Graphical Exploratory Analysis Using Halfspace Depth

Ivan Mizera

University of Alberta
Department of Mathematical and Statistical Sciences
Edmonton, Alberta, Canada

("Edmonton Eulers")

Wien, June 2006

Gratefully acknowledging the support of the
Natural Sciences and Engineering Research Council of Canada

Bivariate halfspace depth (Tukey depth)

Take a fixed collection of **datapoints**:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Given an arbitrary point (x, y):

take all (closed) halfspaces having (x, y) on their boundary;
count how many datapoints lie inside them;
take the minimum of this count over the halfspaces.

That is: the bivariate halfspace depth of a point $\vartheta = (x, y)$ is the minimal number of the datapoints lying in a closed halfspace containing ϑ (on its boundary).

$$\boxed{D(\vartheta) = \inf_{u \neq 0} \#\{i\colon u^T(z_i - \vartheta) \geqslant 0\},}$$

where $z_i = (x_i, y_i)$, $\vartheta = (x, y)$, and $\#\{\cdot\} = \mathrm{card}\{\cdot\}$.

Depth = 0 (movie)

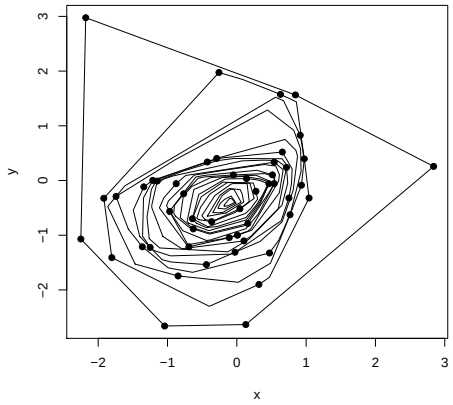
Depth = 1 (movie)

Depth = 2 (movie)

Tukey depth contours

Depth contour of level $k \equiv$ set of points with depth $\geqslant k$.

Nested, convex,...



Bagplot

Rousseeuw, Ruts, and Tukey (1999): a bivariate boxplot

Bag: depth contour containing about 1/2 of observations

Tukey median: a point selected from the contour with maximal depth (various methods possible, the Steiner point is our choice)

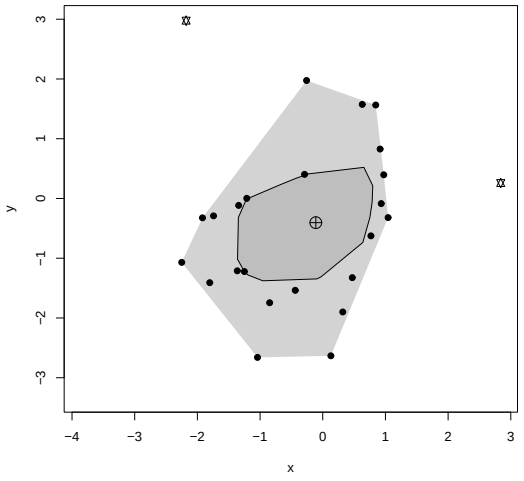
Fence: magnified bag (by fudge factor 3, with Tukey median as center)

Outliers: datapoints outside the fence

Loop: the convex hull of the datapoints inside the fence

Bagplot in action

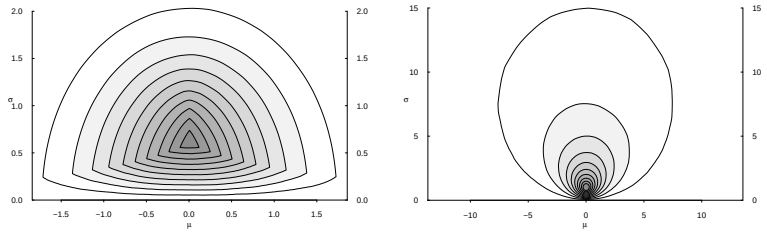
```
> library(depth)
> bagplot(x,y)
```



Student depth (location-scale)

Rousseeuw and Hubert (1998), Mizera (2002).

Mizera and Müller (2004): halfspace depth in the Lobachevski geometry of the location-scale space (a shortest, but perhaps not the most understandable definition).

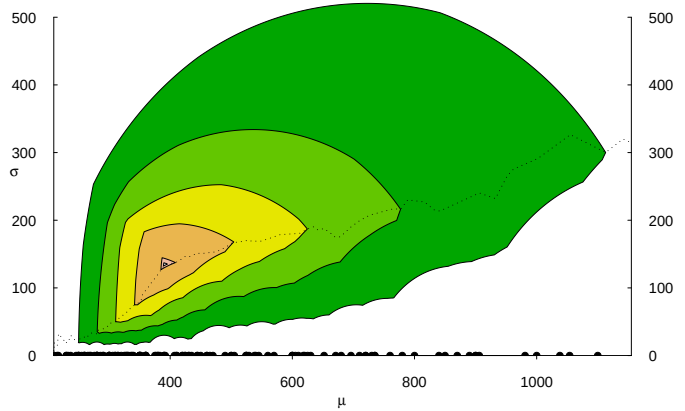



```
> plot(lsdc(rnorm(100000),'dozen'),maxline=F)
> plot(lsdc(rt(100000,1),'dozen'),maxline=F)
```


Depth = 2 (movie)

Student depth contours

```
> plot(lsdc(rivers,"six",maxline = T),paint=terrain.colors(6))
> points(rivers,rivers*0,pch=16)
```



Computer science

In general, NP hard. But plotting fortunately only dim 2.

Student depth contours: $\mathrm{O}(\mathfrak{n})$, apart from the initial $\mathrm{O}(\mathfrak{n} \log \mathfrak{n})$ sorting.

Tukey depth: all contours $\mathrm{O}(\mathfrak{n}^2)$ (but who needs them all?)

Individual depth contours: better? Yes - at least in theory...

Practical algorithm (jointly with David Eppstein): a dynamic convex hull structure (updating strategy).

Implementation: R / ... ?

Interpreted languages (Matlab, R, Python, Lisp) are fun ...

... but slow. Compiled languages (machine code, assembly, FORTRAN, C(++), Java) are fast...

... but are work (= no fun).

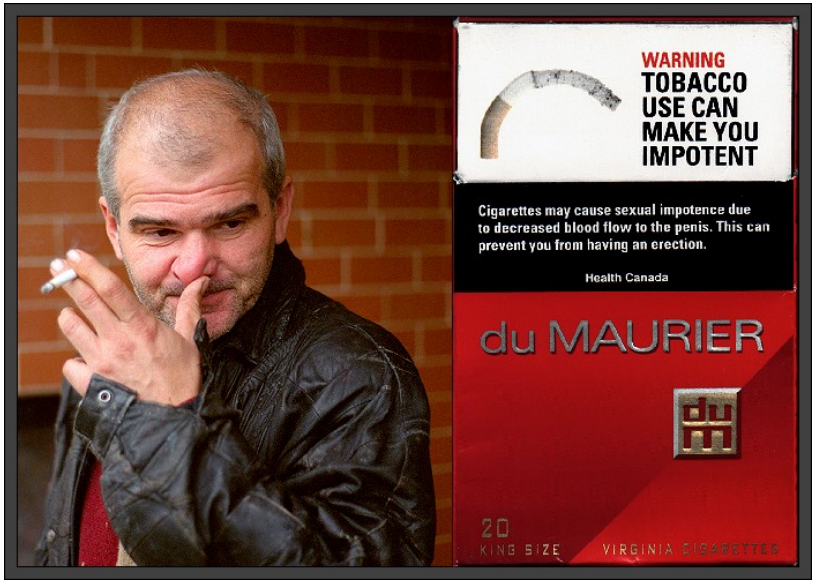
A case study of useR psychoanalysis ($n = 1$)

- FORTRAN avoided (trauma from childhood).
- C routines running (translated from MATLAB, a labor therapy).
- Python prototypes of my co-author David Eppstein deciphered (still waking up at night).
- `Segmentation fault` for $n > 100000$ taken care of (thanks to Duncan Temple Lang for the `S_alloc` command!)
- The next use of `S_alloc` command successfully guessed (without finding any documentation or asking DTL once again).
- Poor Man's Zoom - a Wittgensteinian approach to graphics.
- Eventually, learned how to pass R CMD check (man gets accustomed even to gallows, a Slovak proverb).
- And never ever asked anything on `R-help`.
- It's almost done. (By the anniversary of October revolution?)

Frustrations of a random sample unit: in the search of identity

- (Pressburger blut or Midwesterner in a broad sense?.)
- Computational statistician? Oh, no FORTRAN, thanks...
- UseR from 1998? Bring two witnesses, please.
(UseR $<$ 2000 $\approx$ NSDAP $<$ 1933 or
Czechoslovak Communist Party $<$ 1948)
- Besides, useRs don't worry about things like segmentation faults and `S_alloc` documentation.
- DevelopeR then? Oh, don't make me blushing...
- **AbuseR**. Self-promotion, albeit with attacks of guilty feelings (will a confession get me a pardon?).
- "Don't work on software, work on ideas" (Rich Sutton, a computer science Zen Master from Edmonton).

Warning



Warning

ALTHOUGH ABUSING R
WAS NOT PROVED TO BE ADDICTIVE,
IT SHOULD BE NOTED
THAT IT OFTEN LEADS TO HARDER STUFF.

Viennese epilogue

Stefan Zweig

Theodor Herzl

Some ideas carry a lot of power...

...and the genie is out of the bottle.

Also:

"That what is, often prevails over what could, or even over what should be."

Is it Fellini? (A reward offered for help with this.)