Plug-in Bandwidth Selectors for Bivariate Kernel Density Estimation

Tarn Duong
Supervisor: Dr Martin Hazelton

May 2002
Outline

1. Motivation

2. Basics of KDE

3. Plug-in bandwidth selectors
   - Existing and proposed methods

4. Simulation and real data results

5. Summary
Motivation
Forming KDE (1)

Full bandwidths
Forming KDE (2)
Equation for KDE

Kernel density estimate is

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

where

- $X_1, X_2, \ldots, X_n$ is sample of $n$ data points
- $\mathbf{H}$ is bandwidth matrix
- $K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i)$ is normal pdf with mean $\mathbf{X}_i$ and variance $\mathbf{H}$
Advantages of KDE

• does not require parametric form i.e. is non-parametric

• always guaranteed to be a proper pdf
MISE

Mean Integrated Squared Error (MISE) is

\[
\text{MISE} \hat{f}(\cdot; \mathbf{H}) = \int_{\mathbb{R}^2} \text{MSE} \hat{f}(\mathbf{x}; \mathbf{H}) \, d\mathbf{x}
\]

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

where

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

Asymptotic Mean Integrated Squared Error (AMISE) is

\[
\text{AMISE } \hat{f}(\cdot; \mathbf{H}) = n^{-1}(4\pi)^{-1}|\mathbf{H}|^{-1/2} + \frac{1}{4}(\text{vech}^T\mathbf{H})\Psi(\text{vech} \mathbf{H})
\]

where

- \text{vech} \mathbf{H} is vector of lower triangular half of \mathbf{H}
  
  i.e. if \( \mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix} \) then \( \text{vech} \mathbf{H} = [h_1^2 \ h_{12} \ h_2^2] \)

- \( \Psi \) is matrix of functionals that depend on \( f \)
Plug-in selector

1. choose diagonal or full bandwidth matrix

2. estimate $\Psi$

3. PI is AMISE with $\Psi$ replaced with $\hat{\Psi}$

$$\text{PI}(H) = n^{-1}(4\pi)^{-1}|H|^{-1/2} + \frac{1}{4} (\text{vech}^T H) \hat{\Psi} (\text{vech} H)$$

4. optimal bandwidth is $\hat{H}_{PI} = \arg\min_H \text{PI}(H)$
Existing and proposed methods

- **Existing**
  1. use diagonal bandwidth matrix
  2. estimate $\Psi$ element-wise

- **Proposed**
  1. use full bandwidth matrix
  2. estimate $\Psi$ matrix-wise
‘Old Faithful’ geyser data
Diagonal and full bandwidth matrices

Full bandwidths

Diagonal bandwidths
Explicit expression for $\Psi$

- partial derivative: 
  \[ f^{(r_1,r_2)}(x) = \frac{\partial^{(r_1+r_2)}}{\partial x_1^{r_1} \partial x_2^{r_2}} f(x) \]

- $\psi$ functional: 
  \[ \psi_{r_1,r_2} = \int_{\mathbb{R}^2} f^{(r_1,r_2)}(x) f(x) \, dx \]

- matrix: 
  \[
  \Psi = \begin{bmatrix}
  \psi_{40} & 2\psi_{31} & \psi_{22} \\
  2\psi_{31} & 4\psi_{22} & 2\psi_{13} \\
  \psi_{22} & 2\psi_{13} & \psi_{04}
  \end{bmatrix}
  \]
Estimating $\psi_{r_1,r_2}$

- $\psi_{r_1,r_2} = \mathbb{E} f^{(r_1,r_2)}(X)$

- $\hat{\psi}_{r_1,r_2} = n^{-1} \sum_{i=1}^{n} \hat{f}^{(r_1,r_2)}(X_i; g)$
  where $g$ is a (scalar) pilot bandwidth

- choose $g$ via minimising MSE $\psi_{r_1,r_2}$
Estimating $\psi_{r_1, r_2}$ - existing method (1)

Asymptotic MSE of $\hat{\psi}_{r_1, r_2}$ is

\[
\text{AMSE} \hat{\psi}_{r_1, r_2}(g) = (4\pi)^{-1} n^{-2} g^{-10} \\
+ \left[ n^{-1} g^{-6} K^{(r_1, r_2)}(0) + \frac{1}{2} g^2 (\psi_{r_1+2, r_2} + \psi_{r_1, r_2+2}) \right]^2
\]
Estimating $\psi_{r_1,r_2}$ - existing method (2)

If $r_1, r_2$ are both even then

$$g_{r_1,r_2} = \left[ \frac{-2K(r_1,r_2)(0)}{(\psi_{r_1+2,r_2} + \psi_{r_1,r_2+2})n} \right]^{1/8}$$

If $r_1, r_2$ are both odd then

$$g_{r_1,r_2} = \left[ \frac{5\psi_{0,0}}{2\pi(\psi_{r_1+2,r_2} + \psi_{r_1,r_2+2})^2n^2} \right]^{1/14}$$
Estimating $\Psi$ - existing method

$$\hat{\Psi} = \begin{bmatrix} \hat{\psi}_{40}(g_{40}) & 2\hat{\psi}_{31}(g_{31}) & \hat{\psi}_{22}(g_{22}) \\ 2\hat{\psi}_{31}(g_{31}) & 4\hat{\psi}_{22}(g_{22}) & 2\hat{\psi}_{13}(g_{13}) \\ \hat{\psi}_{22}(g_{22}) & 2\hat{\psi}_{13}(g_{13}) & \hat{\psi}_{04}(g_{04}) \end{bmatrix}$$

$\hat{\Psi}$ can be

- not $+$ve definite so no finite min for $\text{PI}(H)$
- nearly singular so numerical instability in $\text{PI}(H)$
Estimating $\psi_{r_1,r_2}$ - proposed method (1)

Sum of AMSE is

$$\text{SAMSE}(g) = \sum_{r_1=0}^{4} \text{AMSE} \hat{\psi}_{r_1,r_2}(g) \quad \text{where} \quad r_2 = 4 - r_1$$

$$= n^{-2} g^{-12} A_2 + n^{-1} g^{-4} A_3 + \frac{1}{4} g^4 A_4$$

where $A_2, A_4 > 0$ and $A_3 < 0$. 
Estimating $\psi_{r_1,r_2}$ - proposed method (2)

\[
g = \left[ \frac{6A_2}{(-A_3 + \sqrt{A_3^2 + 3A_1 A_2})n} \right]^{1/8}
\]
Estimating $\Psi$ - proposed method

$\hat{\Psi} = \begin{bmatrix} \hat{\psi}_{40}(g) & 2\hat{\psi}_{31}(g) & \hat{\psi}_{22}(g) \\ 2\hat{\psi}_{31}(g) & 4\hat{\psi}_{22}(g) & 2\hat{\psi}_{13}(g) \\ \hat{\psi}_{22}(g) & 2\hat{\psi}_{13}(g) & \hat{\psi}_{04}(g) \end{bmatrix}$

- $\hat{\Psi}$ is +ve definite.
- lose a little efficiency compared to existing method
Simulation results (1)

• normal mixture densities
  – wide range of features
  – closed form for $\text{ISE} \hat{f}(\cdot; \mathbf{H}) = \int_{\mathbb{R}^2} [\hat{f}(\mathbf{x}; \mathbf{H}) - f(\mathbf{x})]^2 d\mathbf{x}$

• use $\log(\text{ISE})$ to compare performance between
  – diagonal selector with different pilots
  – full selector with different pilots
  – full selector with single pilot

• 400 simulations, each of size $n = 1000$
Simulation results (2)

Highly correlated normal density - contour plot
Simulation results (3)

Highly correlated normal density - boxplot
Simulation results (4)

Trimodal density - contour plot
Simulation results (5)

Trimodal density - boxplot
Real data KDE - existing method

Density estimate of geyser eruption data
Real data KDE - proposed method

Density estimate of geyser eruption data

- Typeset by Foil\TeX -
# Summary

<table>
<thead>
<tr>
<th>Existing method</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagonal $\mathbf{H}$</td>
<td>Full $\mathbf{H}$</td>
</tr>
<tr>
<td>Different pilots</td>
<td>Single pilot</td>
</tr>
<tr>
<td>$\hat{\Psi} \not\approx 0$</td>
<td>$\hat{\Psi} &gt; 0$</td>
</tr>
<tr>
<td>Poor estimation of data not aligned to axes</td>
<td>Good estimation of data not aligned to axes</td>
</tr>
</tbody>
</table>