Applications of Bivariate Kernel Density Estimators

Tarn Duong
University of Western Australia

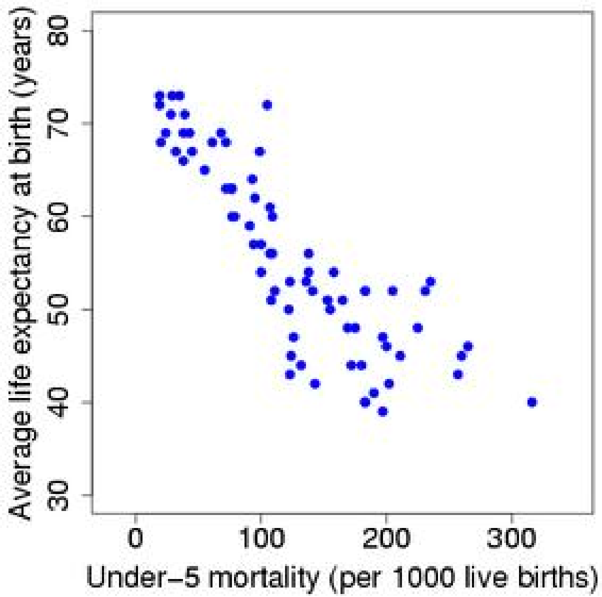
December 2003

Outline

1. Motivation
2. KDE basics
3. Application to real data
4. Application to simulated data
5. Extension to discriminant analysis
6. Summary

Motivation

UNICEF child mortality data



Properties of KDE

- non-parametric
- easy to compute
- easy to interpret

Equation for KDE

Kernel density estimate is

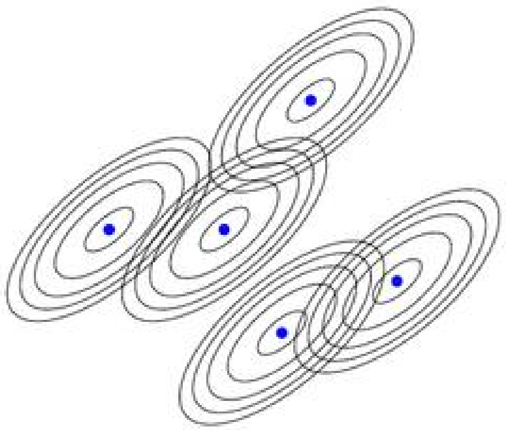
$$\hat{f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$$

where

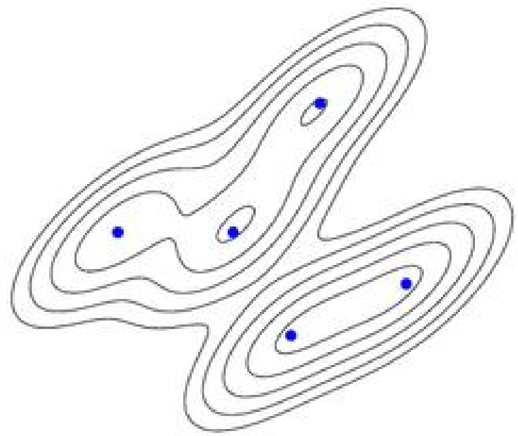
- $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is random sample of n bivariate data points
- $\mathbf{H}$ is **bandwidth** matrix parameter which is estimated from data
- $K_{\mathbf{H}}(\cdot)$ is normal pdf with mean $\mathbf{0}$ and variance $\mathbf{H}$

Constructing KDE

Individual kernels



Averaged kernels = KDE

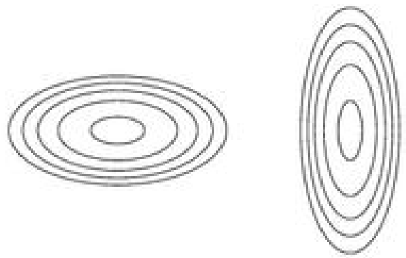


Bandwidth selection

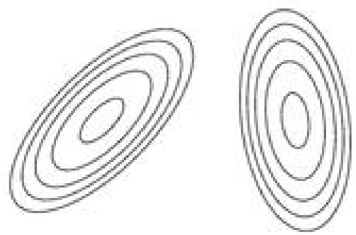
- single most important factor affecting performance of KDE
- induce orientation of kernel
- control spread of kernel
- diagonal bandwidth $\begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix}$ or full bandwidth $\begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}$

Kernel orientation

Diagonal bandwidth matrix

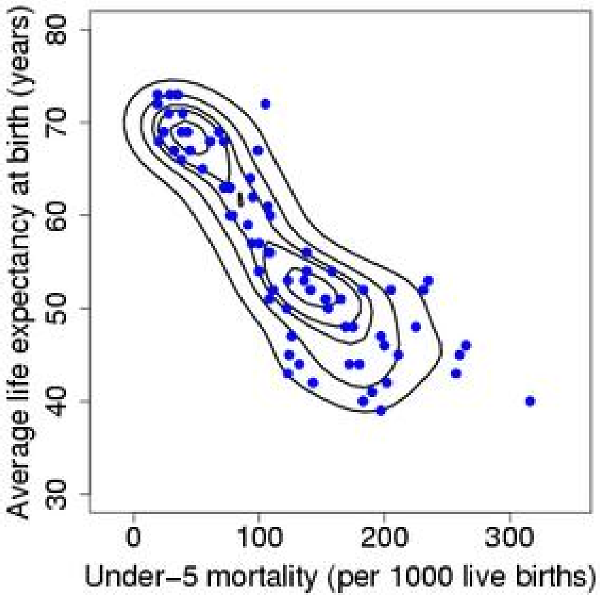


Full bandwidth matrix

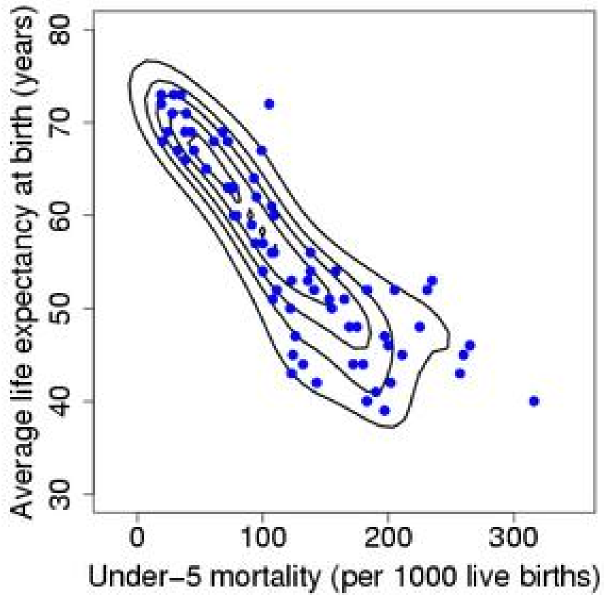


KDE of UNICEF data

Diagonal bandwidth matrix

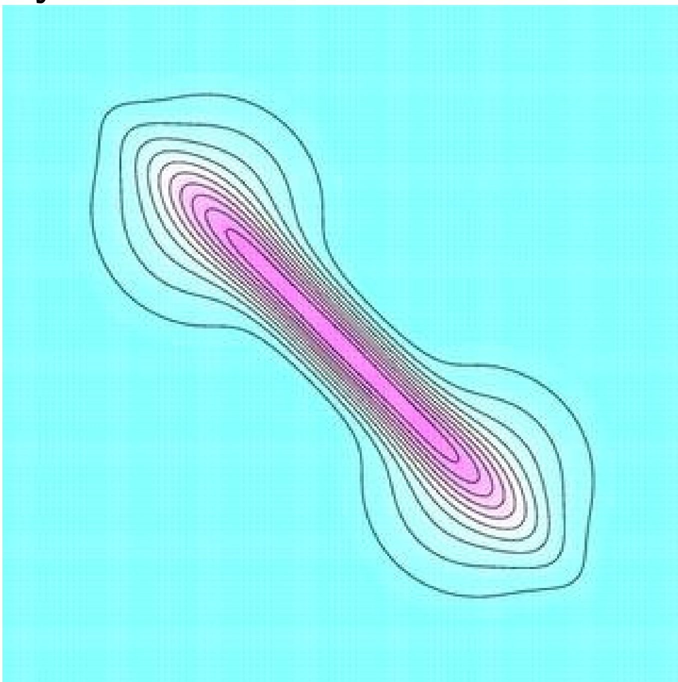


Full bandwidth matrix



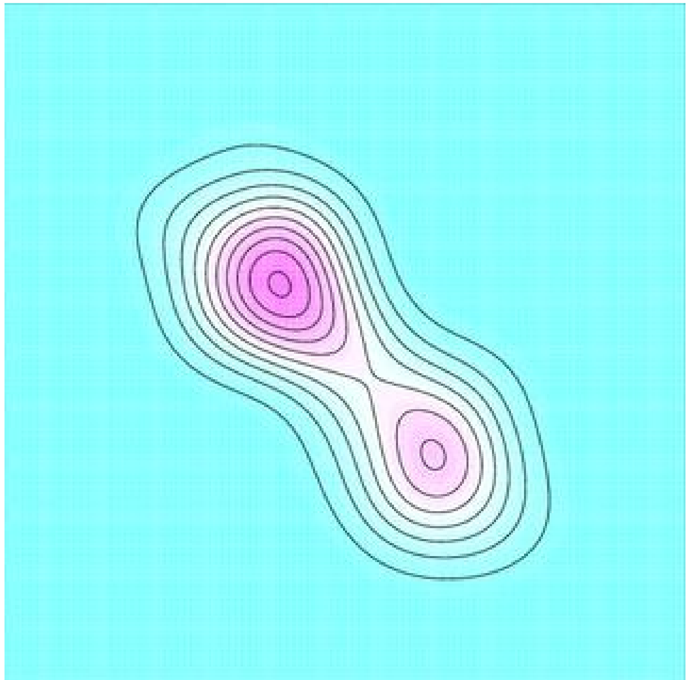
Dumbbell density

Known density with similar structure to UNICEF data

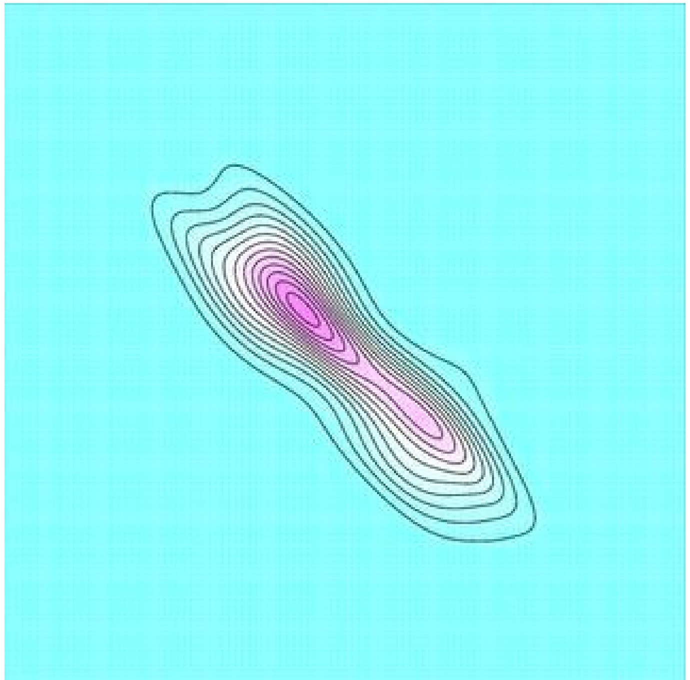


KDE of simulated dumbbell density

Diagonal bandwidth matrix



Full bandwidth matrix

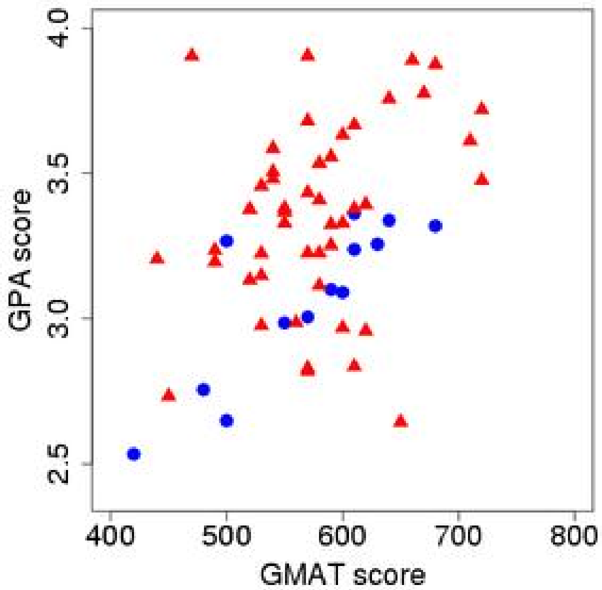


Non-parametric discriminant analysis (1)

- Two training sets
  - $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n_1}$ drawn from density f_1 with prior prob π_1
  - $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{n_2}$ drawn from density f_2 with prior prob π_2
- Want to classify $\mathbf{z}$ to group 1 or 2
- Compute KDE of f_1 and f_2 and decide that
  - $\mathbf{z}$ belongs to group 1 if $\pi_1 \hat{f}_1(\mathbf{z}; \mathbf{H}_1) > \pi_2 \hat{f}_2(\mathbf{z}; \mathbf{H}_2)$
  - $\mathbf{z}$ belongs to group 2 otherwise

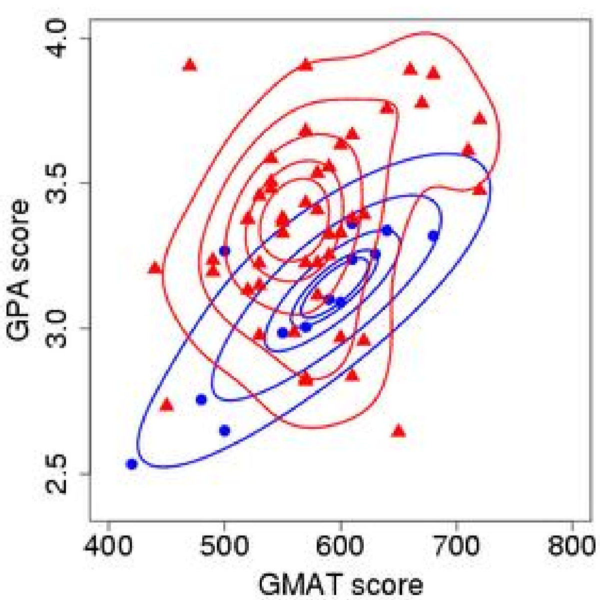
Non-parametric discriminant analysis (2)

NYU student data

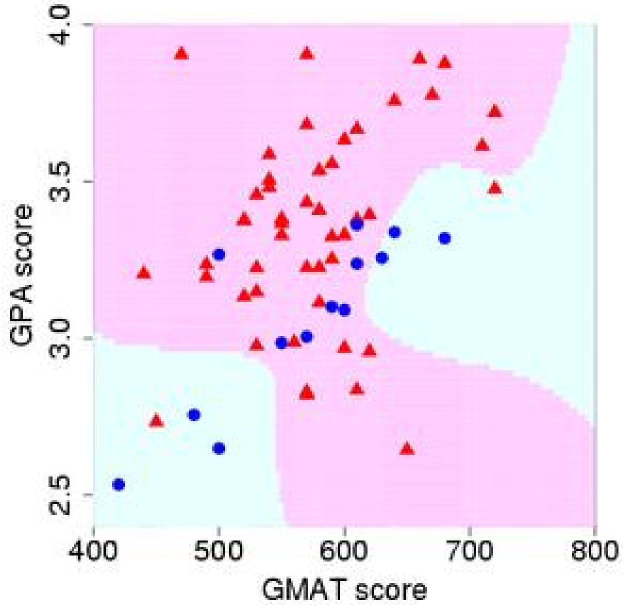


Non-parametric discriminant analysis (3)

Density estimates



Classification regions



Summary

- KDE is easy to use, easy to interpret estimation technique
- more research for multivariate KDE with full bandwidth matrices
- have shown some promising results so far
- KDE is useful in its own right and also for extensions e.g. non-parametric discriminant analysis