From Histograms to Multivariate Polynomial Histograms and Shape Estimation

Assoc Prof Inge Koch

Statistics, School of Mathematical Sciences
University of Adelaide

Motivation: determine the *shape* of data

We have 12 measurements on each of 27,994 blood cells

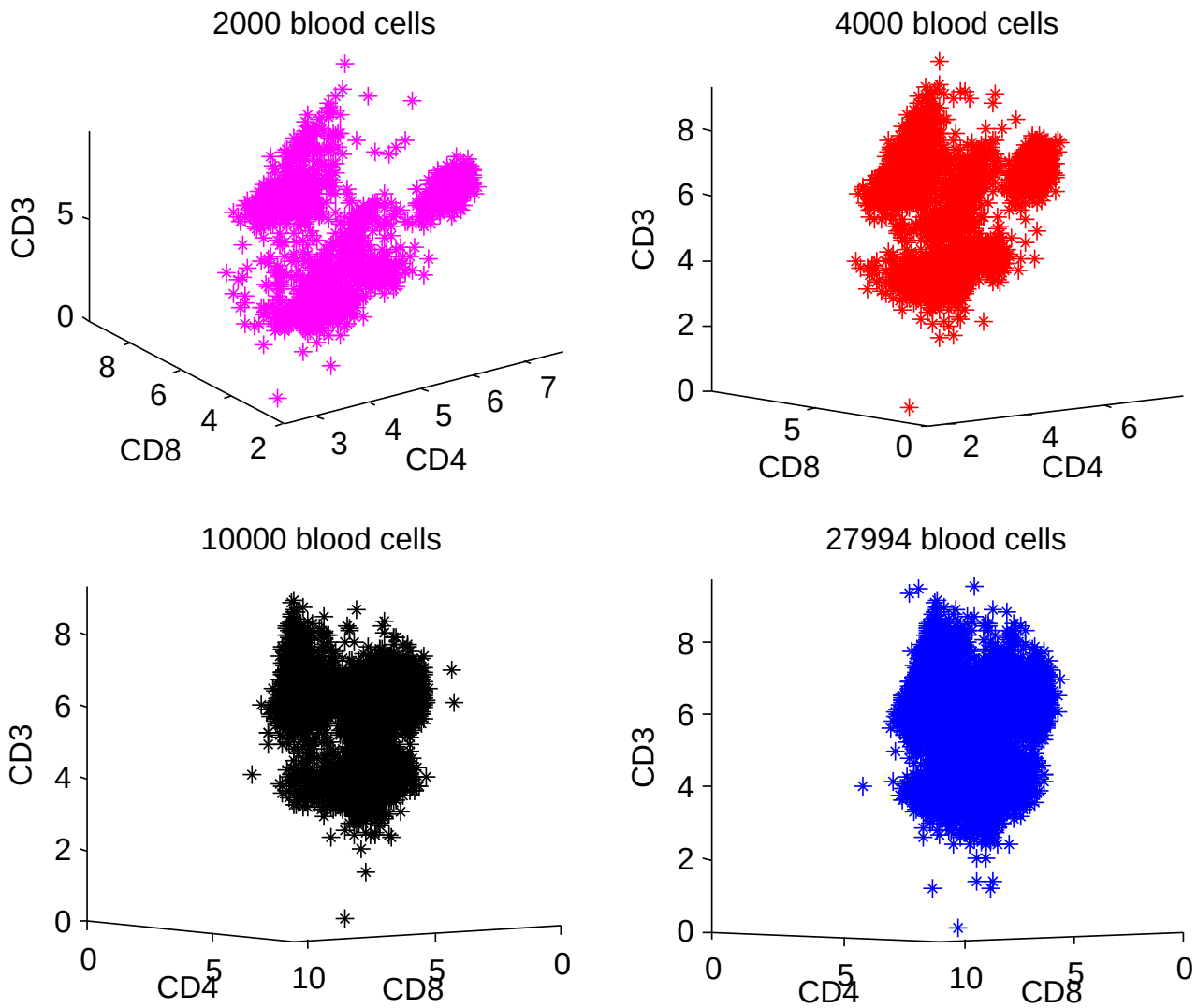
How many cluster?

How big are they and where are they?

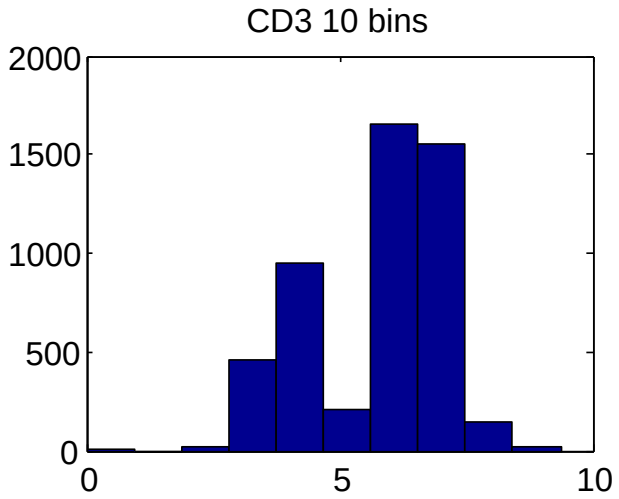
Data: Centre for Immunology, St Vincent Hospital, Sydney

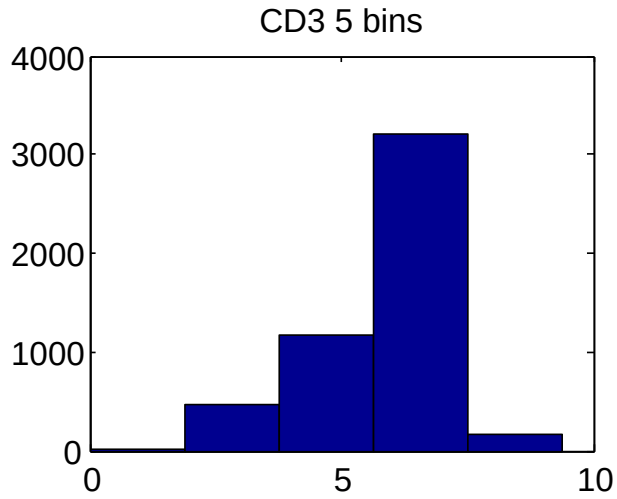
Immunologists want to differentiate between
healthy individuals from those with HIV^+.

Look at the (Log-Data)

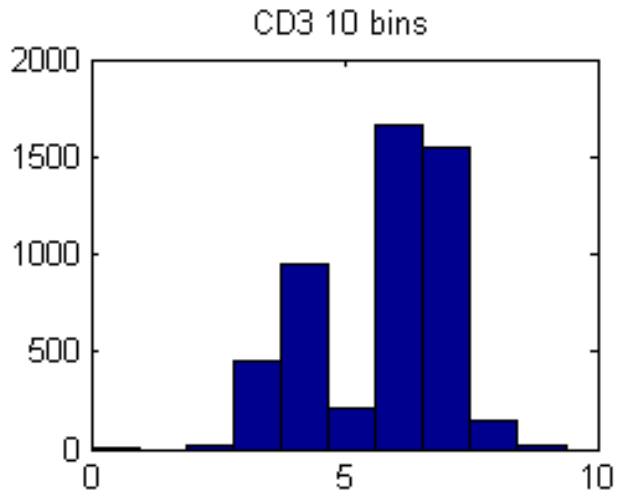


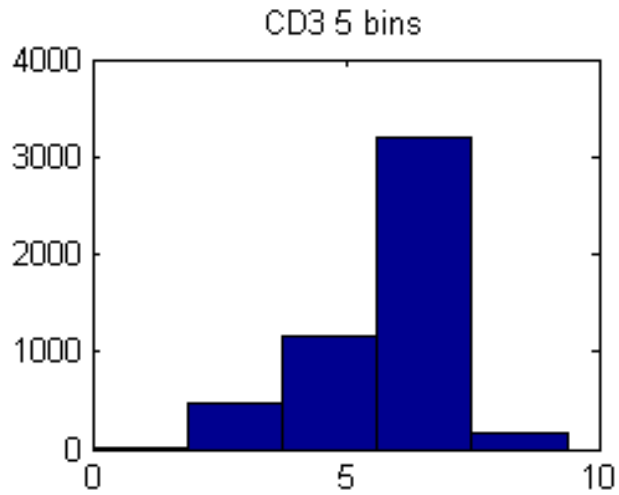
Histograms of the (Log-Data)

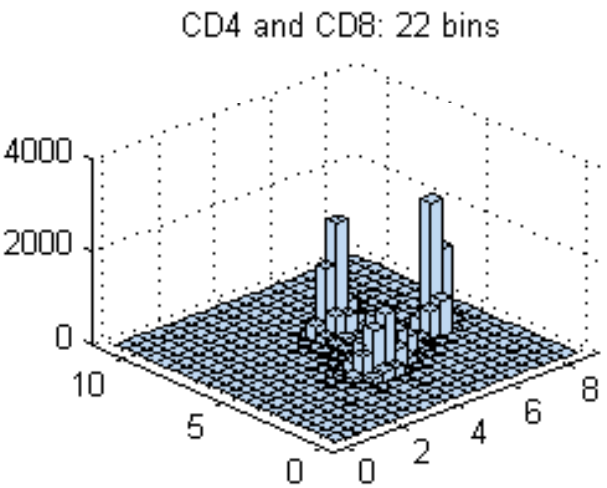


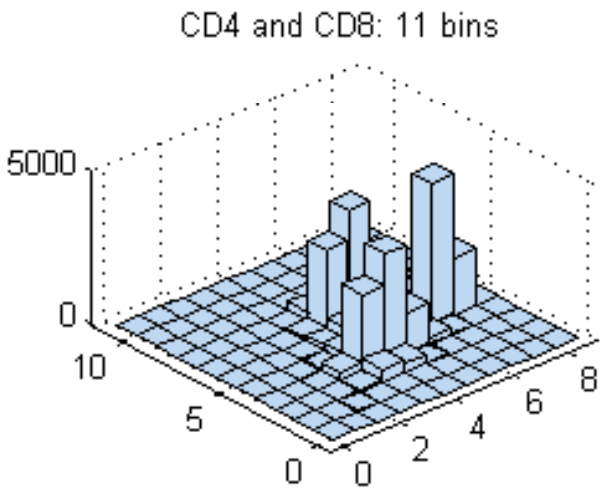


Histograms of the (Log-Data)









How Many Cluster are in the Data?

- One-dimensional data: 1 or 2 modes;
- Two-dimensional data: 1 to 3 or 4 modes;
- How many clusters are in the 12-dimensional data?

If the measurements were independent,
then the number of modes would be the product

$\rightarrow$ but this is not the case in our data

Can you think of a 3D example with k modes such that the 2D projections have $k-1$ modes?

Polynomial Histogram Estimators

Main idea

- histograms have **flat tops**, so instead of
- only estimating the number of points in each bin
- estimate the **shape** separately in each bin

What are Polynomial Histogram Estimators?

Number of observations n, dimension d, binwidth h
$B_\ell = h^d$ a bin with n_ℓ observations

The model for each bin B_ℓ

1. histogram estimators (Hist) $f_0(\mathbf{x}) = a_0$
2. first-order polynomial histogram estimator (Fophe)

$$f_1(\mathbf{x}) = a_0 + \mathbf{a}^T\mathbf{x}$$

3. second-order polynomial histogram estimator (Sophe)

$$f_2(\mathbf{x}) = a_0 + \mathbf{a}^T\mathbf{x} + \mathbf{x}^T A\mathbf{x}$$

Relationships for Coefficients

In each bin B_ℓ the estimate f_k satisfies

1. proportion of data

$$\int_{B_\ell} f_k(\mathbf{x}) d\mathbf{x} = \frac{n_\ell}{n}$$

2. local mean

$$\int_{B_\ell} \mathbf{x} f_k(\mathbf{x}) d\mathbf{x} = \frac{n_\ell}{n} \bar{\mathbf{x}}_\ell$$

3. local second moment

$$\int_{B_\ell} \mathbf{x}\mathbf{x}^T f_k(\mathbf{x}) d\mathbf{x} = \frac{n_\ell}{n} M_\ell$$

The New Estimators

In each bin B_ℓ with bin centre $\mathbf{t}_\ell$

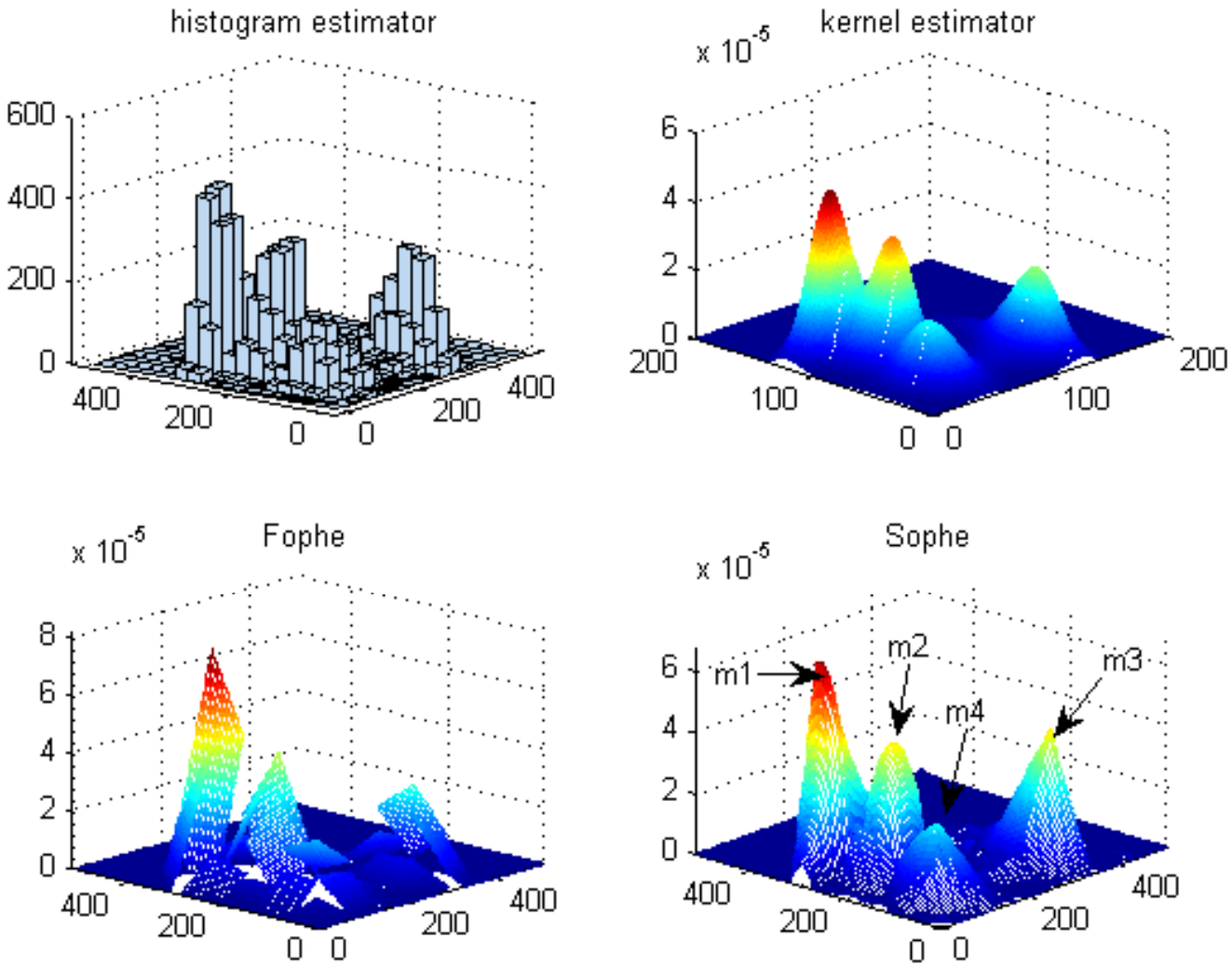
- Fophe

$$\widehat{f}_1(\mathbf{x}) = \frac{1}{h^{d+2}} \frac{n_\ell}{n} \left[h^2 + 12(\bar{\mathbf{x}}_\ell - \mathbf{t}_\ell)^T (\mathbf{x} - \mathbf{t}_\ell) \right]$$

- Sophe

$$\widehat{f}_2(\mathbf{x}) = \frac{1}{h^{d+4}} \frac{n_\ell}{n} \times$$
$$\left\{ \frac{(4+5d)}{4} h^4 - 15h^2 \ tr(S_\ell) + 12h^2 (\mathbf{x} - \mathbf{t}_\ell)^T (\bar{\mathbf{x}}_\ell - \mathbf{t}_\ell) \right.$$
$$\left. + (\mathbf{x} - \mathbf{t}_\ell)^T \left[72 S_\ell + 108 \ diag(S_\ell) - 15h^2 I \right] (\mathbf{x} - \mathbf{t}_\ell) \right\}.$$

Roederer Data: 10,000 observations, CD4 & CD8



The performance of estimators

We assess the performance of estimators with the MSE.

Let $\widehat{\theta}$ be an estimator for a true quantity θ. Then

$$\text{MSE}(\widehat{\theta}) = \left[\text{bias}(\widehat{\theta})\right]^2 + \text{var}(\widehat{\theta})$$

$$\text{bias}(\widehat{\theta}) = \mathbb{E}\widehat{\theta} - \theta$$

$$\text{var}(\widehat{\theta}) = \left\{\mathbb{E}\left[\widehat{\theta} - \mathbb{E}\widehat{\theta}\right]\right\}^2 = \mathbb{E}\left[\widehat{\theta}^2\right] - \left[\mathbb{E}\widehat{\theta}\right]^2$$

Sophe's Performance

For a fixed point $\mathbf{x} \in B_\ell$ we want the bias of $\widehat{f} = \widehat{f}_2$ at $\mathbf{x}$

Consider

$$\mathbb{E}\left[\widehat{f}(\mathbf{x})\right] = \mathbb{E}\left(\frac{1}{h^{d+4}}\frac{n_\ell}{n}\times \right.$$
$$\left\{\frac{(4+5d)}{4}h^4 - 15h^2\ tr(S_\ell) + 12h^2(\mathbf{x}-\mathbf{t}_\ell)^T(\bar{\mathbf{x}}_\ell - \mathbf{t}_\ell)\right.$$
$$\left.\left. + (\mathbf{x}-\mathbf{t}_\ell)^T\left[72S_\ell + 108\ diag(S_\ell) - 15h^2 I\right](\mathbf{x}-\mathbf{t}_\ell)\right\}\right)$$

Some Expectation Calculations I

We show that

$$\mathbb{E}\left[\frac{n_\ell}{n}(\bar{\mathbf{x}}_\ell - \mathbf{t}_\ell)\right] = \int_{B_\ell} (\mathbf{y} - \mathbf{t}_\ell) f(\mathbf{y}) d\mathbf{y}$$

and so

$$\mathbb{E}\left[\frac{12h^2}{h^{d+4}}(\mathbf{x} - \mathbf{t}_\ell)^T \frac{n_\ell}{n}(\bar{\mathbf{x}}_\ell - \mathbf{t}_\ell)\right] = \frac{12}{h^{d+2}}(\mathbf{x} - \mathbf{t}_\ell)^T \int_{B_\ell} (\mathbf{y} - \mathbf{t}_\ell) f(\mathbf{y}) d\mathbf{y}$$

then use a Taylor expansion of f about the bin centre $\mathbf{t}_\ell$

$$f(\mathbf{y}) = f(\mathbf{t}_\ell) + (\mathbf{y} - \mathbf{t}_\ell) Df(\mathbf{t}_\ell) + \frac{1}{2}(\mathbf{y} - \mathbf{t}_\ell)^2 D^2 f(\mathbf{t}_\ell)$$
$$+ \frac{1}{6}(\mathbf{y} - \mathbf{t}_\ell)^3 D^3 f(\mathbf{t}_\ell) + o\left(\|\mathbf{y} - \mathbf{t}_\ell\|^3\right)$$

Some Expectation Calculations II

The first non-zero integral gives

$$\mathbb{E}\left[\frac{12}{h^{d+2}}(\mathbf{x}-\mathbf{t}_\ell)^T\frac{n_\ell}{n}(\bar{\mathbf{x}}_\ell-\mathbf{t}_\ell)\right] \approx (\mathbf{x}-\mathbf{t}_\ell)^T Df(\mathbf{t}_\ell)$$

We prove similar results for all terms contributing to $\mathbb{E}\left[\widehat{f}(\mathbf{x})\right]$

. . . and finally get

The Bias

$$\mathbb{E}[\widehat{f}(\mathbf{x})] = f(\mathbf{t}_\ell) + (\mathbf{x}-\mathbf{t}_\ell)^T Df(\mathbf{t}_\ell) + \frac{1}{2}(\mathbf{x}-\mathbf{t}_\ell)^2 D^2 f(\mathbf{t}_\ell)$$
$$+ \frac{h^2}{12}(\mathbf{x}-\mathbf{t}_\ell)^T \left(\frac{\sum_i f_{uii}}{2} - \frac{f_{uuu}}{5} \right) + o(h^3)$$

Taylor expansion of f about the bin centre $\mathbf{t}_\ell$

$$f(\mathbf{x}) = f(\mathbf{t}_\ell) + (\mathbf{x}-\mathbf{t}_\ell) Df(\mathbf{t}_\ell) + \frac{1}{2}(\mathbf{x}-\mathbf{t}_\ell)^2 D^2 f(\mathbf{t}_\ell)$$
$$+ \frac{1}{6}(\mathbf{x}-\mathbf{t}_\ell)^3 D^3 f(\mathbf{t}_\ell) + o\left(\|\mathbf{x}-\mathbf{t}_\ell\|^3\right)$$

so bias$[\widehat{f}(\mathbf{x})]$ depends on difference of 3^{rd} order derivatives

Moving on . . .

and making some big leaps

We have the following steps in the performance calculations

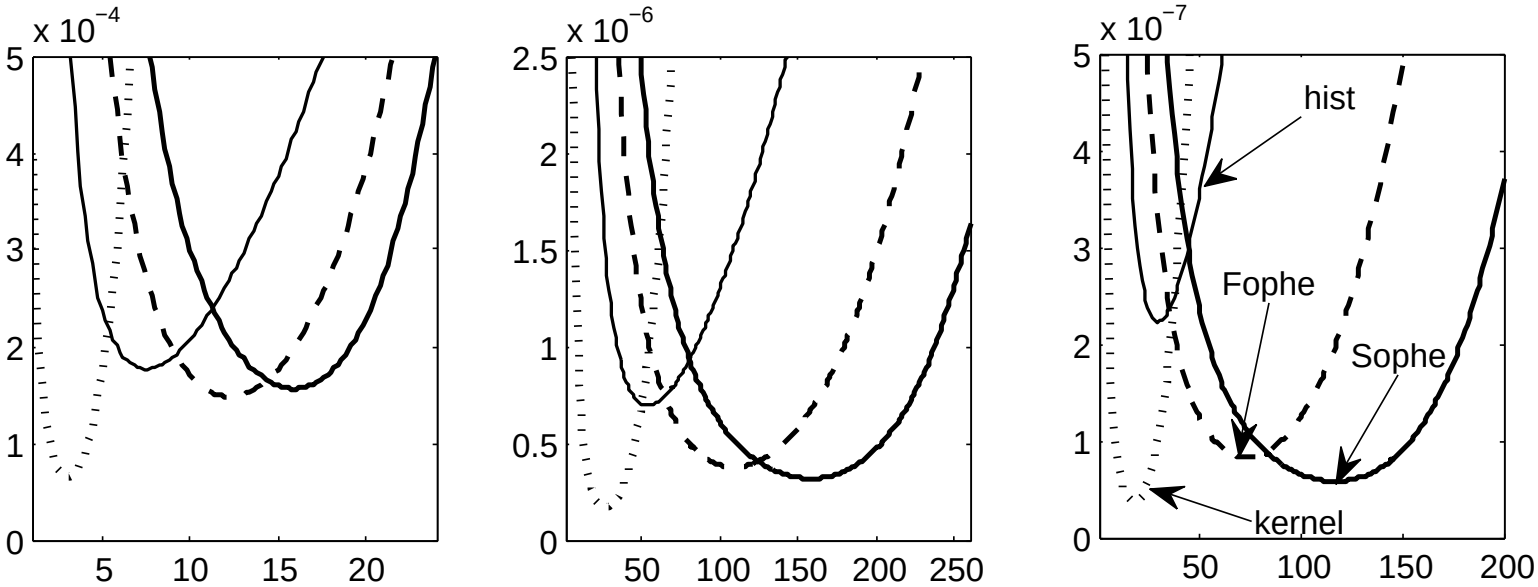
1. pointwise bias and variance $\rightarrow$ MSE at $\widehat{f}(\mathbf{x})$
2. integrated squared bias and integrated variance of $\widehat{f}$ over all $\mathbf{x}$
3. finally some asymptotics when $n \rightarrow \infty$

We want to know how Fophe and Sophe depend on the sample size n, the binwidth h, and the dimension d

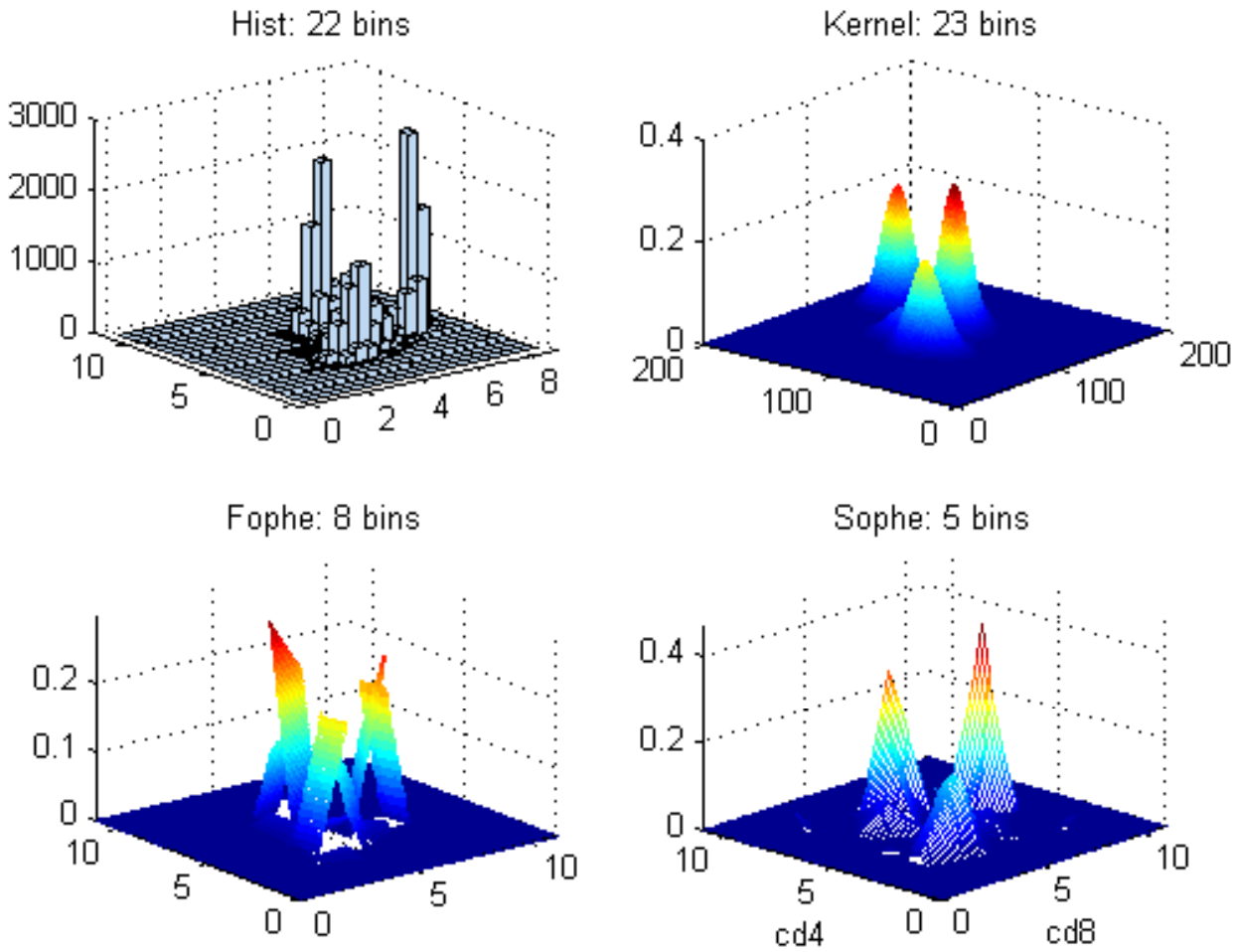
How Good are Fophe and Sophe

	$Bias^2$	Variance	Rate of Convergence
hist	$C_H h^2$	$\frac{1}{nh^d}$	$n^{-2/(d+2)}$
kernel	$C_K h^4$	$\frac{R(K)}{nh^d}$	$n^{-4/(d+4)}$
fophe	$C_F h^4$	$\frac{d+1}{nh^d}$	$n^{-4/(d+4)}$
sophe	$C_S h^6$	$\frac{(d+1)(d+2)}{2nh^d}$	$n^{-6/(d+6)}$

Performance for 200, 1000 and 10000 Observations



27,994 obs: Kernel est. takes 92× Sophe



Advantages of Fophe and Sophe

Computational advantages

1. a smaller number of bins is required
2. number of bins only needs to be approximately correct

Sophe better than Fophe in visual and computational aspects
$\rightarrow$ use **Sophe** for data

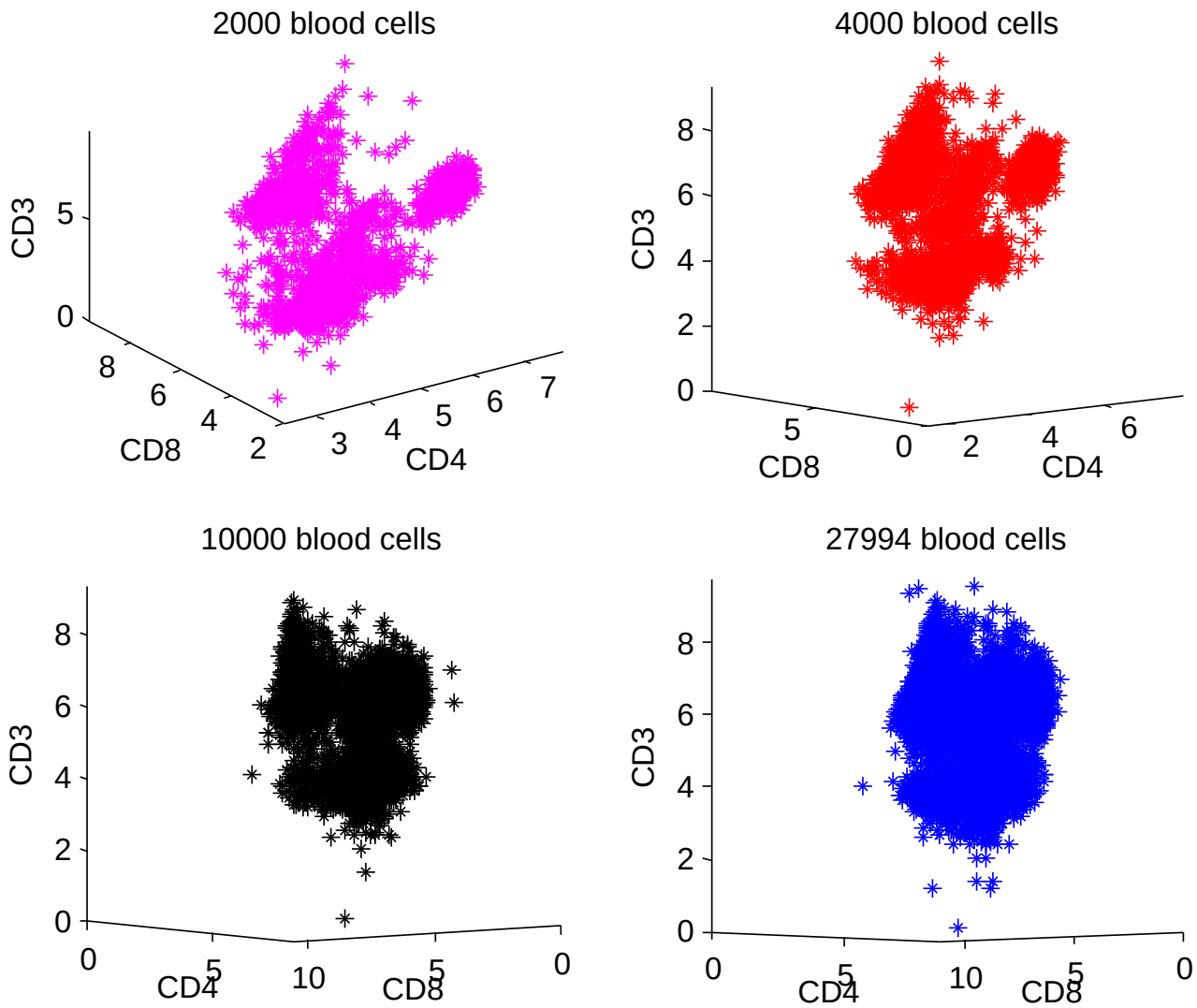
Finding Modes with the Sophe

1. Fix binwidth h_0, # of bins ν_{bin}, thresholds θ_0, and κ.
2. Find bins with high density.
   1. Find n_ℓ in each bin, and discard bins that contain fewer than θ_0 observations. Let $\mathcal{B}_0 = \{B_\ell \ : \ n_\ell > \theta_0\}$.
   2. Sort bins in $\mathcal{B}_0$ by # of observations, starting with largest.
3. Determine modes from $\mathcal{B}_0$ using (1) or (2) below.
   1. For $i, j = 1 \ldots, \kappa$ calculate pairwise distances $\Delta_{(i,j)}$ between the bin centres. For i consider the set of nearest neighbours

$$\mathfrak{nn}_{(i)} = \left\{(\Delta_{(i,j)}, \ n_{(j)}) \ : \Delta_{(i,j)} \leq h_0\right\}.$$

      $B_{(i)}$ contains a mode, if $n_{(i)}$ is maximum over $\mathfrak{nn}_{(i)}$.
   2. If matrix $A_{(j)}$ is negative definite, then $B_{(j)}$ contains a mode.

Look at the (Log-Data)



Modes for 12-Dimensional Data

Use 5 bins in each variable

compare # of modes and % of non-empty bins

# variables	# modes	# of bins	% non-empty
CDs 3,4,8	3	125	39.2
+ CDs 14, 19, 56	5	15625	2.6
all 12	9	244,140,625	0.0015

The End

J Jing, I Koch and K Naito (2009). Polynomial Histograms for Multivariate Density and Mode Estimation *preprint*.

Thank you