A tour of kernel smoothing

Tarn Duong

Institut Pasteur

October 2007

The journey up till now

- 1995–1998 Bachelor, Univ. of Western Australia, Perth
- 1999–2000 Researcher, Australian Bureau of Statistics, Canberra and Sydney
- 2001–2004 PhD, Univ. of Western Australia, Perth
- 2005 Lecturer, Macquarie Univ., Sydney
- 2005–2007 Post-doc, Univ. of New South Wales, Sydney
- 2007– present Post-doc, Institut Pasteur, Paris

Research interests

- Kernel smoothing
- Nonparametric statistics
- Statistical software

Today

- Kernel density estimation (KDE)
  - 1st stage of inference (estimation)
  - translation is Éstimation de densité à noyau
- Feature significance
  - 2nd stage of inference (formal inference)
  - translation is ?
  - extension of density estimation to significance testing

Kernel (1)

- NOT cell nucleus
- NOT kernel of an operating system
- NOT kernel/nullspace of a matrix $\mathbf{A}$: $\{\boldsymbol{x} : \mathbf{A}\boldsymbol{x} = \mathbf{0}\}$

Kernel (2)

Kernel $K : \mathbb{R}^d \to \mathbb{R}$ is

- $K(\boldsymbol{x}) \geq 0$
- $\int_{\mathbb{R}^d} K(\boldsymbol{x}) \, d\boldsymbol{x} = 1$
- K is symmetric about $\mathbf{0}$

Kernel density estimation

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample drawn from a common density f. A kernel density estimate $\hat{f}$ is

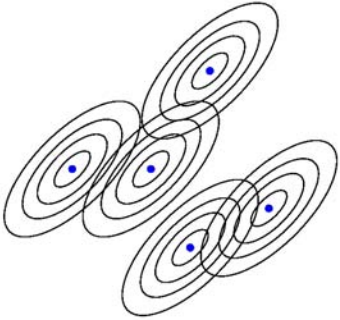
$$\hat{f}(\boldsymbol{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$$

where

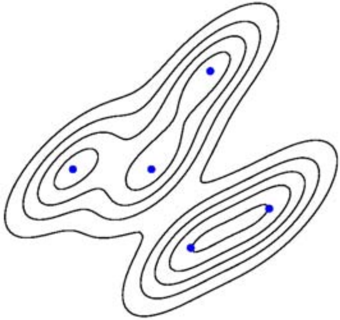
$$\begin{aligned} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) &= \text{normal (Gaussian) pdf with mean } \boldsymbol{X}_i, \text{variance } \mathbf{H} \\ \mathbf{H} &= \text{bandwidth or window width (fenetre)} \end{aligned}$$

Graphical illustration

Scaled kernels $K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i)$



Kernel density estimate $\hat{f}$



Advantages of kernel density estimates

- non-parametric
- easy to construct
- easy to interpret
- suitable for multivariate data
- smooth, no discretisation effects
- no anchor points effects

Bandwidth selectors

- single most important factor effecting performance of $\hat{f}$
- ideal bandwidth selector: $\mathbf{H}_0 = \operatorname*{argmin}_{\mathbf{H}} \mathrm{AMISE}(\mathbf{H})$

  where AMISE = asymptotic $\int_{\mathbb{R}^d} \mathbb{E}[\hat{f}(\boldsymbol{x};\mathbf{H}) - f(\boldsymbol{x})]^2 \, d\boldsymbol{x}$
- data-driven selector: $\hat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{H}} \widehat{\mathrm{AMISE}}(\mathbf{H})$

Relative convergence rates (1)

- a data-driven selector $\hat{\mathbf{H}} = \underset{\mathbf{H}}{\operatorname{argmin}}\, \widehat{\mathrm{AMISE}}(\mathbf{H})$ converges to $\mathbf{H}_0$ with rate $n^{-\alpha}, \alpha > 0$ if

$$\operatorname{vech}(\hat{\mathbf{H}} - \mathbf{H}_0) = O_p(n^{-\alpha}\mathbf{J}) \operatorname{vech} \mathbf{H}_0$$

where O_p is order in probability, $\mathbf{J}$ = matrix of ones, and

$$\operatorname{vech} \begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Relative convergence rates (2)

- $\hat{\mathbf{H}}$ converges to $\mathbf{H}_0$ with rate $n^{-\alpha}$ if

$$\begin{aligned}\text{MSE}(\hat{\mathbf{H}}) &= \text{Var}(\hat{\mathbf{H}}) + \text{Bias}(\hat{\mathbf{H}})\,\text{Bias}^T(\hat{\mathbf{H}}) \\ &= O(n^{-2\alpha})(\text{vech}\,\mathbf{H}_0)(\text{vech}^T\,\mathbf{H}_0)\end{aligned}$$

Relative convergence rates (3)

Easier(?!) to compute

$$\mathrm{Bias}(\hat{\mathbf{H}}) = O\left(\mathbb{E}\left[\frac{\partial}{\partial \operatorname{vech} \mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_0)\right]\right)$$

$$\mathrm{Var}(\hat{\mathbf{H}}) = O\left(\mathrm{Var}\left[\frac{\partial}{\partial \operatorname{vech} \mathbf{H}}(\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_0)\right]\right)$$

Table of convergence rates

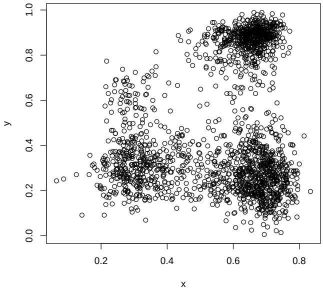
	Convergence rate	
Selector	$d = 1$	$d > 1$
Plug-in 1 (1994)	$n^{-4/13}$	$n^{-4/(d+12)}$
Plug-in 2 (2003)	$n^{-2/7}$	$n^{-2/(d+6)}$
CV 1 (1982, 1984)	$n^{-1/10}$	$n^{-min(d,4)/(2d+8)}$
CV 2 (1994)	$n^{-1/10}$	$n^{-min(d,4)/(2d+8)}$
CV 3 (1992, 2004)	$n^{-5/14}$	$n^{-2/(d+6)}$

Software

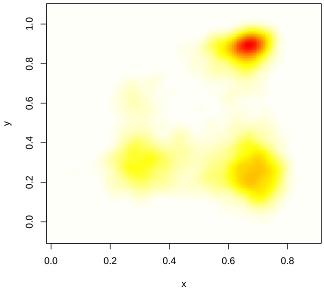
- ks: R library available on CRAN `www.r-project.org`
- comprehensive package for kernel density estimation and bandwidth selection

Flow cytometry (FACS) data (1)

Data sample

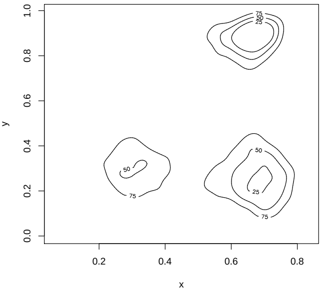


KDE

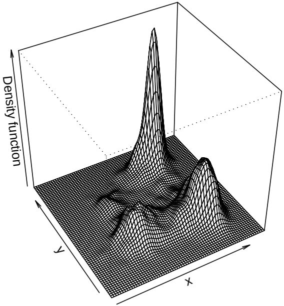


Flow cytometry (FACS) data (2)

Contour plot



Wireframe plot



Independent citations in other fields

- Zago, A. and Dongili P. (2006) Bad loans and efficiency in Italian banks, Working paper no. 28, Università di Verona
- Fieberg, J. (2007) Kernel density estimators of home range: smoothing and the autocorrelation red herring. *Ecology*, **88**, 1059–1066
- Peng T.G., Wang Y.H. and Wu T.H. (2007) Mean shift algorithm equipped with the intersection of confidence intervals rule for image segmentation. *Pattern Recognition Letters*, **28**, 268–277

Features

- $d = 1, 2$: mode, valley, saddle-point, ridge etc.
- $d > 2$: mode

Modes and modal regions

- mode $\boldsymbol{x}^*$ of function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$
  - $\mathrm{D}\, f(\boldsymbol{x}^*) = \mathbf{0}, \mathrm{D}^2\, f(\boldsymbol{x}^*) < 0$
  - $\mathrm{D}\, f(\boldsymbol{x}^*) = \mathbf{0}$, eigenvalues $\lambda_1(\boldsymbol{x}^*), \lambda_2(\boldsymbol{x}^*), \ldots, \lambda_d(\boldsymbol{x}^*)$ of $\mathrm{D}^2\, f(\boldsymbol{x}^*) < 0$
- modal region M of f
  - $M = \{\boldsymbol{x} : \|\mathrm{D}\, f(\boldsymbol{x})\| \leq \delta, -\varepsilon \leq \lambda_j(\boldsymbol{x}) \leq 0\}$
  - δ, ε 'small' positive

Kernel density derivative estimation

- density (zero-th derivative):

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

- gradient (first derivative):

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

- curvature (second derivative):

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

Kernel curvature estimators

- asymptotic distribution:
  $\operatorname{vech} \widehat{\mathsf{D}^2 f}(\boldsymbol{x}; \mathbf{H}) \overset{\text{approx.}}{\sim} N(\operatorname{vech} \mathsf{D}^2 f(\boldsymbol{x}), \Sigma(\boldsymbol{x}))$
- local null hypothesis: $H_0(\boldsymbol{x}) : \operatorname{vech} \mathsf{D}^2 f(\boldsymbol{x}) = \mathbf{0}$
- null distribution: $\operatorname{vech} \widehat{\mathsf{D}^2 f}(\boldsymbol{x}; \mathbf{H}) \overset{\text{approx.}}{\sim} N(\mathbf{0}, \Sigma(\boldsymbol{x}))$
- test statistic:
  $W(\boldsymbol{x}) = \|\Sigma(\boldsymbol{x})^{-1/2} \operatorname{vech} \widehat{\mathsf{D}^2 f}(\boldsymbol{x}; \mathbf{H})\|^2 \overset{\text{approx.}}{\sim} \chi^2_{d(d+1)/2}$

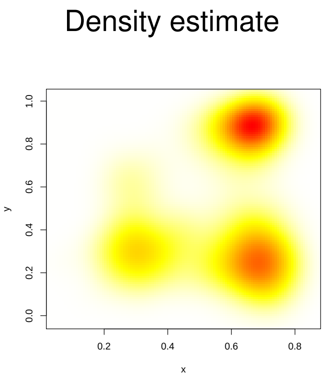
Significant curvature regions

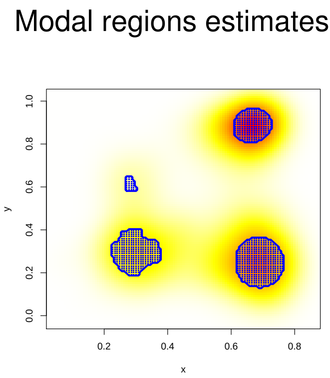
- extension of kernel density estimation suited to finding modal regions
- modal region estimate at significance level α: significant curvature region $\hat{M} = \{\boldsymbol{x} : W(\boldsymbol{x}) \geq \chi^2_{d(d+1)/2;1-\alpha'}\}$
- α' is adjusted significance level to account for multiple hypothesis tests

Software

- feature: R library available on CRAN

Flow cytometry (FACS) data (3)





Summary

- Multivariate kernel density estimators
  - theoretical development of optimal bandwidth selectors
  - software implementation
- Feature significance
  - some theoretical development of multivariate modal region estimation
  - software implementation

Future directions

- Comparing two kernel density estimators
- Optimal bandwidth selection for kernel density derivative estimators

Acknowledgements

- Kernel density estimation
  - Prof. Martin Hazelton, then Univ. of Western Australia, now at Massey Univ. (New Zealand), as PhD supervisor
- Feature significance
  - Dr Inge Koch, Univ. of New South Wales (Australia),
  - Prof. Matt Wand, then Univ. of New South Wales (Australia), now at Univ. of Wollongong (Australia)