Estimating highest density difference regions using generalised chi-squared tests

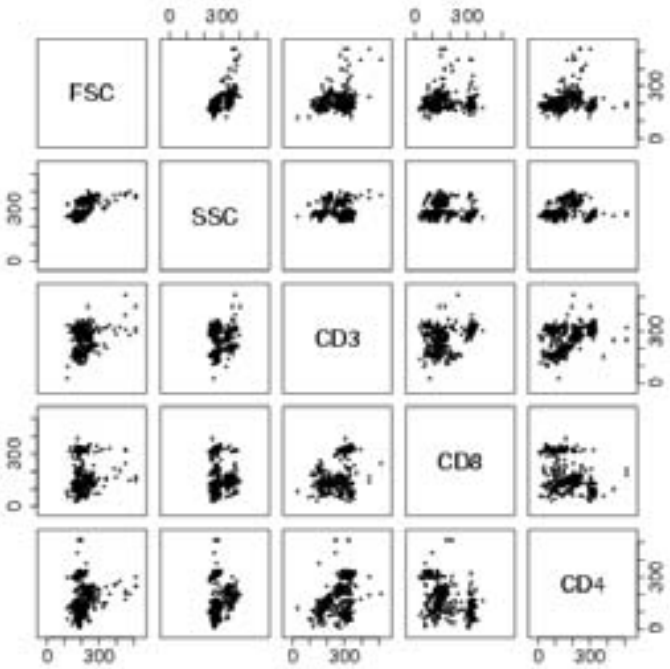
Tarn Duong, Inge Koch & Matt Wand

Department of Statistics
University of New South Wales
Sydney, Australia

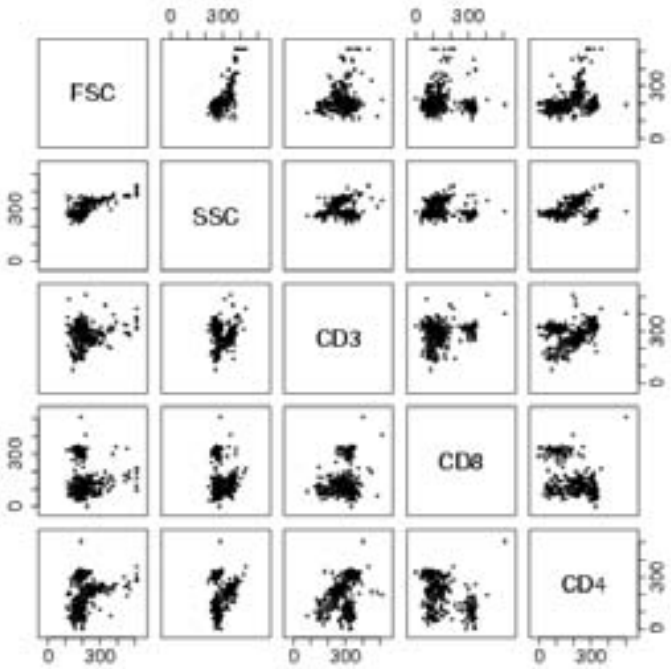
November 2006

Motivating data

HIV+ (control)



HIV– (test)



Questions

1. Are there significant differences between these two samples?
2. If so, **where** are they different?

Existing approaches for general data analysis

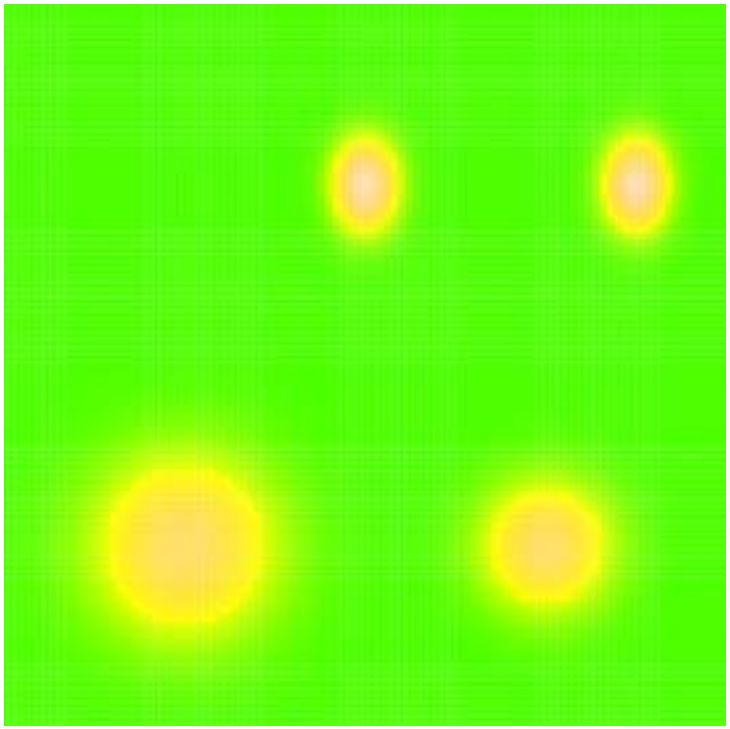
- t-test (normal data)
- Kolmogorov-Smirnov test (maximum distance)
- Mann-Whitney test (ranks)
- Pearson χ^2 test (counts)

Existing approaches for flow cytometry data analysis

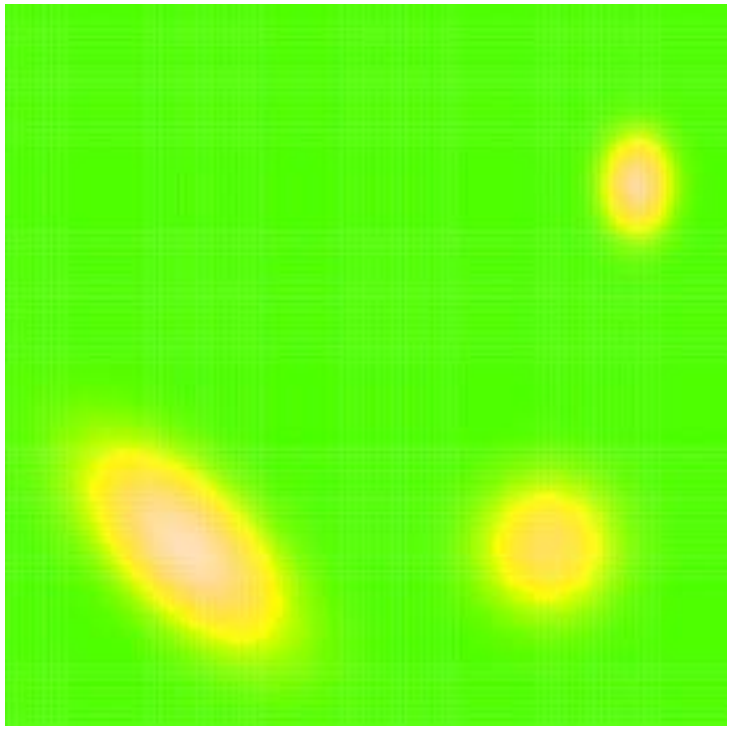
- Classical Pearson χ^2 test (Roederer et al. 2001)
- Good starting point but has deficiencies
  1. Poorly defined problem
  2. Largely unsuitable for samples size $>$ 10 000
  3. Largely unsuitable for dimensions $>$ 2

Control and test density functions

Control density

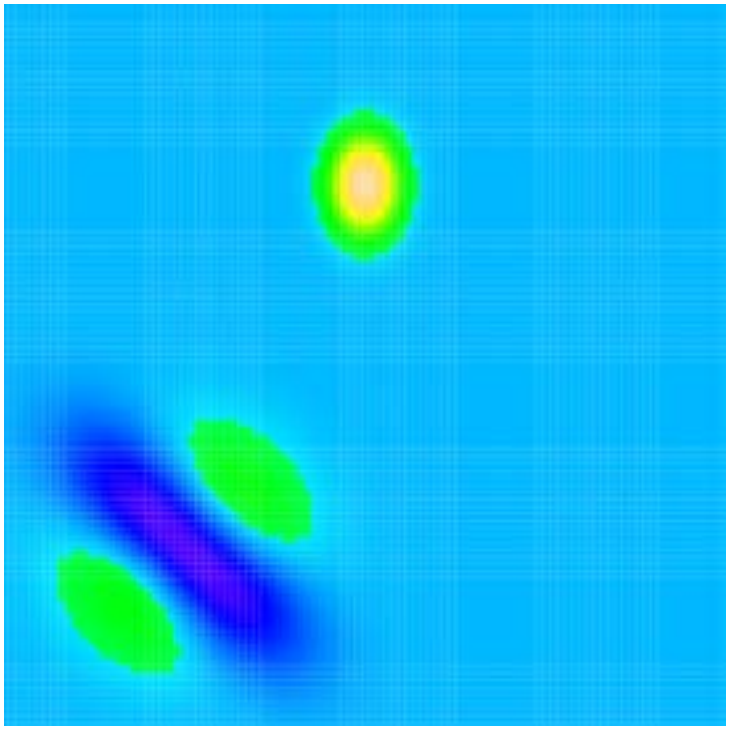


Test density

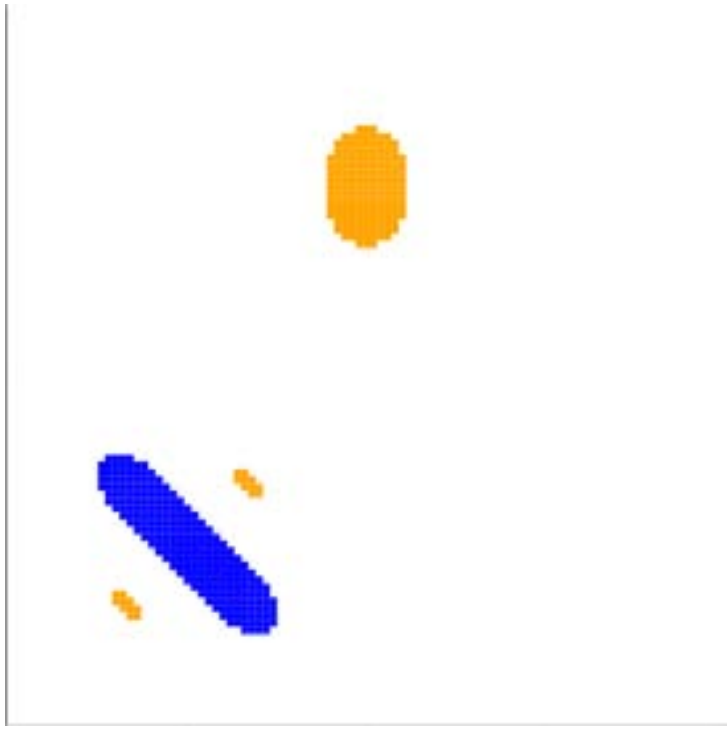


Density difference function

Control – Test



Excess sets



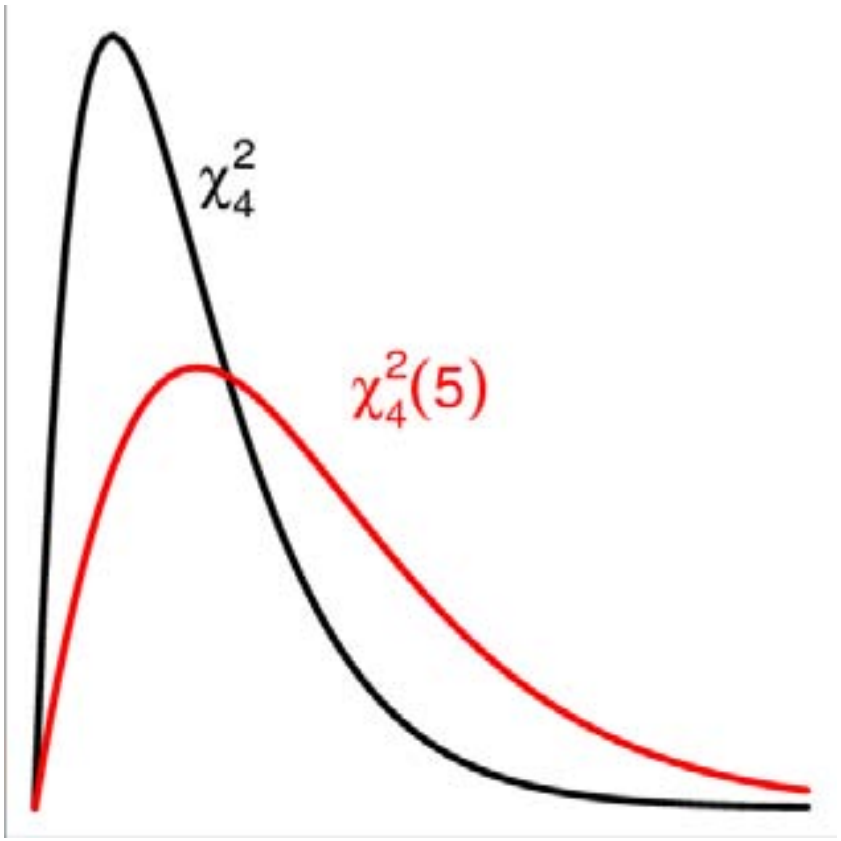
Improvements to flow cytometry data analysis (1)

1. Well-defined problem
2. Largely unsuitable for samples size > 10 000
3. Largely unsuitable for dimensions > 2

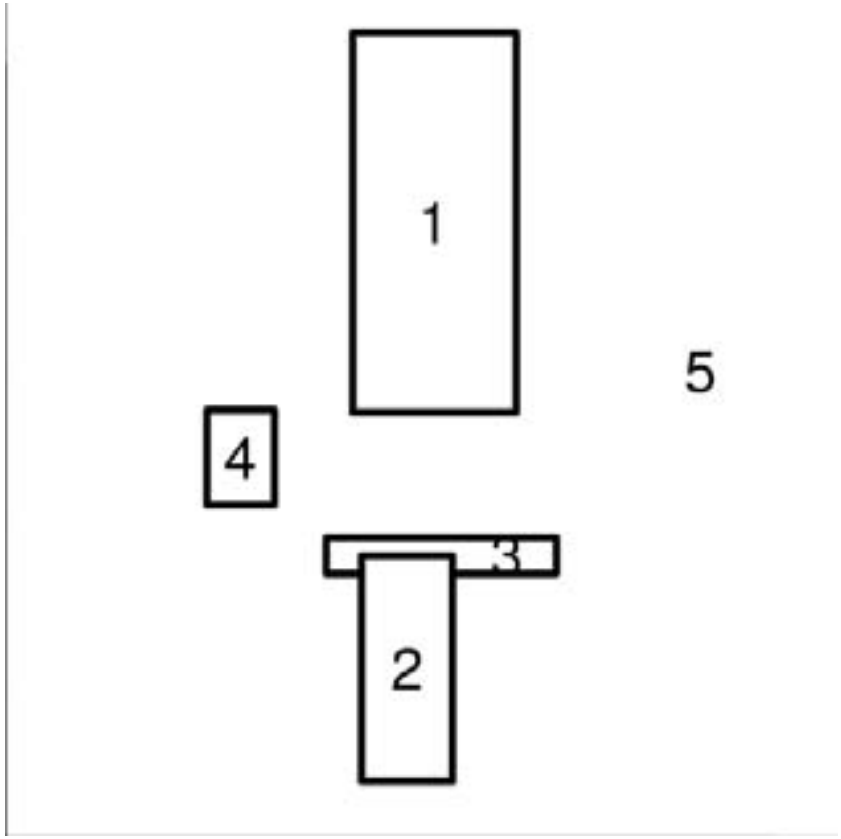
Generalised chi-squared test

- Classical chi-squared tests suited to modest sample sizes (< 1000)
- Generalised chi-squared tests are modification for large sample sizes ($> 10\,000$)

Central and non-central chi-squared distributions

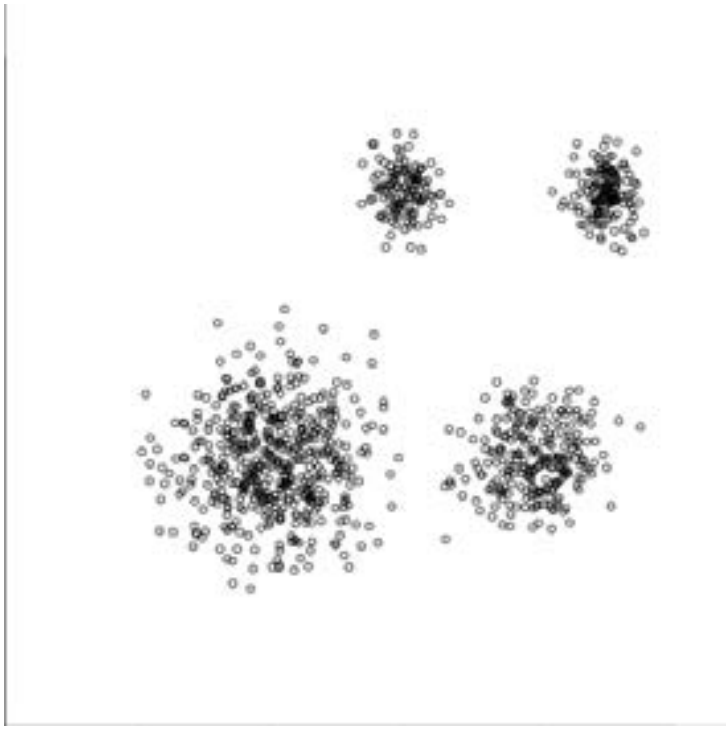


Partition

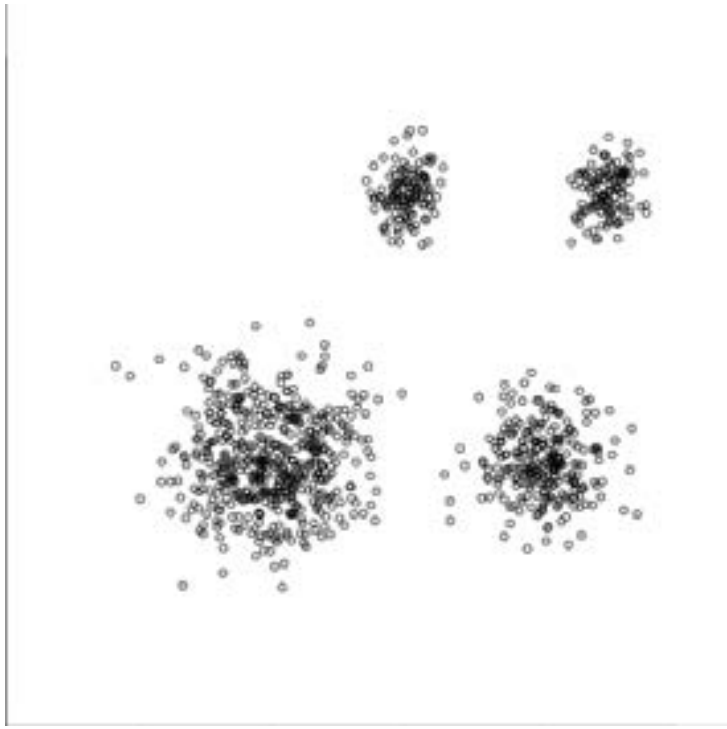


Two control samples

Control sample 1



Control sample 2



Classical vs generalised test

H_0 : control 1 = control 2

Box	1	2	3	4	5	Total
Control 1	12 858	2721	4030	3200	77 191	100 000
Control 2	12 775	2931	3944	3286	77 064	100 000
X^2	0.5393	15.046	1.875	2.251	0.2093	19.92

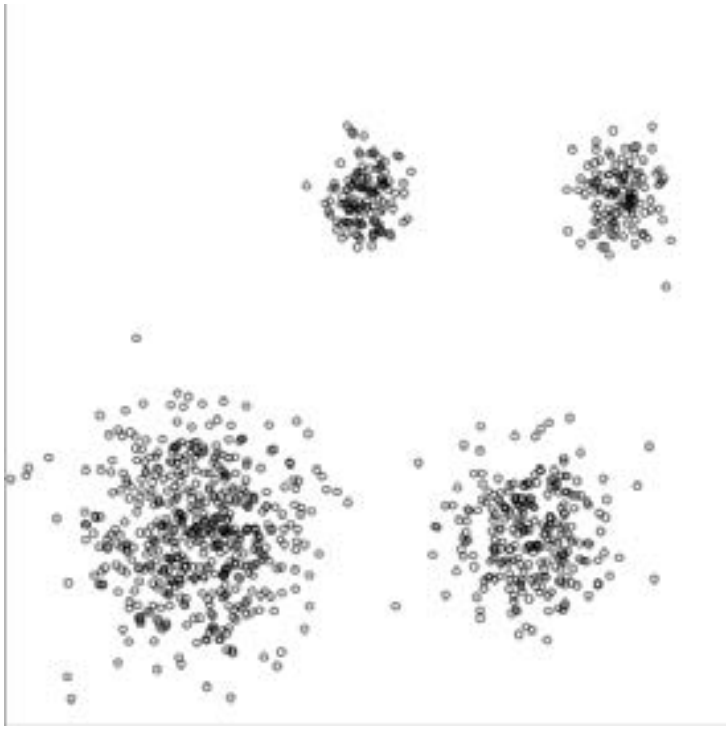
$X^2 = 19.92$

Classical test: 95% crit. value of $\chi^2_4 = 9.488 \Rightarrow$ reject H_0

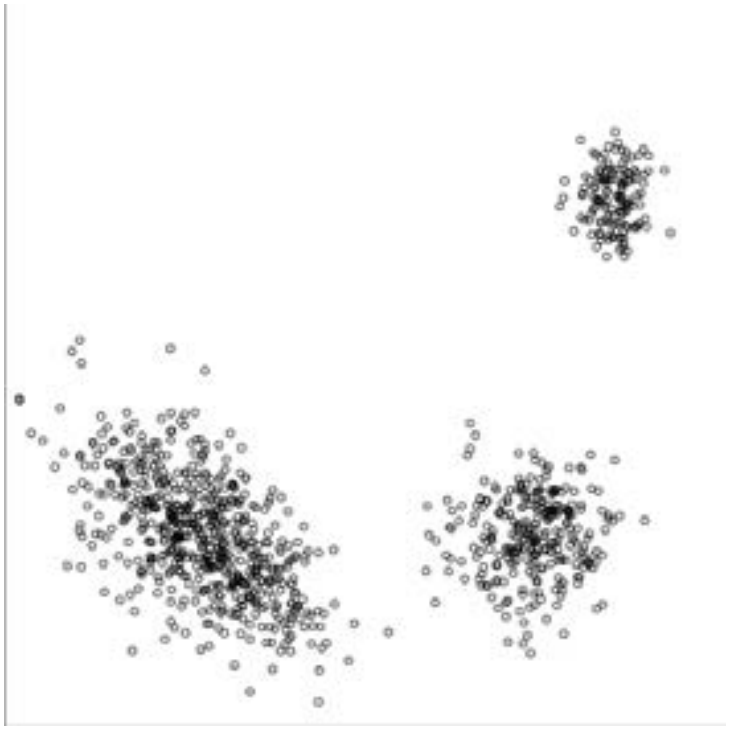
Gen. test: 95% crit. value of $\chi^2_4(1250) = 1372 \Rightarrow$ accept H_0

Control and test sample

Control sample



Test sample



Generalised chi-squared test statistic

H_0 : control = test

Box	1	2	3	4	5	Total
Control	0	777	629	774	7820	10000
Test	1257	244	391	319	7789	10000
X^2	1257	1164	144	649	0.12	3215

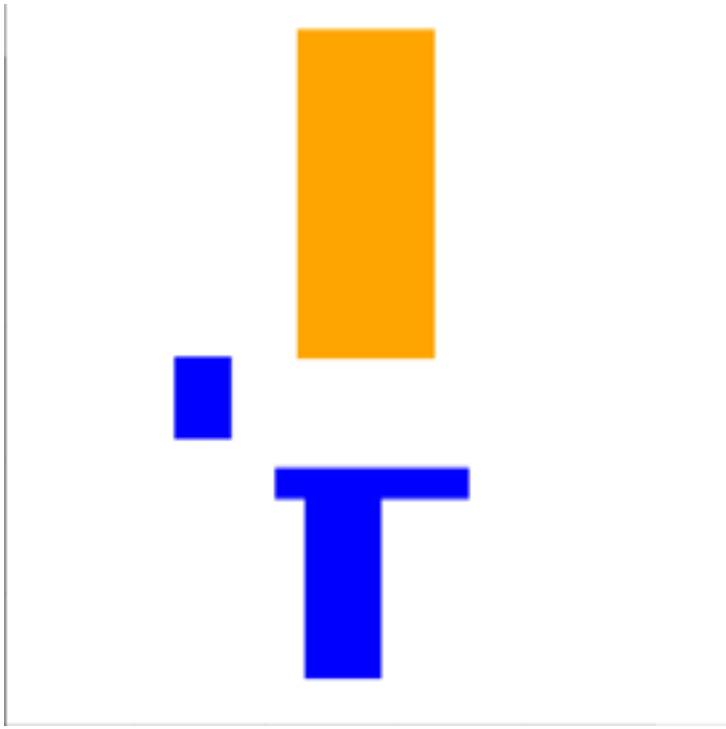
$X^2 = 3215$
95% critical value of $\chi^2_4(125) = 166.70$
Reject H_0 (as expected).

Post-hoc test

- Now have rejected H_0, can use post-hoc tests

Box	1	2	3	4	5	Total
Control	0	777	629	774	7820	10000
Test	1257	244	391	319	7789	10000
X^2	1257	1164	144	649	0.12	3215
Post-hoc	control	test	test	test	equal	

Highest density difference regions



- orange = (control > test)
- blue = (control < test)
- aka Frequency Difference Gate (FDG)

Improvements to flow cytometry data analysis (2)

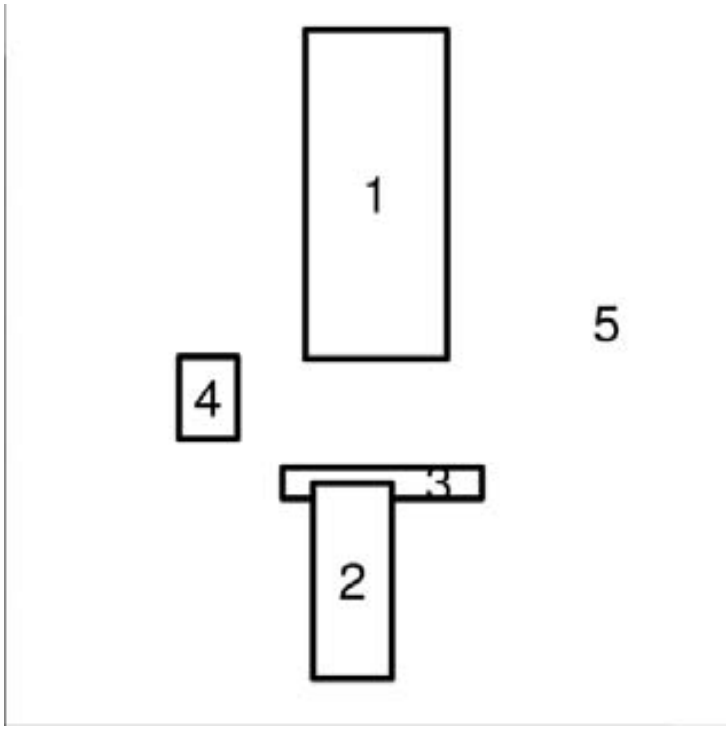
1. Well-defined problem
2. Suitable for samples size > 10 000
3. Largely unsuitable for dimensions > 2

Partitioning algorithms

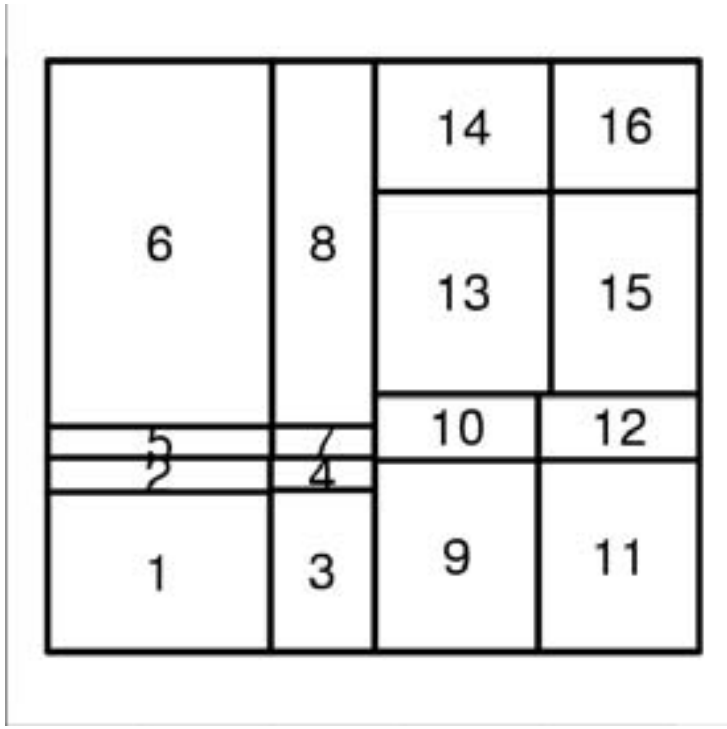
- Patient Rule Induction Method (PRIM) (Friedman & Fisher 1999)
- Probability binning (Roederer et al. 2001)

Partitions

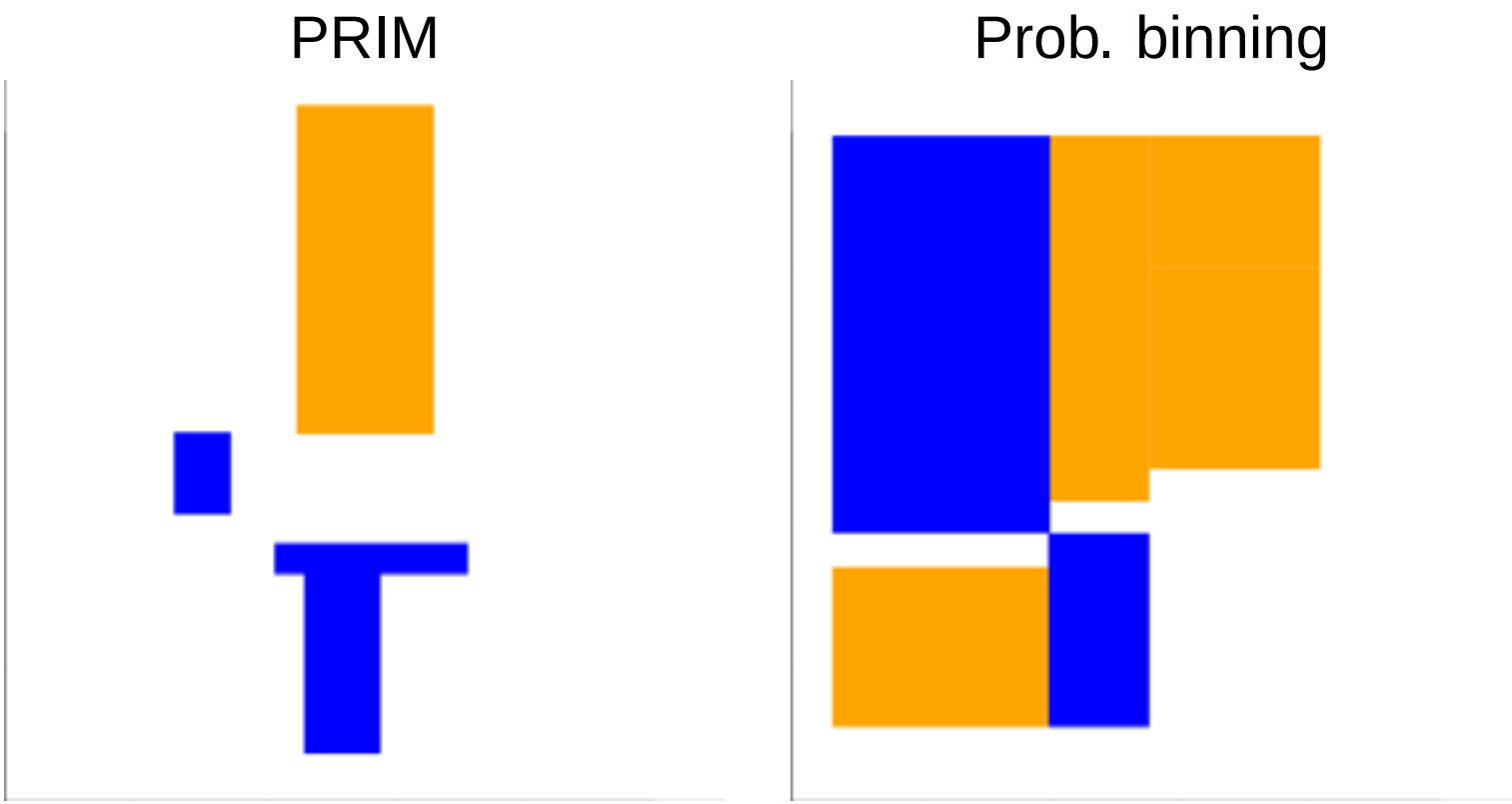
PRIM



Prob. binning



Highest density difference regions



Error measure

$$\text{err}(R, \hat{R}) = \int_{R \Delta \hat{R}} \left[\frac{1}{2} f^c(\mathbf{x}) + \frac{1}{2} f^t(\mathbf{x}) \right] \, d\mathbf{x}$$

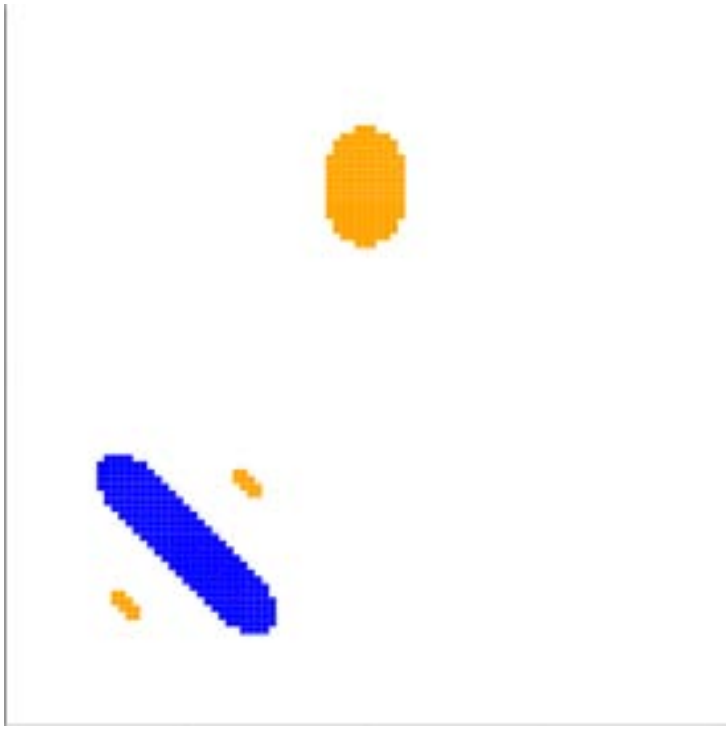
where

$$\begin{aligned} R &= \text{true excess set} \\ \hat{R} &= \text{estimated excess set} \\ R \Delta \hat{R} &= \text{symmetric difference between } R \text{ and } \hat{R} \\ f^c(\mathbf{x}) &= \text{control density} \\ f^t(\mathbf{x}) &= \text{test density} \end{aligned}$$

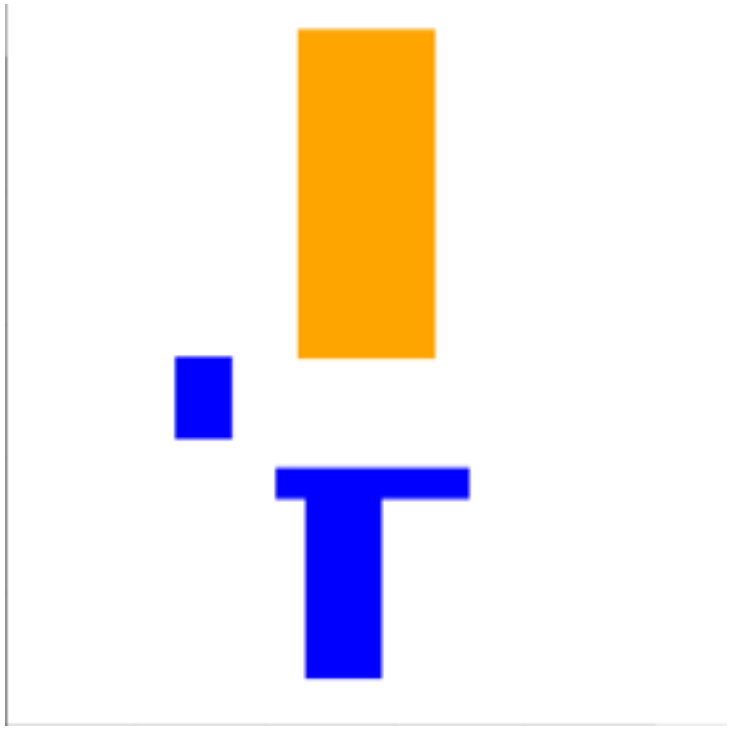
Error for PRIM estimates

$$\text{err}(R, \hat{R}) = 0.2217$$

True



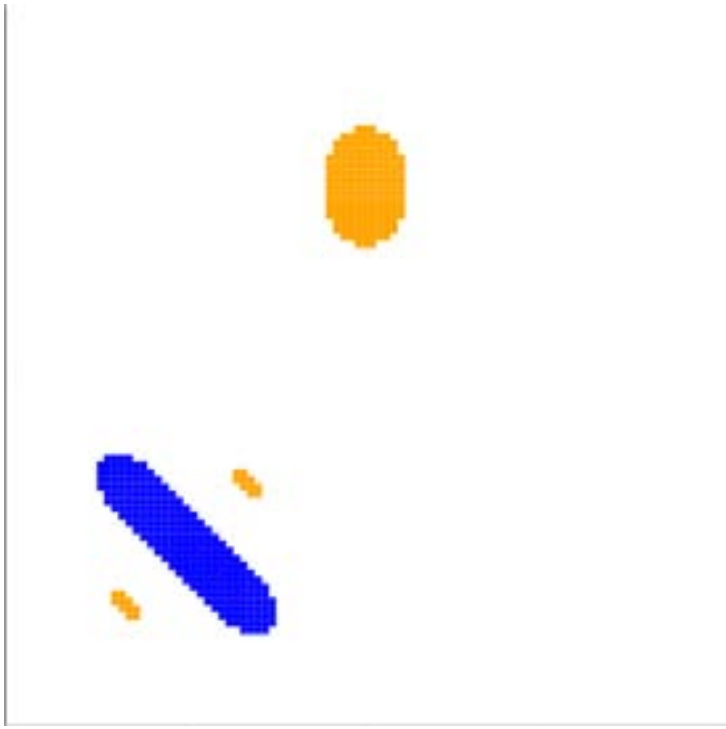
PRIM estimated



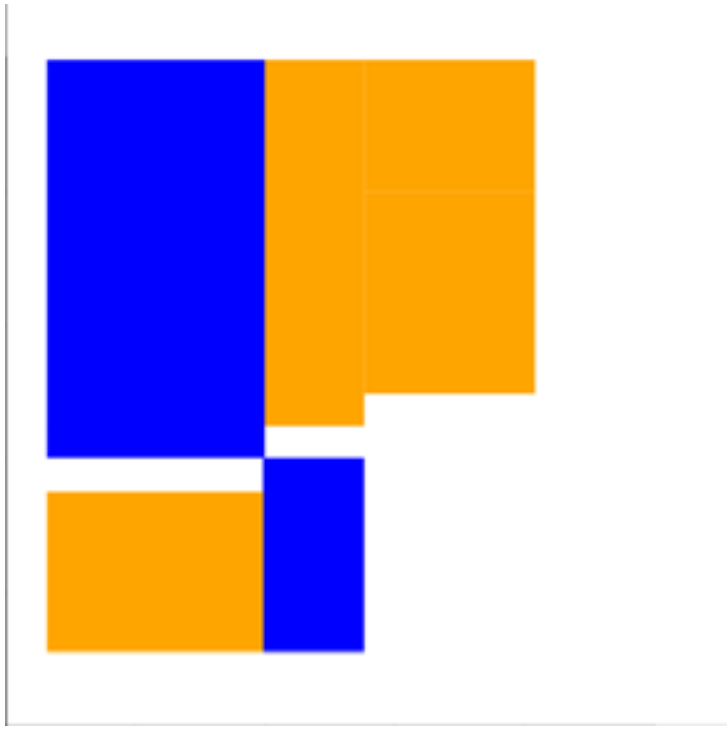
Error for prob. binning estimates

$$\mathrm{err}(R, \hat{R}) = 0.2387$$

True



Prob. binning estimated

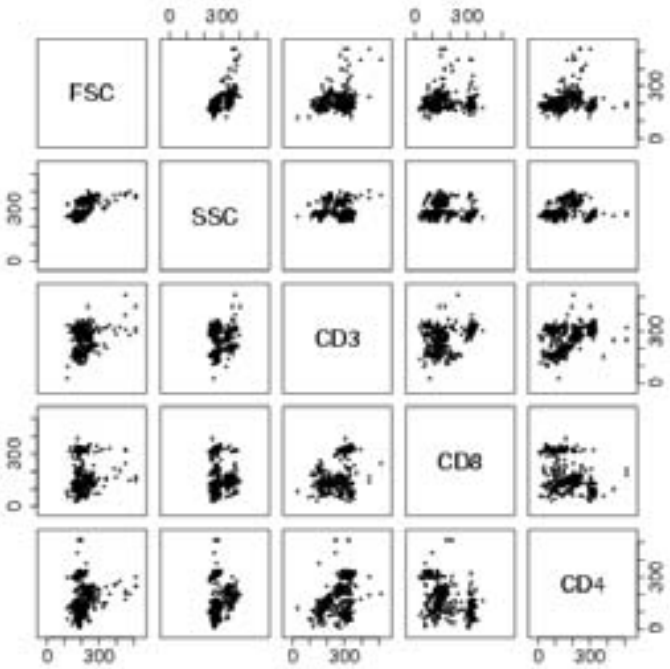


Improvements to flow cytometry data analysis (3)

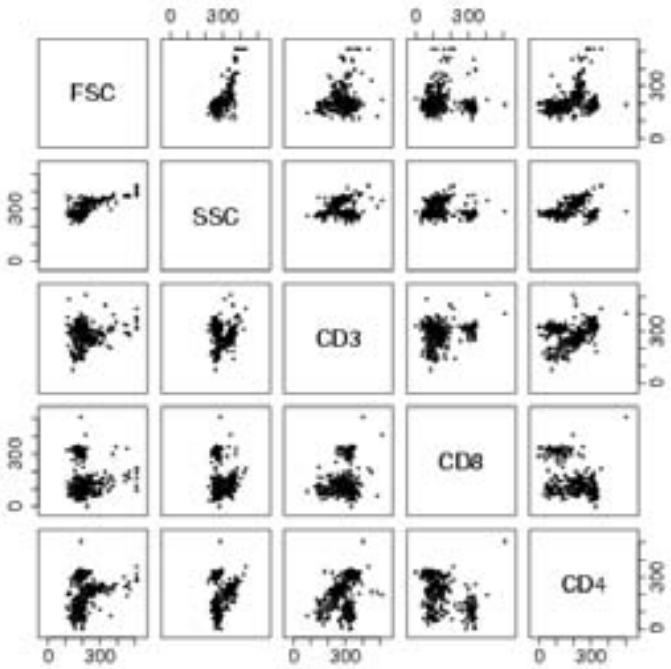
1. Well-defined problem
2. Suitable for samples size $>$ 10 000
3. Suitable for dimensions $>$ 2

HIV data

HIV+ (control)

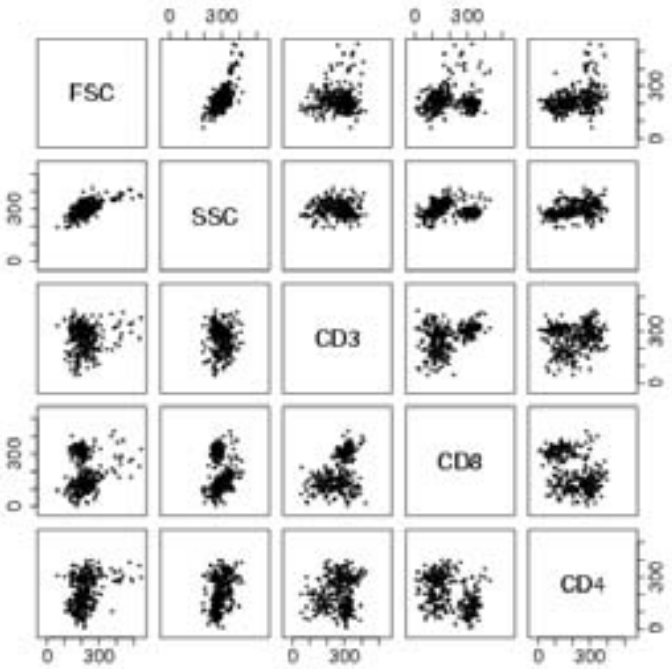


HIV– (test)

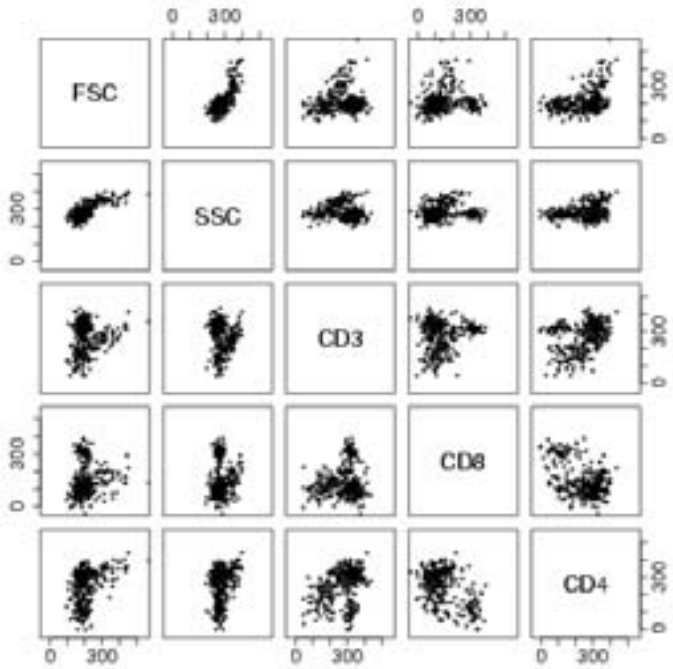


Simulated HIV data

HIV+ (control)

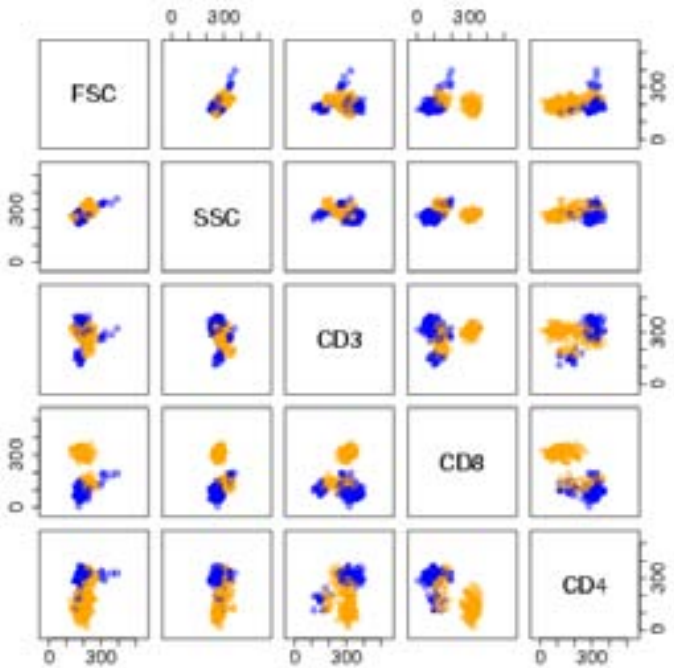


HIV– (test)

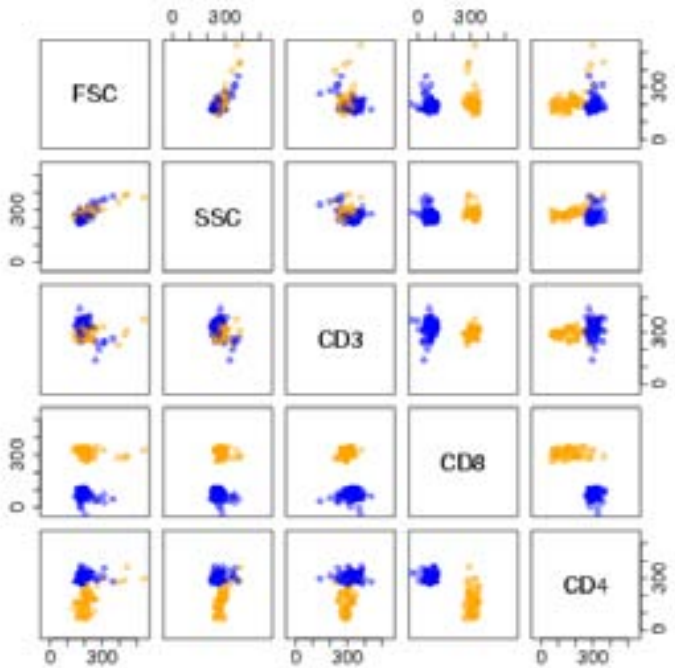


Highest density difference regions

True



Estimated



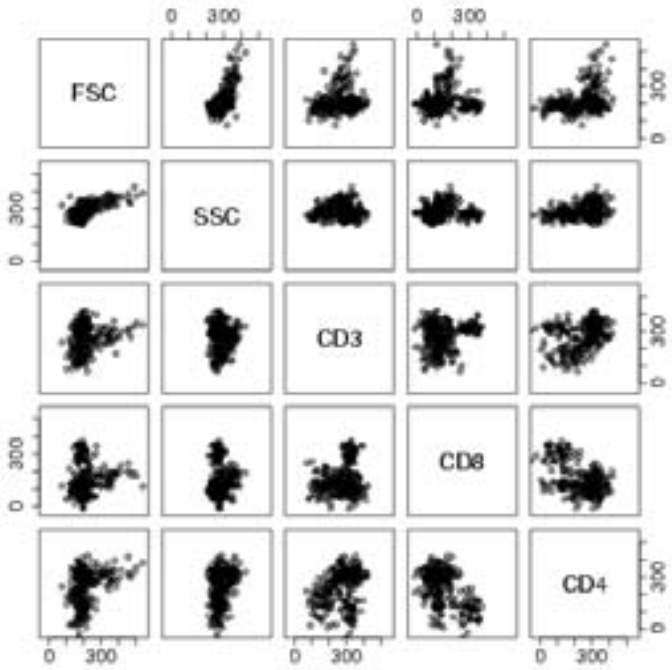
Interpreting highest density difference regions

$$\begin{bmatrix} & min & max \\ FSC & 0 & 725 \\ SSC & 268 & 493 \\ CD3 & 167 & 550 \\ CD8 & 282 & 350 \\ CD4 & 140 & 580 \end{bmatrix} \setminus \begin{bmatrix} & min & max \\ FSC & 33 & 657 \\ SSC & 236 & 456 \\ CD3 & 155 & 309 \\ CD8 & 265 & 348 \\ CD4 & 60 & 516 \end{bmatrix} \Rightarrow +1$$

$$\begin{bmatrix} & min & max \\ FSC & 166 & 725 \\ SSC & 85 & 392 \\ CD3 & 0 & 550 \\ CD8 & 24 & 77 \\ CD4 & 263 & 358 \end{bmatrix} \setminus \begin{bmatrix} & min & max \\ FSC & 138 & 217 \\ SSC & 228 & 302 \\ CD3 & 280 & 383 \\ CD8 & 0 & 114 \\ CD4 & 261 & 365 \end{bmatrix} \Rightarrow -1$$

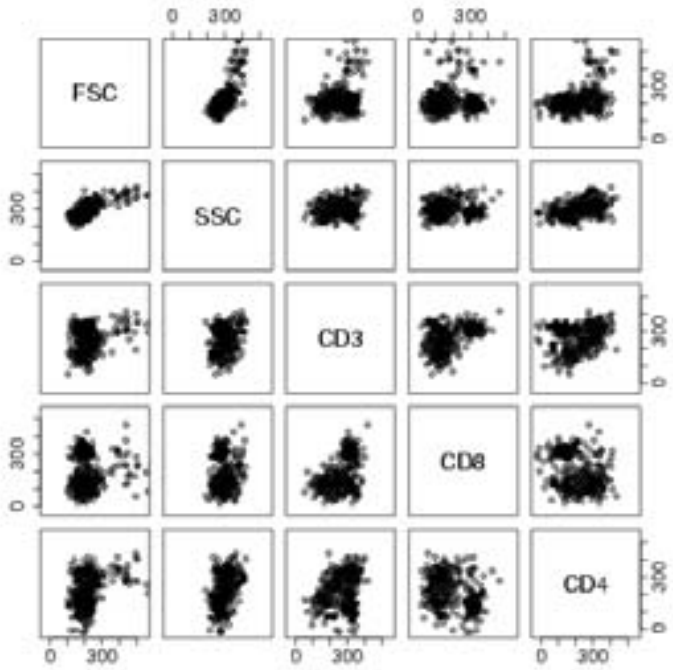
Testing

(HIV+)



Score = 1095 $\Rightarrow$ HIV+

(HIV–)



Score = –1159 $\Rightarrow$ HIV–

Summary and future research

- Summary
  - Technique for finding significantly different regions between two samples
  - Generalised chi-squared $\rightarrow$ suitable for large sample sizes ($>$ 10 000)
  - PRIM $\rightarrow$ suitable for moderate dimensions ($>$ 2)
- Future research
  - Needs more research to be completely data-driven
  - Alternative: estimate density difference function directly