Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation

Tarn Duong[1], Martin Hazelton[2]

[1]University of New South Wales, Sydney, Australia

[2]Massey University, Palmerston North, New Zealand

August 2006

Kernel density estimator

Kernel density estimate $\hat{f}$ of target density f is

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

where

- $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \sim f$ is d-dim. random sample of size n
- $\mathbf{H}$ is **bandwidth** matrix
- $K_{\mathbf{H}}(\cdot)$ is normal pdf with mean $\mathbf{0}$, variance $\mathbf{H}$

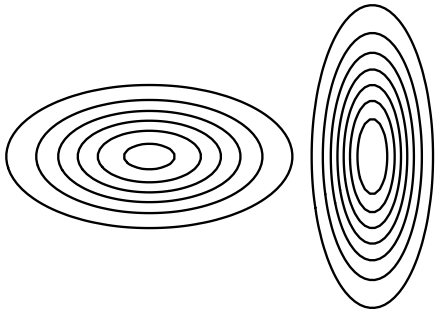
Bandwidth matrix

- single most important factor affecting performance of KDE
- induces orientation of kernel
- controls spread of kernel

Bandwidth matrix parameterisation

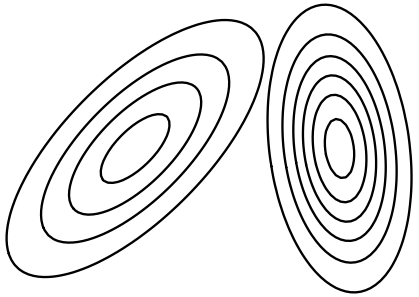
Diagonal bandwidth

$$\begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix}$$



Full bandwidth

$$\begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}$$



Convergence rate

- data-driven bandwidth $\hat{\mathbf{H}}$
- ideal bandwidth $\mathbf{H}_{\text{ideal}}$
- $\hat{\mathbf{H}} \to \mathbf{H}_{\text{ideal}}$ at rate $n^{-\alpha}, \alpha > 0$.

Convergence rates as at Jan 2001

Selector	Rate
Plug-in (diag)	$n^{-\min(8,d+4)/(2d+12)}$
Plug-in (full)	$n^{-4/(d+12)}$
Unbiased/Biased CV (diag)	$n^{-d/(2d+8)}$
Smoothed CV ($d = 1$)	$n^{-(d+4)/(2d+12)}$

- Plug-in: Wand & Jones (1994)
- Unbiased/Biased CV : Sain, Baggerly & Scott (1994)
- Smoothed CV ($d = 1$ only): Hall, Marron & Park (1992)

Comparison of convergence rates

Let $d = 2$ then

- plug-in rate (diag) is $n^{-3/8}$
- UCV/BCV rates (diag) are $n^{-1/6}$
- to achieve order = 0.1, plug-in: $n = 10^{8/3} \approx 464$
- to achieve order = 0.1, CV: $n = 10^6$
- ratio $\approx$ 5200

Summary of results

- Convergence rate for full SCV selector
- Convergence rate for improved full plug-in selector
- Convergence rates for full BCV and UCV selectors
- Finite sample behaviour

Bandwidth selectors

- $\mathrm{MISE}(\mathbf{H}) = \mathbb{E}\left[\int_{\mathbb{R}^d} (\hat{f}(\boldsymbol{x}; \mathbf{H}) - f(\boldsymbol{x}))^2 \, d\boldsymbol{x}\right]$
- ideal bandwidth $\mathbf{H}_{\mathrm{MISE}} = \operatorname*{argmin}_{\mathbf{H}} \mathrm{MISE}(\mathbf{H})$
- $\mathrm{AMISE}(\mathbf{H}) = \text{ Asymptotic } \mathrm{MISE}(\mathbf{H})$
- proxy ideal $\mathbf{H}_{\mathrm{AMISE}} = \operatorname*{argmin}_{\mathbf{H}} \mathrm{AMISE}(\mathbf{H})$
- data-driven bandwidth $\hat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{H}} \widehat{\mathrm{AMISE}}(\mathbf{H})$

Order in probability (univariate)

- Let $\{a_n\}, \{b_n\}$ be sequence of numbers. Then

$$a_n = O(b_n) \text{ if } \exists M, L \, : \, n > M \Rightarrow |a_n| < L|b_n|$$

- Let $\{A_n\}$ be sequence of random variables. Then

$$A_n = O_p(b_n) \text{ if } \forall \epsilon > 0 \, \exists M, L \, : n > M \Rightarrow \mathbb{P}(|A_n| < L|b_n|) > 1-\epsilon$$

Order in probability (multivariate)

Let $\{\mathbf{A}_n\}, \{\mathbf{B}_n\}$ be sequences of matrices of the same dimensions. Then

$$\mathbf{A}_n = O_p(\mathbf{B}_n) \text{ if } a_{n,ij} = O_p(b_{n,ij})$$

for all elements $a_{n,ij}$ of $\mathbf{A}_n$ and $b_{n,ij}$ of $\mathbf{B}_n$.

Convergence rate definition

$\hat{\mathbf{H}} \to \mathbf{H}_{\mathrm{AMISE}}$ with (relative) rate $n^{-\alpha}$ if

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

where

- $\operatorname{vech} \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix} = \begin{bmatrix} a & b & c & d & e & f \end{bmatrix}^T$
- $\mathbf{J}$ is $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix of ones

Strategy (1)

- $\mathrm{MSE}(\hat{\mathbf{H}}) = \mathbb{E}[\mathrm{vech}(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}})\,\mathrm{vech}^T(\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}})]$
- If we can write

$$\mathrm{MSE}(\hat{\mathbf{H}}) = O(n^{-2\alpha}\mathbf{J})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\,\mathbf{H}_{\mathrm{AMISE}})$$

- then $\hat{\mathbf{H}} \to \mathbf{H}_{\mathrm{AMISE}}$ at rate $n^{-\alpha}$

Strategy (2)

Decomposition of MSE:

$$\mathrm{MSE}(\hat{\mathbf{H}}) = \mathrm{Var}(\hat{\mathbf{H}}) + [\mathrm{Bias}(\hat{\mathbf{H}})][\mathrm{Bias}^T(\hat{\mathbf{H}})]$$

where

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

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

Summary of strategy

To find convergence rate of $\hat{\mathbf{H}} = \operatorname*{argmin}_{\mathbf{H}} \widehat{\mathrm{AMISE}}(\mathbf{H})$ to $\mathbf{H}_{\mathrm{AMISE}}$:

1. Compute order of expected value and variance of

$$\frac{\partial}{\partial \operatorname{vech} \mathbf{H}} (\widehat{\mathrm{AMISE}} - \mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}}).$$

2. Compute order of $\mathrm{MSE}(\hat{\mathbf{H}})$ from Step 1. Convergence rate is square root of order of $\mathrm{MSE}(\hat{\mathbf{H}})$.

Smoothed cross validation

- $$\begin{aligned}\mathrm{SCV}(\mathbf{H};\mathbf{G}) &= \widehat{\mathrm{AMISE}}(\mathbf{H}) \\ &= n^{-1}R(K)|\mathbf{H}|^{-1/2} \\ &\quad + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j)\end{aligned}$$

  where K is normal kernel

- $\hat{\mathbf{H}} = \underset{\mathbf{H}}{\operatorname{argmin}}\ \mathrm{SCV}(\mathbf{H};\mathbf{G})$

SCV selectors

- Hall, Marron & Park (1992): univariate with optimal pilot selector g indep. of h
- Sain, Baggerly & Scott (1994): multivariate, diagonal matrix with sub-optimal pilot selector $\mathbf{G}$ set equal to $\mathbf{H}$
- Proposed: multivariate, full matrix with optimal pilot selector $\mathbf{G} = g^2\mathbf{I}$ indep. of $\mathbf{H}$

Convergence rate for SCV selector (1)

Step 1. (a)

$$\mathbb{E}[\mathrm{SCV}(\mathbf{H}; g^2\mathbf{I})]$$

$$= n^{-1}R(K)|\mathbf{H}|^{-1/2} + n^{-1}(K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}})(\mathbf{0})$$

$$+ n^{-2}\,\mathbb{E}\left[\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{2\mathbf{H}+2\mathbf{G}} - 2K_{\mathbf{H}+2\mathbf{G}} + K_{2\mathbf{G}})(\boldsymbol{X}_i - \boldsymbol{X}_j)\right]$$

$$\vdots$$

$$= \mathrm{AMISE}(\mathbf{H}) + O((g^2 + n^{-1}g^{-d-4})\|\mathrm{vech}\,\mathbf{H}\|^2)$$

$$\Rightarrow \mathbb{E}\left[\frac{\partial}{\partial\,\mathrm{vech}\,\mathbf{H}}(\mathrm{SCV} - \mathrm{AMISE})(\mathbf{H})\right] = O((g^2 + n^{-1}g^{-d-4})\mathbf{J})\,\mathrm{vech}\,\mathbf{H}$$

Convergence rate for SCV selector (2)

Step 1. (b)

$$\operatorname{Var}\left[\frac{\partial}{\partial \operatorname{vech}\mathbf{H}}(\mathrm{SCV}-\mathrm{AMISE})(\mathbf{H})\right]$$

$$= \operatorname{Var}\left[\frac{\partial}{\partial \operatorname{vech}\mathbf{H}}\mathrm{SCV}(\mathbf{H})\right]$$

$$= n^{-4}\operatorname{Var}\left[\frac{\partial}{\partial \operatorname{vech}\mathbf{H}}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}(K_{2\mathbf{H}+2\mathbf{G}}-2K_{\mathbf{H}+2\mathbf{G}}+K_{2\mathbf{G}})(\boldsymbol{X}_i-\boldsymbol{X}_j)\right]$$

$$\vdots$$

$$= O((n^{-2}g^{-d-8}+n^{-1})\mathbf{J})(\operatorname{vech}\mathbf{H})(\operatorname{vech}^T\mathbf{H})$$

Convergence rate for SCV selector (3)

Step 2. We have $g = O(n^{-1/(d+6)})$ so

$$\mathrm{Bias}(\hat{\mathbf{H}}) = O(n^{-2/(d+6)}\mathbf{J})\,\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}}$$

$$\mathrm{Var}(\hat{\mathbf{H}}) = O(n^{-4/(d+6)}\mathbf{J})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\,\mathbf{H}_{\mathrm{AMISE}})$$

$$\Rightarrow \mathrm{MSE}(\hat{\mathbf{H}}) = O(n^{-4/(d+6)}\mathbf{J})(\mathrm{vech}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vech}^T\,\mathbf{H}_{\mathrm{AMISE}})$$

$$\Rightarrow \hat{\mathbf{H}} \to \mathbf{H}_{\mathrm{AMISE}} \text{ at rate } n^{-2/(d+6)}$$

Unbiased (Least Squares) cross validation

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

compared to

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

Biased cross validation and plug-in

$$\mathrm{BCV}(\mathbf{H}) = n^{-1} R(K) |\mathbf{H}|^{-1/2} + \frac{1}{4} [\mathrm{vech}^T \mathbf{H}] \hat{\mathbf{\Psi}}_4(\mathbf{H}) [\mathrm{vech}\, \mathbf{H}]$$

compared to

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

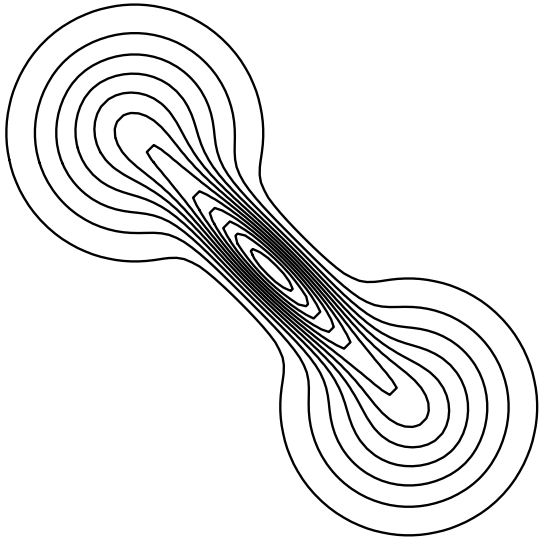
Convergence rates as at Oct 2004

Selector	Rate (2001)	Rate (2004)
Plug-in$_1$ (diag)	$n^{-\min(8,d+4)/(2d+12)}$	unchanged
Plug-in$_1$ (full)	$n^{-4/(d+12)}$	unchanged
Plug-in$_2$ (full)	-	$n^{-2/(d+6)}$
UCV/BCV (diag)	$n^{-d/(2d+8)}$	$n^{-\min(d,4)/(2d+8)}$
UCV/BCV (full)	-	$n^{-\min(d,4)/(2d+8)}$
SCV ($d = 1$)	$n^{-(d+4)/(2d+12)}$	unchanged
SCV ($d > 1$, full)	-	$n^{-2/(d+6)}$

Data example - ‘dumbbell’ density (1)

$$\frac{4}{11}N\left(\begin{bmatrix}-2\\2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)+\frac{3}{11}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}0.8&-0.72\\-0.72&0.8\end{bmatrix}\right)+\frac{4}{11}N\left(\begin{bmatrix}2\\-2\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$$

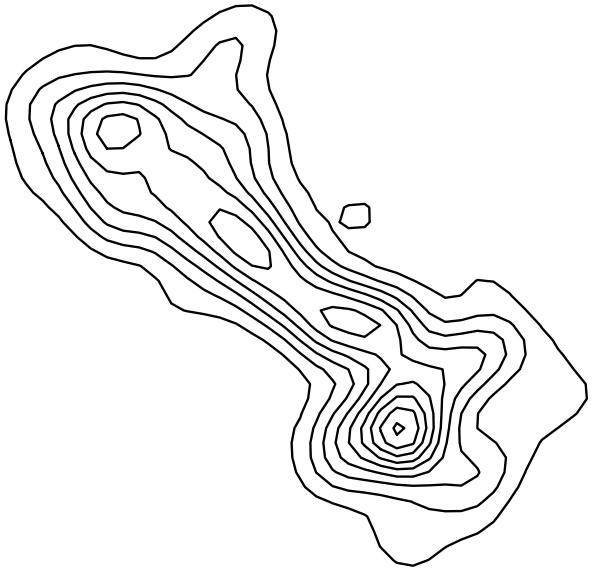
Contour plot



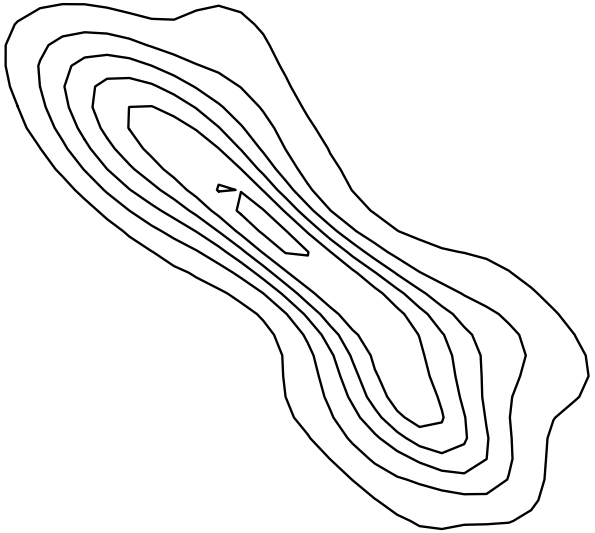
Scatter plot

Data example - ‘dumbbell’ density (2)

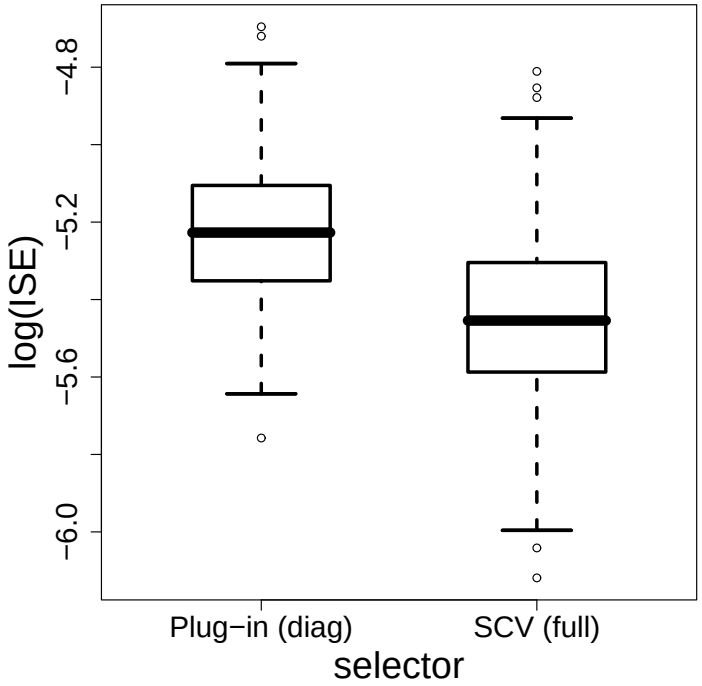
diagonal plug-in



full SCV



Data example - ‘dumbbell’ density (3)



R Software

- ks 1.4.2 on CRAN http://www.r-project.org
- 2- to 6-dim data
- kernel density estimation (KDE) and kernel discriminant analysis (KDA)

Publications

- Duong, T. and Hazelton, M.L. (2005) Journal of Multivariate Analysis, **93**, 417 - 433
- Duong, T. and Hazelton, M.L. (2005) Scandinavian Journal of Statistics, **32**, 485 - 506
- Duong, T. (2004) Ph.D. Thesis, University of Western Australia.

Summary

- Unified framework for computing convergence for any multivariate bandwidth selector
  - e.g. improved plug-in, UCV, BCV, SCV
- Finite sample behaviour
- General recommendation: SCV or improved plug-in selector