Software for multivariate kernel smoothing

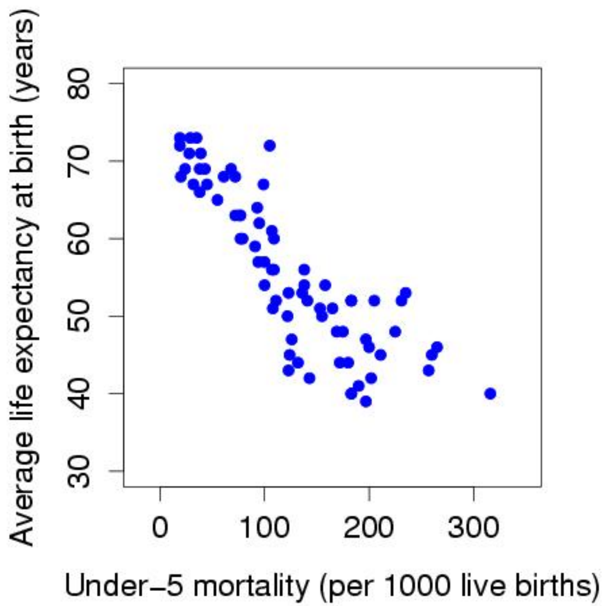
Tarn Duong
Department of Statistics
University of New South Wales

September 2005

Outline

- Kernel density estimation
  - Motivating example
  - Bandwidth selection
  - Software
- Kernel discriminant analysis
  - Motivating example
  - Software
- Extensions & summary

Motivation – density estimation



Given data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are drawn from density f, what is plausible estimate $\hat{f}$?

Existing R libraries for kernel smoothing

- KernSmooth (Wand & Jones) - up to bivariate, restricted bandwidths

- sm (Bowman & Azzalini) - up to trivariate, restricted bandwidths

- ash (Scott) - up to bivariate, restricted bandwidths

ks - R library for kernel smoothing

- kernel density estimation
- kernel discriminant analysis
- 2 – 6 dimensional data
- general bandwidths

Equation for KDE

Kernel density estimate (KDE) is

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

where

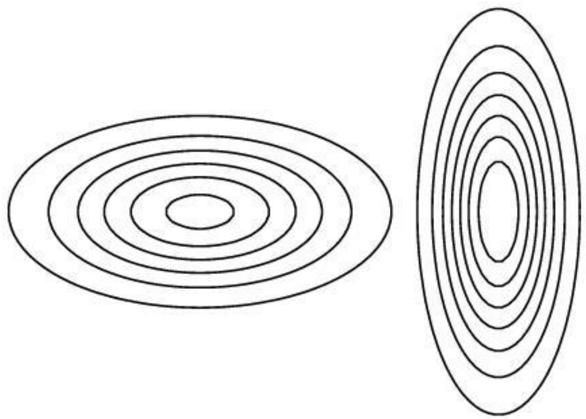
- $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is random sample of n d-variate data points
- $\mathbf{H}$ is **bandwidth** matrix parameter
- $K_{\mathbf{H}}(\cdot)$ is normal pdf with mean $\mathbf{0}$ and variance $\mathbf{H}$

Bandwidths

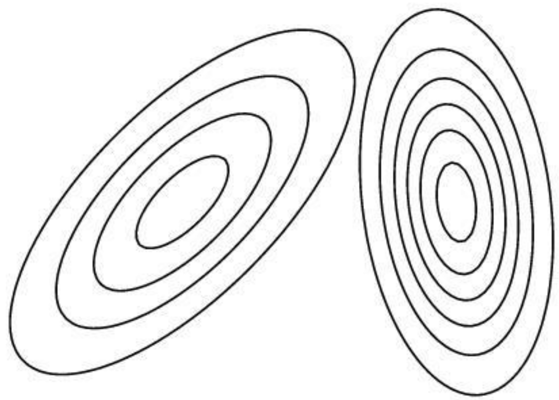
- induces orientation of kernel
- controls spread of kernel
- restricted (diagonal) bandwidth $\begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix}$
- general (full) bandwidth $\begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}$

Restricted vs. general bandwidths

Restricted (diagonal) bandwidth

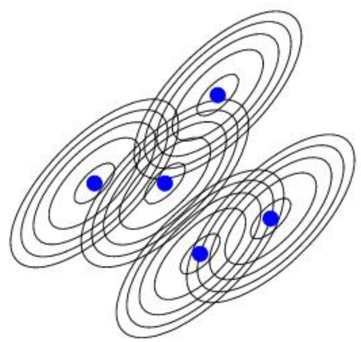


General (full) bandwidth

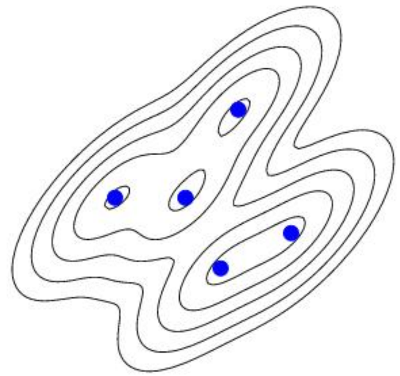


Constructing KDE

Individual (scaled) kernels



Summed (scaled) kernels = KDE



Error measures

Mean Integrated Squared Error (MISE) is

$$\mathrm{MISE}(\mathbf{H}) = \int_{\mathbb{R}^d} \mathbb{E}[\hat{f}(\boldsymbol{x};\mathbf{H}) - f(\boldsymbol{x})]^2 \, d\boldsymbol{x}$$

where

- $\hat{f}(\boldsymbol{x};\mathbf{H})$ is kernel density estimate
- $f(\boldsymbol{x})$ is (unknown) target density

Bandwidth selection

- Ideal: $\mathbf{H}_{\text{MISE}} = \underset{\mathbf{H}}{\operatorname{argmin}} \; \text{MISE}(\mathbf{H})$
- not tractable so need alternative approach
- several flavours of estimation

Bandwidth selection – plug-in

- Plug-in estimate of MISE

$$\mathrm{PI}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + \tfrac{1}{4}\mu_2(K)^2(\mathrm{vech}^T\,\mathbf{H})\hat{\mathbf{\Psi}}_4(\mathbf{G})(\mathrm{vech}\,\mathbf{H})$$

where $\mathbf{G}$ is pilot (or 'helper') bandwidth matrix

- Plug-in selector $\hat{\mathbf{H}}_{\mathrm{PI}} = \underset{\mathbf{H}}{\mathrm{argmin}}\ \mathrm{PI}(\mathbf{H})$

Bandwidth selection – cross validation (1)

- Least squares cross validation (LSCV) estimate of MISE

$$\mathrm{LSCV}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + [n(n-1)]^{-1}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{2\mathbf{H}} - 2K_{\mathbf{H}})(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

- LSCV selector $\hat{\mathbf{H}}_{\mathrm{LSCV}} = \underset{\mathbf{H}}{\mathrm{argmin}}\ \mathrm{LSCV}(\mathbf{H})$

Bandwidth selection – cross validation (2)

- Smoothed cross validation (SCV) estimate of MISE

$$\mathrm{SCV}(\mathbf{H}) = n^{-1}|\mathbf{H}|^{-1/2}R(K) + [n(n-1)]^{-1} \times \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}(K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

where $\mathbf{G}$ is pilot ('helper') bandwidth matrix

- SCV selector is $\hat{\mathbf{H}}_{\mathrm{SCV}} = \underset{\mathbf{H}}{\operatorname{argmin}}\ \mathrm{SCV}(\mathbf{H})$

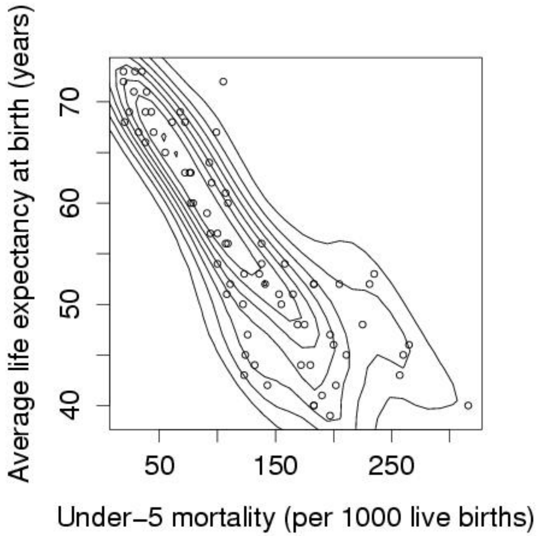
Sample code

- `H <- Hpi(unicef)`
  - plug-in selector
  - unicef data is original motivating data

- `fhat <- kde(unicef, H)`
  - kernel density estimate

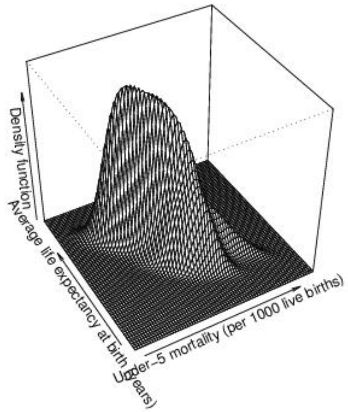
- `plot(fhat, display="slice")`
  - contour or 'slice' plot

Sample output



Sample code and output (2)

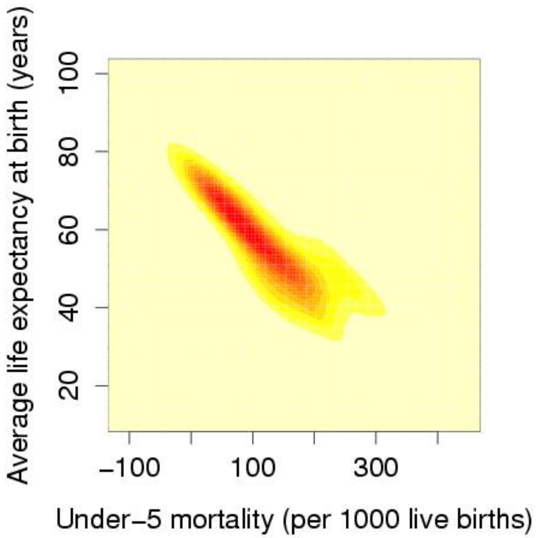
- `plot(kde(unicef, Hlscv(unicef)), display="persp")`
  - least squares cross validation selector
  - perspective or 'wire-frame' plot



–

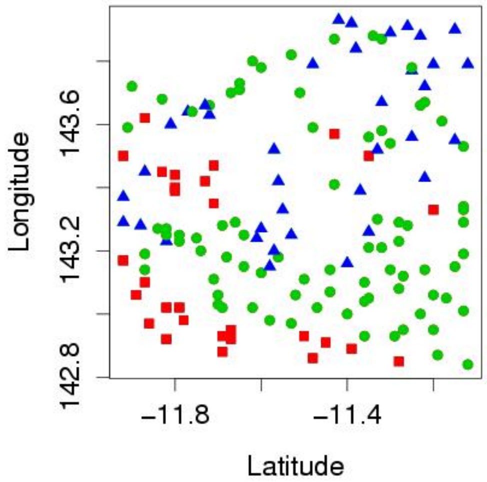
Sample code and output (3)

- `plot(kde(unicef, Hscv(unicef)), display="image")`
  - smoothed cross validation selector
  - image or ‘heat’ plot



–

Motivation – discriminant analysis



Given training data $\{\boldsymbol{X}_{11}, \ldots \boldsymbol{X}_{1n_1}\}$ drawn from f_1, . . . , $\{\boldsymbol{X}_{\nu 1}, \ldots \boldsymbol{X}_{\nu n_1}\}$ drawn from f_ν, what is plausible rule for classifying test data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ to one of groups 1, . . . ν?

Discriminant analysis

Bayes discriminant rule:

$$\boldsymbol{x} \in \text{group } j_0 \text{ if } j_0 = \underset{j \in 1,\ldots,\nu}{\operatorname{argmax}} \, \pi_j f_j$$

where π_j is prior prob. of drawing from density f_j.

Kernel discriminant analysis

Kernel discriminant rule:

$$\boldsymbol{x} \in \text{group } j_0 \text{ if } j_0 = \underset{j \in 1,\ldots,\nu}{\operatorname{argmax}}\ \hat{\pi}_j \hat{f}_j$$

where

- $\hat{f}_j$ is kernel density estimate of f_j based on j-th training data sample $\boldsymbol{X}_{j1}, \ldots, \boldsymbol{X}_{jn_j}$
- $\hat{\pi}_j$ is estimate of π_j (either subjective or sample proportion)

Sample code (4)

- `H <- Hkda(trw, bw="plugin")`
  - plug-in bandwidths for kernel discriminant analysis
  - trw is trawler data from motivating example

- ```
  fhat <- kda.kde(trw, trw.gr, H)
  plot(fhat)
  ```
  - partition based on kernel discriminant rule
  - trw.gr is actual group labels
  ```

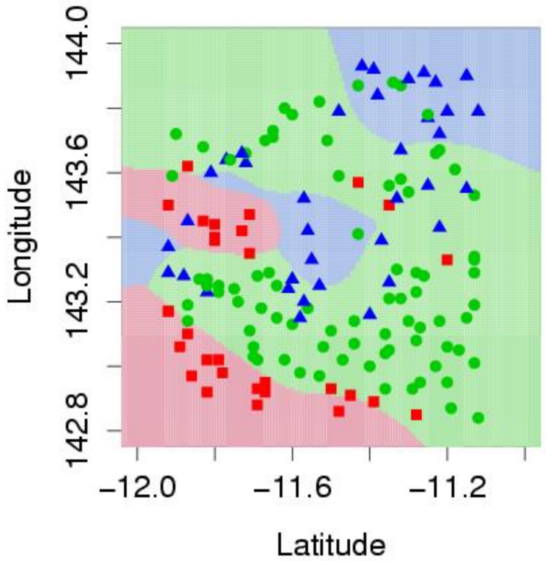
Sample code (5)

- ```
  fhat <- pda.pde(trw, trw.gr, type="quad")
  plot(fhat)
  ```
  - partition based on quadratic discriminant rule

- ```
  compare.kda.cv(trw, trw.gr, bw="plugin")
  compare.pda.cv(trw, trw.gr, type="quad")
  ```
  - cross-validated estimate of misclass. rate since test data = training data

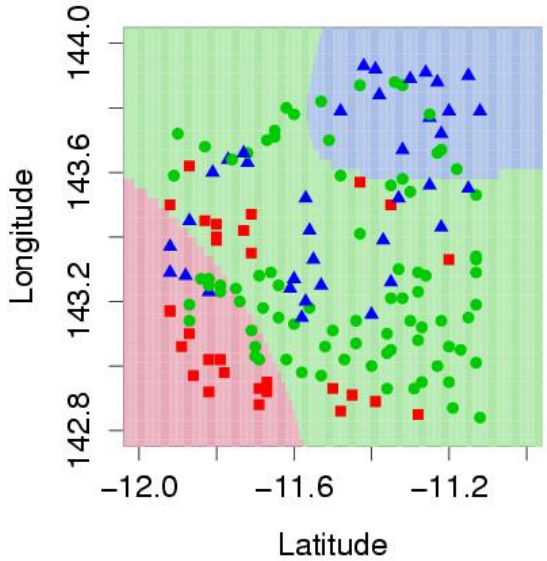
Sample output (6)

Kernel



misclass. rate = 0.309

Quadratic



misclass. rate = 0.430

References

- Plug-in bandwidth matrices for bivariate kernel density estimation, *Journal of Nonparametric Statistics* (2003) **15**, 17–30
- Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation, *Journal of Multivariate Analysis* (2005) **93**, 417–433
- Cross validation bandwidth matrices for multivariate kernel density estimation, *Scandinavian Journal of Statistics* (2005) **32**, 485-506
- CRAN `http://cran.r-project.org` – ks v1.3.2

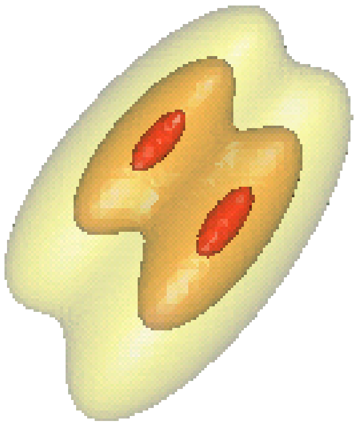
Current development – v1.3.3 (1)

Trivariate data example:

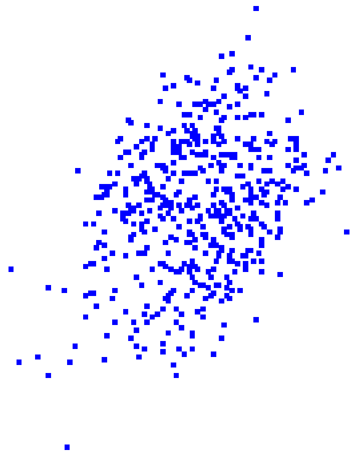
$$f \sim \frac{1}{2} N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix} \right) + \frac{1}{2} N \left(\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix} \right)$$

Current development – v1.3.3 (2)

True density

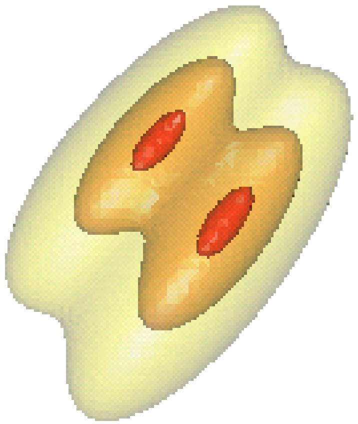


Sample of 500 points

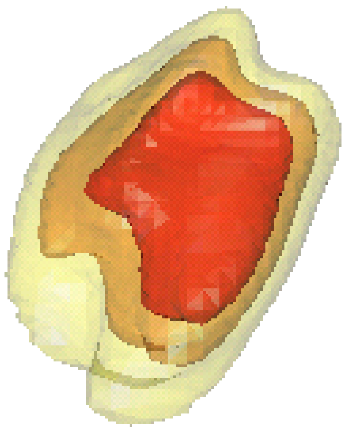


Current development – v1.3.3 (3)

True density



Kernel density estimate (plug-in)



Extensions

- other smoothing problems e.g. regression, distribution function estimation, hazard function estimation etc.
- improve computational efficiency

Summary

Current features of ks:

- general bandwidth selectors (plug-in and cross-validation)
- kernel density estimation and discriminant analysis
- 2- to 6-dimensional data
- graphics for 2- or 3-dimensions