of k in the algorithm can be understood as to balance flexibility and stability against overfitting.

Reducing Predictor Variable Dimension

Considerations for a systematic approach to dealing with many predictors (p) relative to number of observations (n):

- for regression, dedicated approaches exist, e.g. ridge regression, partial least squares. . . ,
- principal components regression is simply the method of replacing the original predictors with the principal components derived from them, [13]

but in regression we often want to use our understanding of the predictors (epi!):

* maybe use the components/factors as guidance for selection of variables or for construction of new variables from existing ones,
* maybe useful to split the multivariate analysis into separate analyses for a number of strata where interpretations of components/factors are more manageable,
* maybe measures of agreement among variables within such strata (e.g., Cronbach's alpha) are useful.

What about classification? — a real issue?

- classifiers based on a large number of predictor variables may be numerically feasible and even more robust,
- for "black-box" predictive systems, misclassifications must be investigated.

[13] This raises the interesting question of whether we can select the most important components for analysis; it has been argued to be invalid: Joliffe (1982), *J. Royal Statist. Soc. C* 31, 300−303.

Spatial Clustering — Overview

Main characteristics of a spatial problem:

- spatial dimension: typically 2 or 3, possibly with added time (spatio-temporal),
- spatial information type: either coordinates (points) or aggregates (area summaries),
- data types, separated by information type:
  - grid data: characteristics (e.g., cases or non-cases) or measurements ($\sim$ usual data types) associated with fixed point locations in a grid,
  - point data: locations of points (possibly with characteristics),
  - aggregated data: counts/proportions of units (individuals) with certain features,
- potential objectives (some limited to certain situations):
  - detection of clustering or spatial correlation (typically by significance tests),
  - quantification of degree of global clustering (by suitably estimates),
  - identification of local clusters or test for clusters at pre-defined locations,
  - spatial interpolation $\sim$ prediction in a spatial context.

  — our main focus here: identification of local clusters based on point-based data.

Spatial analysis is a vast field, and we only aim to understand a small corner:

- many methods require specialized software and skills to handle maps,
- many methods rely on heavy mathematical theory and calculation, and are implemented only in dedicated software and add-ons to standard platforms.

SaTScan Methods

SaTScan (maybe short for: Space and Time Scan): methodology and software implementation of regular-shaped cluster detection in space, time and space-time,

- developed by M. Kulldorff (and collaborators) from 1990s onward, to gradually encompass more models and scenarios,
- freely available in SaTScan software (from `https://www.satscan.org`), for multiple platforms and with ample documentation,
- scan statistic: an estimate of a region (among a class of shapes) that optimizes a criterion distinguishing points inside and outside of the region,
- aim of SaTScan statistics: to determine clusters of circular (ellipsoid, cylindrical) shapes to strongest separate selected parameters (e.g., probabilities or means) within and outside of clusters, and associate P-values with these clusters from statistical testing,
- based on (non-random) point data and classical geometry (Euclidean distances),
- range of models/data types: Poisson and Bernoulli (binary), with extensions to normal, exponential (survival) and continuous Poisson (random data locations),
- simulation-based (also "Monte Carlo") P-values calculated by randomization under the null hypothesis studied.

SaTScan Test Idea

= likelihood-ratio test based on likelihood values for parametric statistical models, for example in a purely spatial Bernoulli model:

- consider a target subset Z, say, of the study area,
- denote by $n_1(Z)$, $n_0(Z)$ and $n(Z) = n_1(Z) + n_0(Z)$ the number of events, non-events and total number of points in Z, respectively, and similarly for Z^c,
- assume independence among point outcomes (event/non-event), so that: $n_1(Z) \sim \mathrm{B}(n(Z), p(Z))$ and $n_1(Z^c) \sim B(n(Z^c), p(Z^c))$, and compute

$$\hat{p}(Z) = n_1(Z)/n(Z), \quad \text{and} \quad \hat{p}(Z^c) = n_1(Z^c)/n(Z^c),$$
$$L(Z) = const \times \hat{p}(Z)^{n_1(Z)} \, (1 - \hat{p}(Z))^{n_0(Z)} \times \hat{p}(Z^c)^{n_1(Z^c)} \, (1 - \hat{p}(Z^c))^{n_0(Z^c)}$$

- employ a search across candidate sets Z to maximize $L(Z)$,
- under $\mathrm{H}_0 : p(Z) = p(Z^c)$, $L(Z)$ no longer depends on Z (say, we call it L_0), and the likelihood-ratio statistic equals $L(\hat{Z})/L_0$,
- simulate random allocations of events/non-events across the study area with equal probability for all points, to determine the simulation-based P-value of the observed $L(\hat{Z})/L_0$, relative to the alternatives of interest,
- continue with a similar procedure for secondary clusters not overlapping the first detected cluster, etc.

SaTScan Example: Bernoulli Model

Demonstration of the tutorial example with the SaTScan software (birth defects in New York state); some comments/observations:

- the input data format is `.dbf` (dBase), but the resulting files (`.cas`, `.ctl`, `.geo`) are in plain text format and simple to construct manually,
- also multiple output formats, with the main log file in plain text format,
- the spatial coordinates can be Cartesian or latitude/longitude[14]
- clusters can be defined by either high or low (or both) proportions of events inside cluster relative to outside,
- a maximum size for clusters is required (default at 50% of points),
- clusters can be circular, or elliptical (with some further controls),
- multiple settings for secondary etc. clusters; simplest is hierarchical with no overlap.

Notes on results:

- population size = **1 237 189** (births in 2005), number of cases = **24 940** $\sim$ **2.0%**,
- number of locations (zip-codes) = **1143** $\Rightarrow$ actually grouped binary data (with an assumed independence within locations),
- **7** clusters were detected, hereof four with elevated risk.

[14] With latitude/longitude coordinates, GIS software can directly display the results, e.g. ArcGIS Earth.