



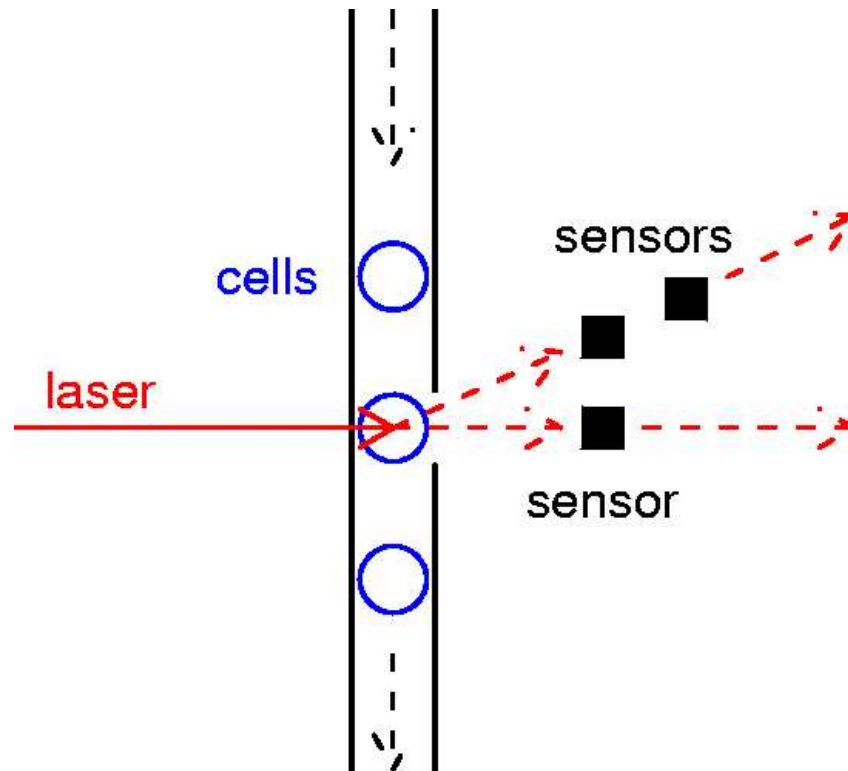
Feature Significance for Multivariate Kernel Density Estimation

Tarn Duong, Arianna Cowling, Inge Koch & Matt Wand
University of New South Wales, Sydney, Australia

July 2006

Institute of Information Sciences and Technology
Massey University, New Zealand

Flow cytometer (1)

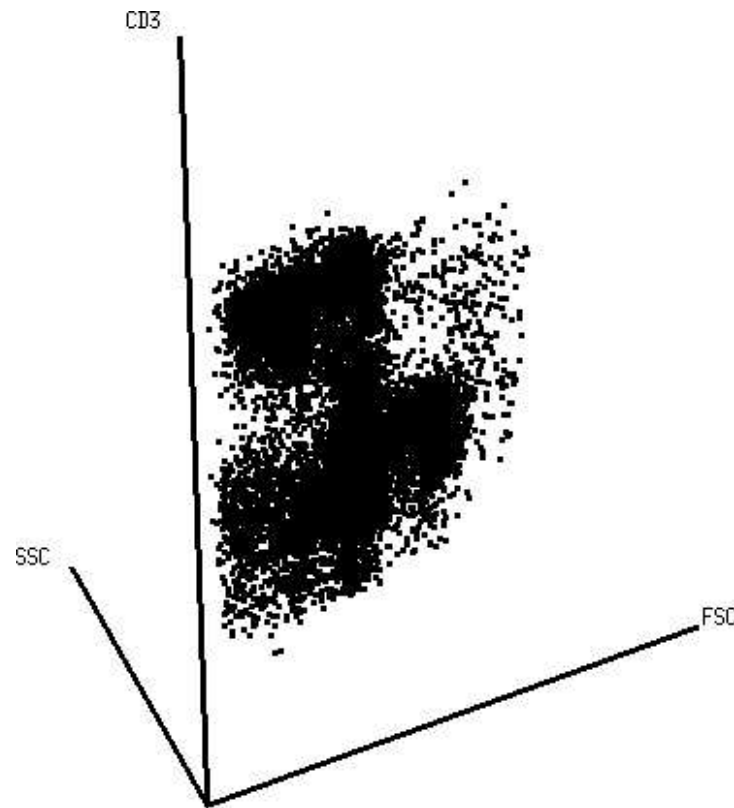


Flow cytometer (2)

- no. sensors = data dim ($3 < \text{dim} < 19$)
- typical data size $\sim 100\ 000$

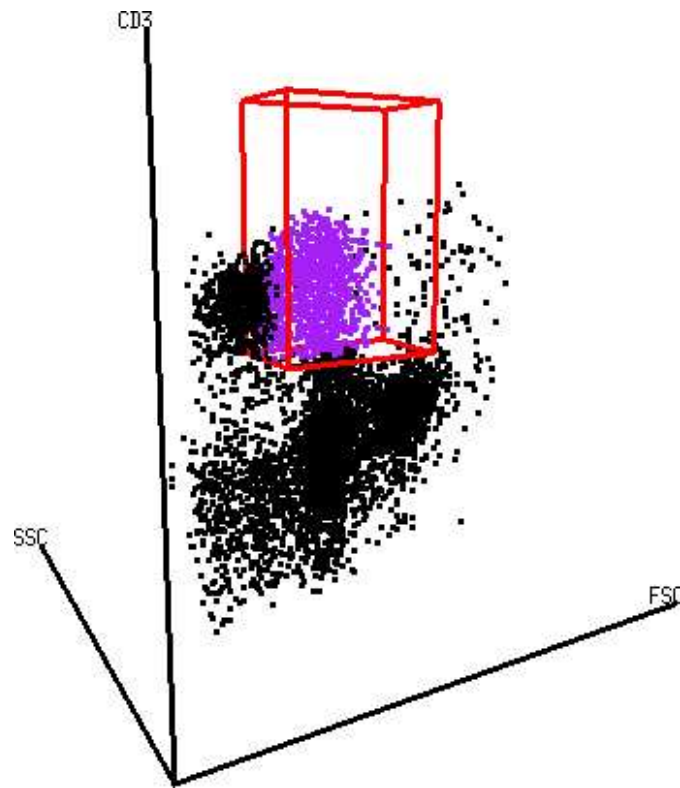
Flow cytometry data (1)

How to isolate lymphocytes (cells of lymph node system) which are important for immune system?



Flow cytometry data (2)

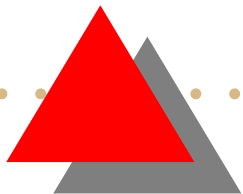
Define gate (red lines) then all cells (purple) inside gate are lymphocytes





Flow cytometry data (3)

- Current methods for defining gates are
 - heuristic (not rigorous)
 - subjective (non-automatic)
- Aim to develop rigorous and automatic gates
- Use kernel density estimation and feature significance



Derivatives

- function $f : \mathbb{R}^d \rightarrow \mathbb{R}$

- gradient $\nabla f = \left[\frac{\partial f}{\partial x_1} \quad \dots \quad \frac{\partial f}{\partial x_d} \right]^T$

- curvature $\nabla^{(2)} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d} & \dots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$



Features (1)

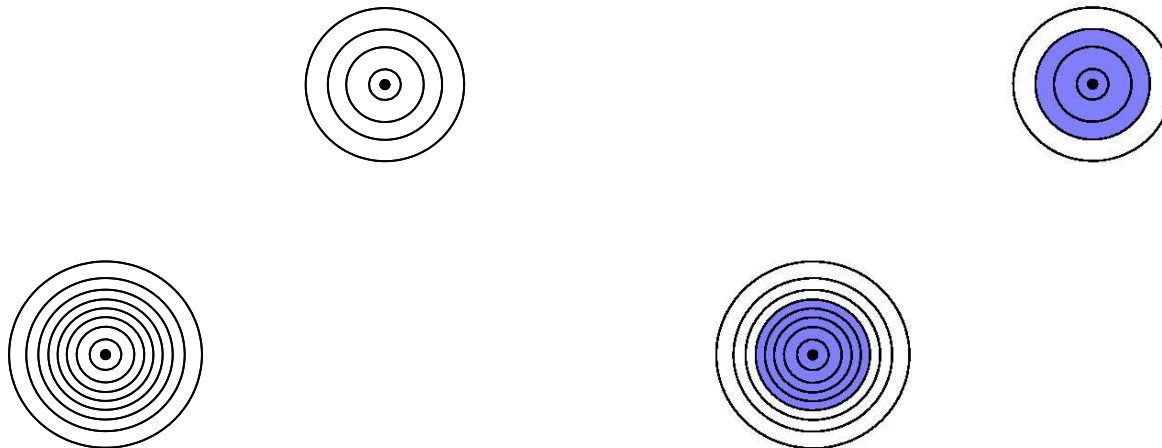
- feature is region with ‘interesting’ gradient and/or curvature
- mode \boldsymbol{x}^*
 - zero gradient $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$
 - negative-definite curvature $\nabla^{(2)} f(\boldsymbol{x}^*) < 0$
- modal region
 - $\{\boldsymbol{x} : \text{small } \nabla f(\boldsymbol{x}), \text{ negative-def. } \nabla^{(2)} f(\boldsymbol{x})\}$

Features (2)

$$f \sim \frac{2}{3}N\left([0\ 0]^T + 0.25\mathbf{I}\right) + \frac{1}{3}N\left([3\ 3]^T + 0.25\mathbf{I}\right)$$

density function

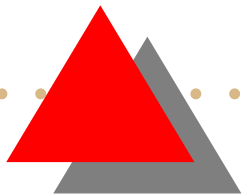
modal regions





Questions

- Q1: How to estimate gradient ∇f and curvature $\nabla^{(2)} f$?
- Q2: How to estimate modal region?
- Q3: How to link modal regions with flow cytometer gates?



Kernel density estimation

Kernel density estimator \hat{f} is

$$\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

where

- $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ is d -dim random sample
- $K_{\mathbf{H}}(\cdot - \boldsymbol{\mu})$ is normal pdf with mean vector $\boldsymbol{\mu}$ and variance matrix \mathbf{H}

Kernel derivative estimation

- kernel gradient estimator

$$\widehat{\nabla} f(\mathbf{x}; \mathbf{H}) = \nabla \hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^n \nabla K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$

- kernel curvature estimator

$$\widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H}) = \nabla^{(2)} \hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^n \nabla^{(2)} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$$



Feature significance

- 1-dim: Chaudhuri & Marron (1999)
- 2-dim: Godtliebsen, Marron & Chaudhuri (2002)
- 3- and 4-dim: current work

Significant gradient

For point \boldsymbol{x} over a grid

- test $H_0(\boldsymbol{x}) : \|\nabla f(\boldsymbol{x})\| = 0$
- using $W^{(1)}(\boldsymbol{x}) = \|\widehat{\boldsymbol{\Sigma}}^{(1)}(\boldsymbol{x})^{-1/2} \widehat{\nabla} f(\boldsymbol{x}; \mathbf{H})\|^2$
where $\widehat{\nabla} f(\boldsymbol{x}; \mathbf{H}) \stackrel{\text{approx.}}{\sim} N(\nabla f(\boldsymbol{x}), \boldsymbol{\Sigma}^{(1)}(\boldsymbol{x}))$
- null distn $W^{(1)}(\boldsymbol{x}) \stackrel{\text{approx.}}{\sim} \chi_d^2$

Significant curvature

For point \boldsymbol{x} over a grid

- test $H_0(\boldsymbol{x}) : \|\text{vech } \nabla^{(2)} f(\boldsymbol{x})\| = 0$ where

$$\text{vech} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = [a \ b \ c \ d \ e \ f]^T$$

- using $W^{(2)}(\boldsymbol{x}) = \|\widehat{\boldsymbol{\Sigma}}^{(2)}(\boldsymbol{x})^{-1/2} \text{vech}[\widehat{\nabla}^{(2)} f(\boldsymbol{x}; \mathbf{H})]\|^2$ where
 $\text{vech}[\widehat{\nabla}^{(2)} f(\boldsymbol{x}; \mathbf{H})] \stackrel{\text{approx.}}{\sim} N(\text{vech}[\nabla^{(2)} f(\boldsymbol{x}; \mathbf{H})], \boldsymbol{\Sigma}^{(2)}(\boldsymbol{x}))$
- null distn $W^{(2)}(\boldsymbol{x}) \stackrel{\text{approx.}}{\sim} \chi_{d(d+1)/2}^2$

Significant modal region

Modal region at signif. level α is

$$\{\mathbf{x} : W^{(1)}(\mathbf{x}) \leq \chi_{d;1-\alpha'}^2, \text{ (non-signif. gradient)}\}$$

and/or

$$W^{(2)}(\mathbf{x}) > \chi_{d(d+1)/2;1-\alpha'}^2,$$

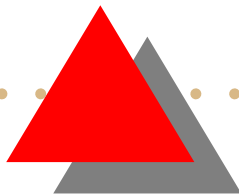
$$\widehat{\nabla^{(2)} f(\mathbf{x}; \mathbf{H})} < 0 \text{ (signif. curvature)}\}$$

α' is adjusted signif. level for testing over grid



Multiple hypothesis testing (1)

- for nearby points x_1 and x_2 , hypotheses $H_0(x_1)$ and $H_0(x_2)$ are highly correlated
- Hochberg's (1988) multiple testing procedure for m correlated tests $H_{0,1}, \dots, H_{0,m}$
- compute corr. p -values P_1, \dots, P_m
- ordered p -values $P_{(1)}, \dots, P_{(m)} \rightarrow$ ordered hypotheses $H_{0,(1)}, \dots, H_{0,(m)}$





Multiple hypothesis testing (2)

- reject $H_{0,(1)}, \dots, H_{0,(j_{\max})}$ where
$$j_{\max} = \operatorname{argmax}_{1 \leq j \leq m} \{P_{(j)} \leq \alpha / (m - j + 1)\}$$
- overall level of sig. is α (no proof given)
- improvement over current method (more rigorous)

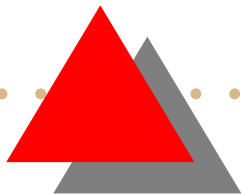
Variance estimation (1)

- $\Sigma^{(1)}(\mathbf{x}) = \text{Var}[\widehat{\nabla} f(\mathbf{x}; \mathbf{H})]$
- $\Sigma^{(1)}(\mathbf{x}) = \mathbf{C}\mathbf{C}^T f(\mathbf{x})$ where \mathbf{C} is const. matrix wrt \mathbf{x}
- estimator $\widehat{\Sigma}^{(1)}(\mathbf{x}) = \mathbf{C}\mathbf{C}^T \widehat{f}(\mathbf{x}; \mathbf{H})$
- $\widehat{\Sigma}^{(1)}(\mathbf{x})$ is non-singular *and* +ve-definite
- $W^{(1)}(\mathbf{x}) = \|\widehat{\Sigma}^{(1)}(\mathbf{x})^{-1/2} \widehat{\nabla} f(\mathbf{x}; \mathbf{H})\|^2$



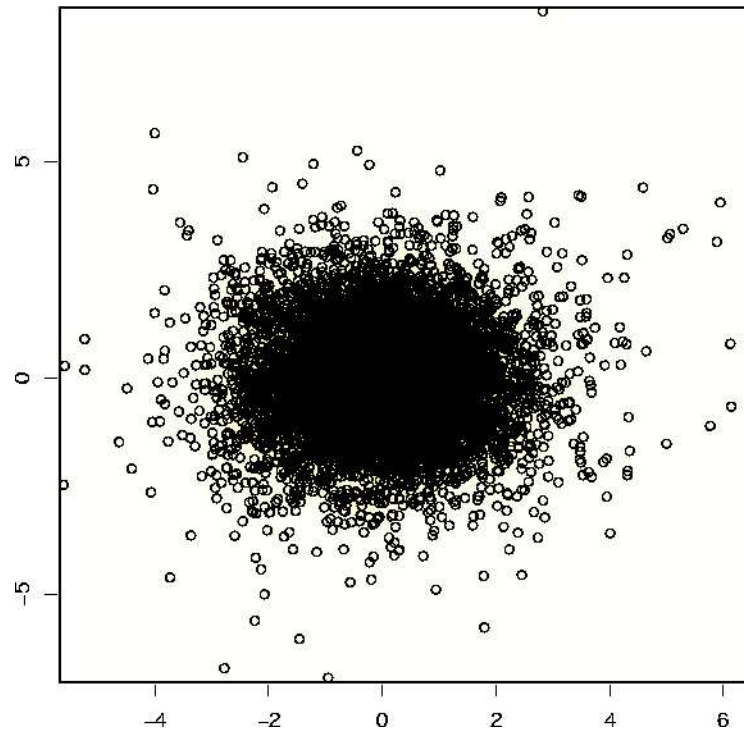
Variance estimation (2)

- similar results for $\Sigma^{(2)}(\mathbf{x}) = \text{Var}\{\text{vech}[\widehat{\nabla^{(2)}} f(\mathbf{x}; \mathbf{H})]\}$
- improvement over current method
- reduced computation



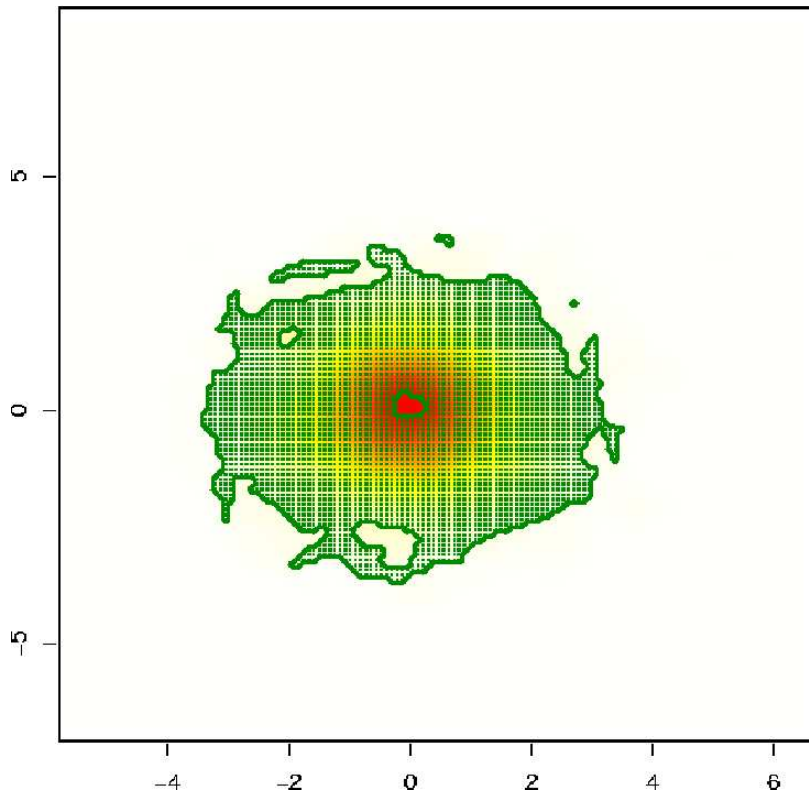
Simulation example - 2-dim (1)

- 10 000 sample \sim 2-dim t -distn with 8 d.f.
- mode at (0,0)

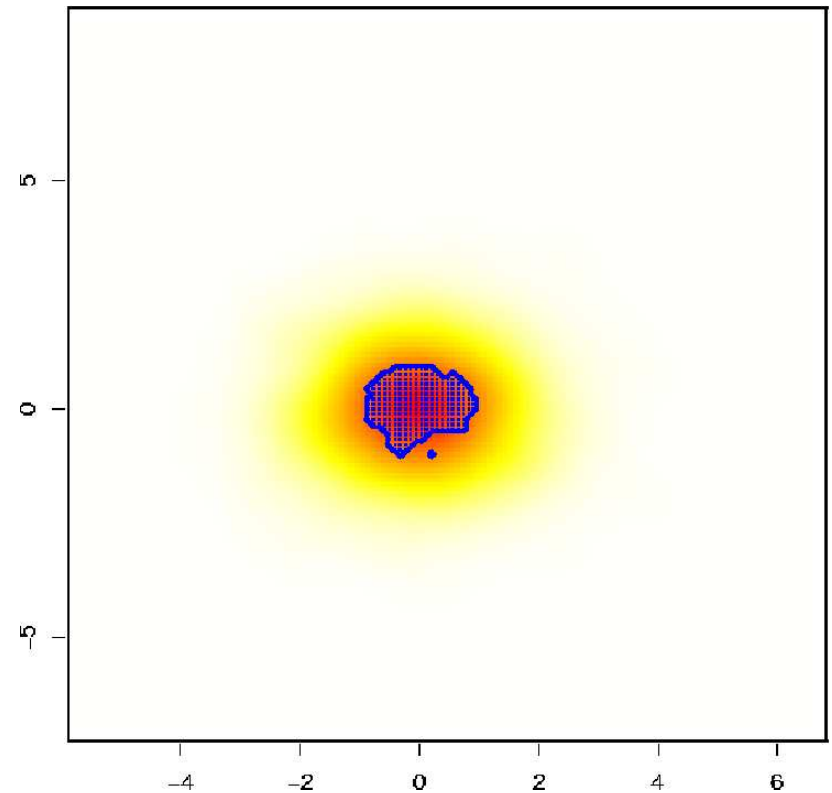


Simulation example - 2-dim (2)

Signif. gradient region



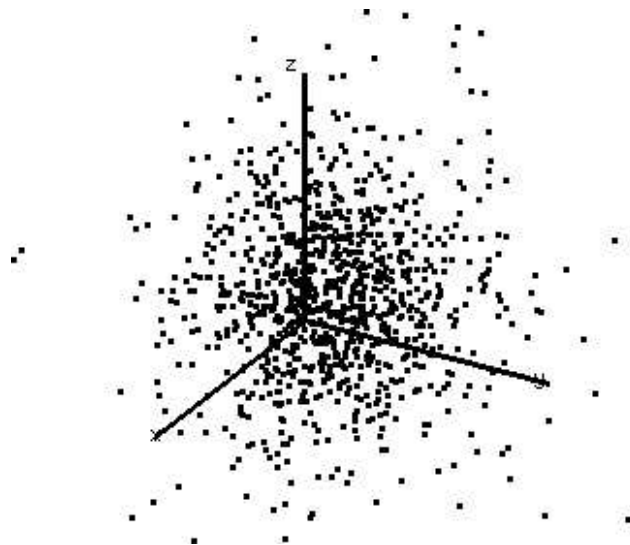
Signif. curvature region



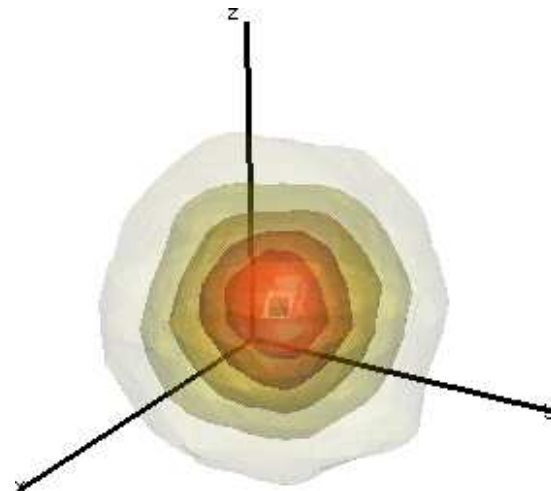
Simulation example - 3-dim (1)

- 10 000 sample \sim 3-dim t -distn with 8 d.f.
- mode at (0,0,0)

Scatter plot

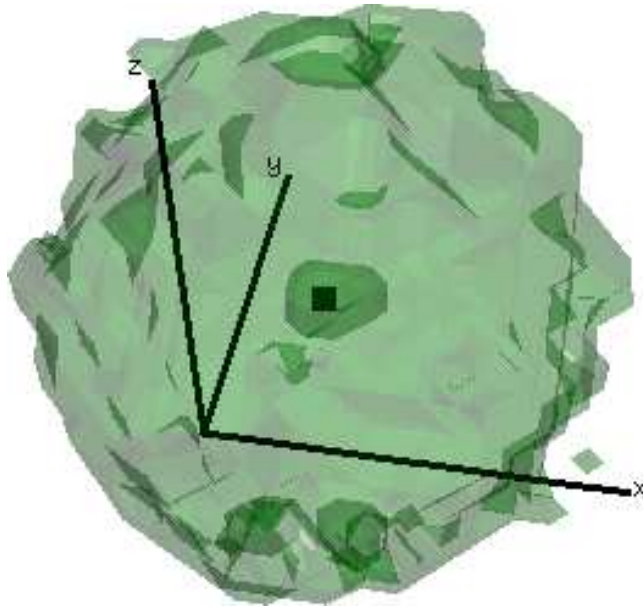


KDE

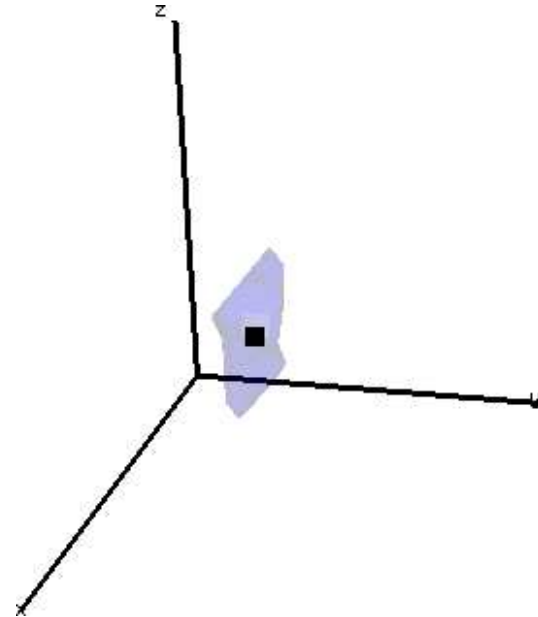


Simulation example - 3-dim (2)

Signif. gradient region

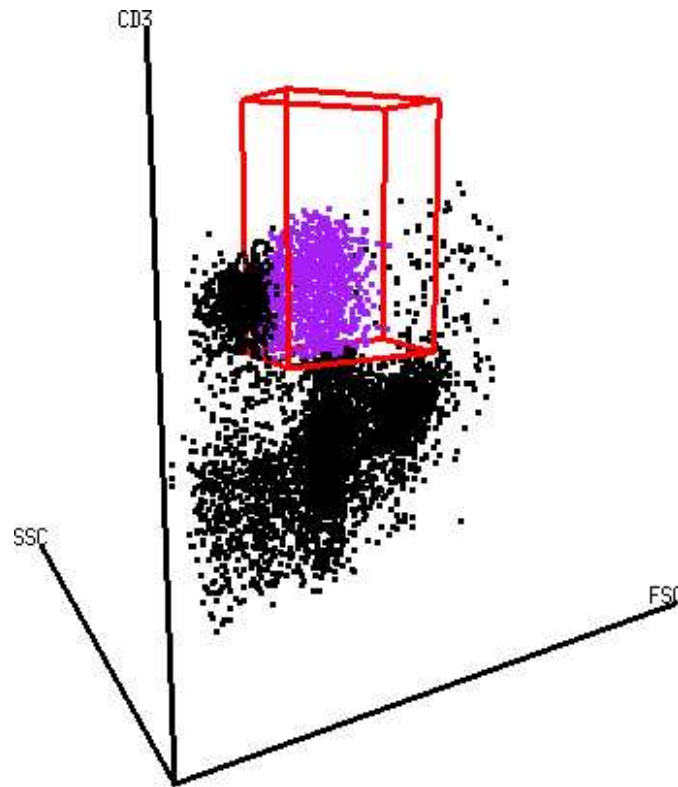


Signif. curvature region



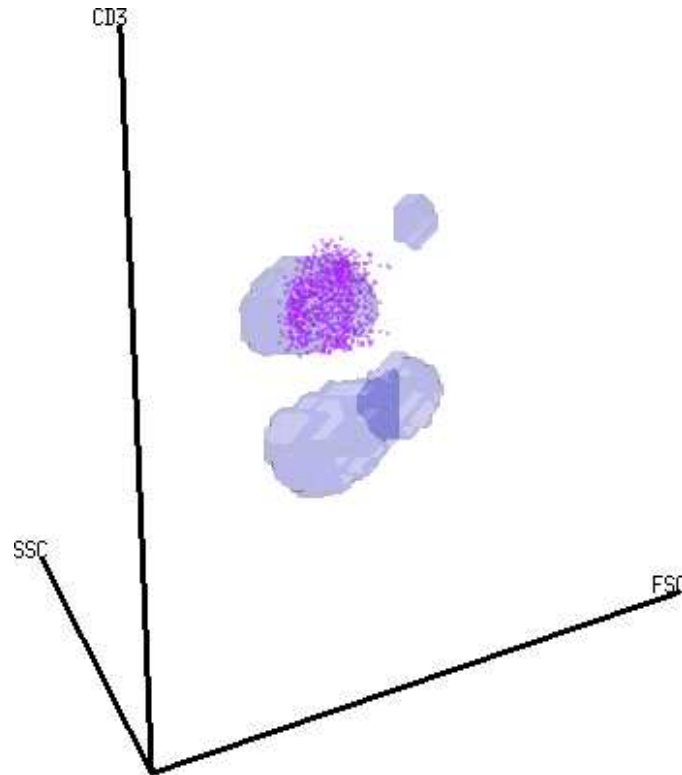
Return to flow data (1)

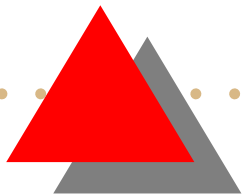
Subjective, manual gate for lymphocytes



Return to flow data (2)

Signif. curvature \leftrightarrow rigorous, automatic gate

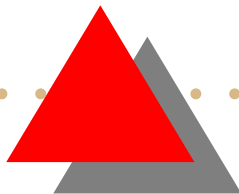






Bandwidth selection

- single optimal bandwidth (matrix) – see Cowling (2005) honours thesis
 - most promising approach for defining flow cytometer gates
- complementary approach is scale space
 - signif. features as function of bandwidth



Scale space

- Click for [scale space movie](#)
- small \rightarrow large \rightarrow small bandwidths



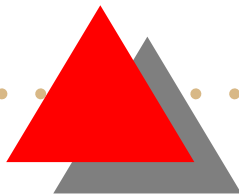
Software

- R package feature
- interactive feature significance for 1- to 4-dim
- released soon on CRAN (R archive)



Questions & Answers

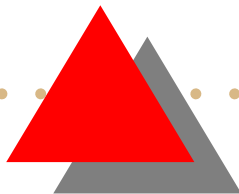
- Q1: How to estimate gradient ∇f and curvature $\nabla^{(2)} f$?
- **A1: Kernel density derivative estimators**
- Q2: How to estimate modal region?
- **A2: Feature significance**
- Q3: How to link modal regions with flow cytometer gates?
- **A3: Ongoing...**





Future research

- higher dim (> 4 -dim) feature signif. diffi cult with KDE
 - problems with data sparseness
 - e.g. try high-dim methods from data mining
- more rigour for signif. feature \leftrightarrow flow cytometer gates
 - e.g. optimal bandwidth selection for kernel density derivative estimators





UNSW flow cytometry webpage

www.maths.unsw.edu.au/~wand/flowcyt.html

