修勻學(Graduation) — 核(Kernel)修勻法

授課教師：余清祥教授

課程日期：2024年10月9日

資料下載：

http://csyue.nccu.edu.tw

估計密度函數

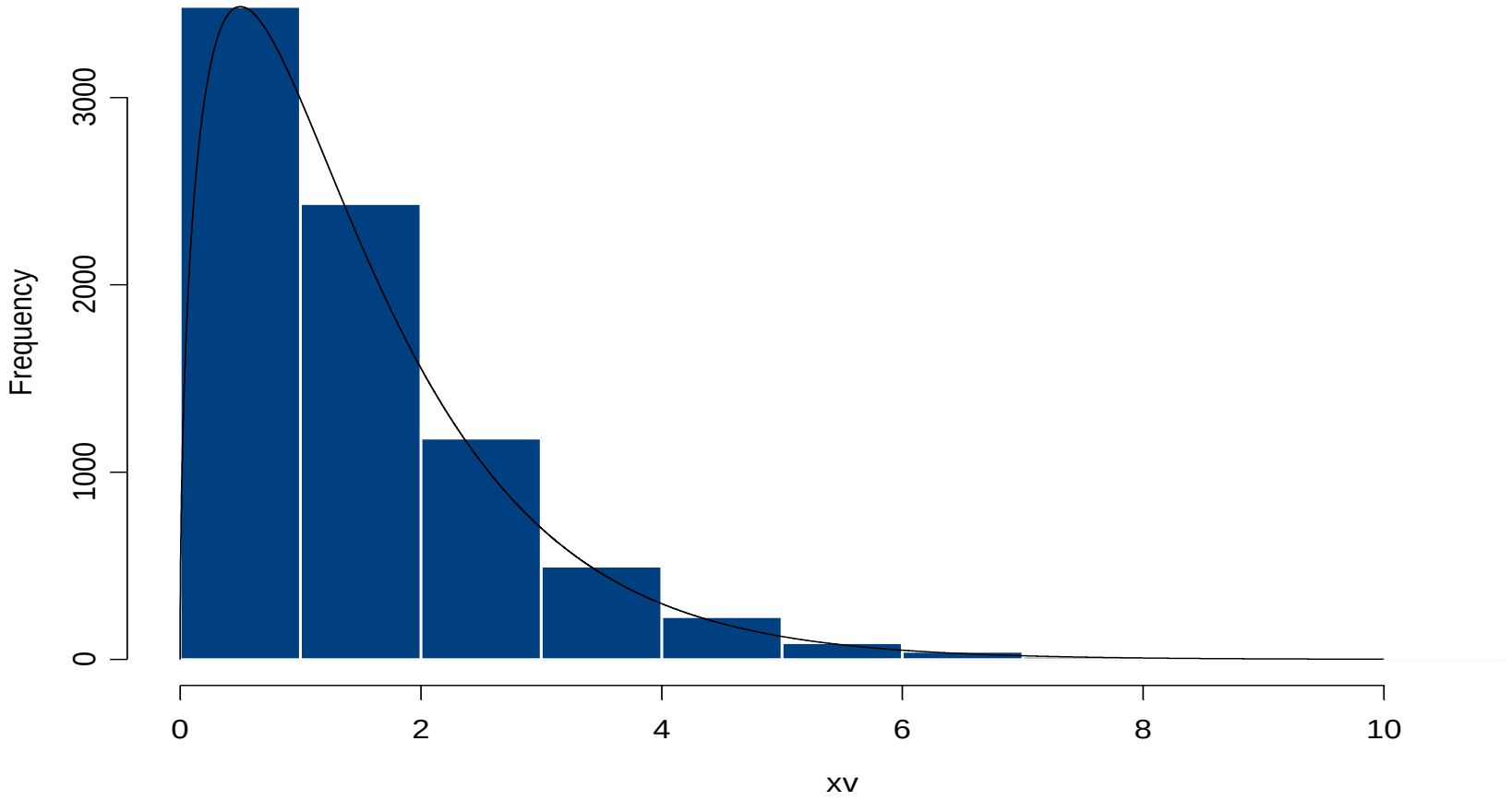
- 死亡率修勻與統計的密度函數(p.d.f.)估計類似，以無母數方法找出觀察值來自哪一種類型的分配。

→分配函數 (c.d.f. $F(x_0)$)的估計可藉由不大於 x_0的觀察值個數，同理，p.d.f.可藉由

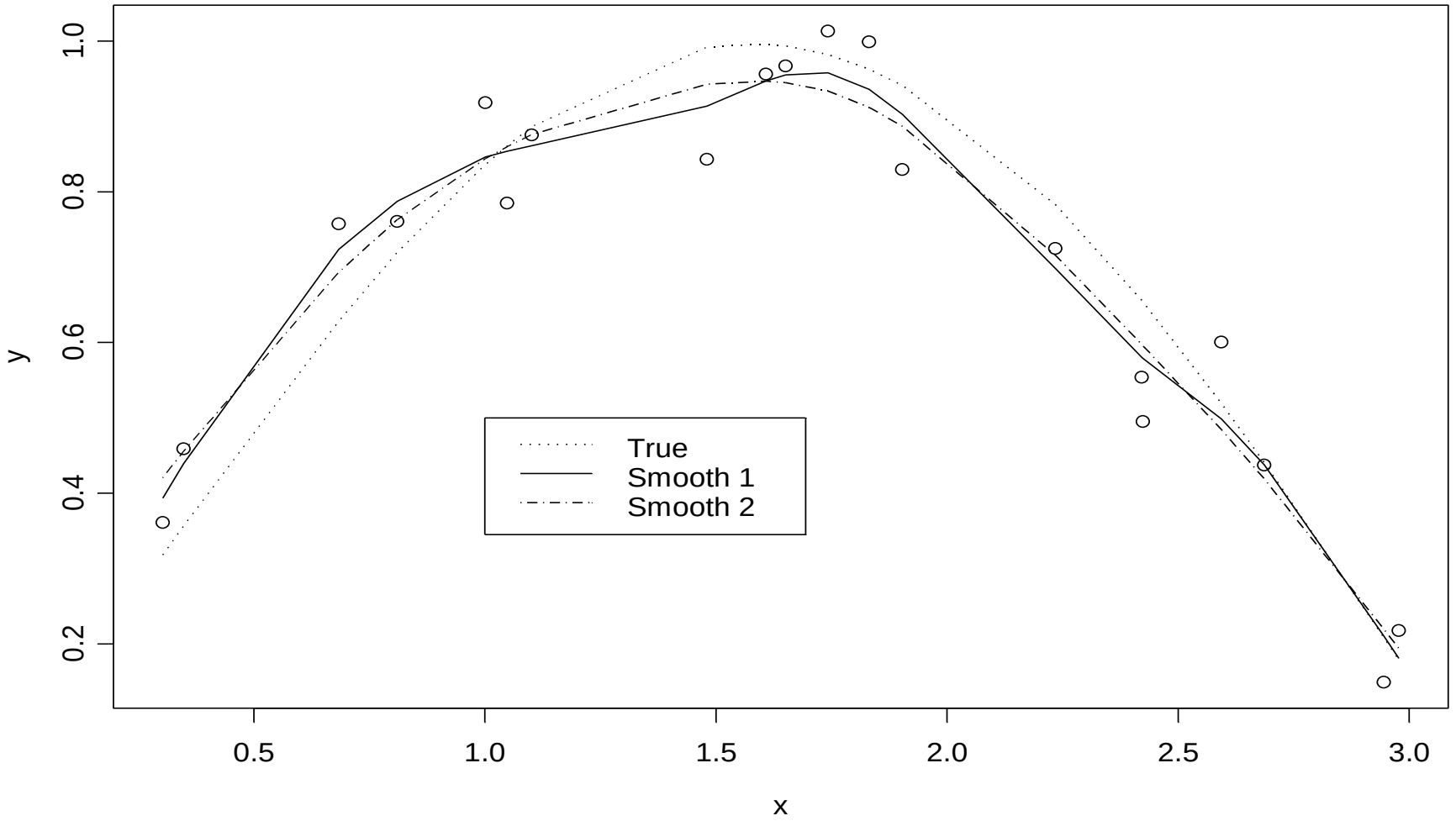
$$\hat{f}(x_0) = \sum I\{x_i = x_0\}/n.$$

推得，實際上卻不然，因為任何一個觀察值的發生機率為0。

直觀而言，在觀察值附近的點應該可以給定一個非零的加權數，而且權數應與觀察值的距離成正比（當然，也不一定要如此！）



- 平滑化(Smoothing)在統計的應用多與估計密度函數有關，也可用於保險領域，修匀死亡率曲線。



- 直方圖(Histogram)

→直方圖是最簡單的密度估計函數

→ 操作時將區間 $[a,b]$分成 m 個長度皆為 h的子區間， $a = a_0 < a_1 < \cdots < a_{m-1} < a_m = b.$

則區間任何一點 $x \in [a,b]$的密度函數為

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{m} \frac{n_j}{h} \cdot I\{x \in [a_{j-1}, a_j]\},$$

n_j 是區間 $x \in [a_{j-1}, a_j]$內的觀察值個數。

(註： $h = (a\text{-}b)/m$)

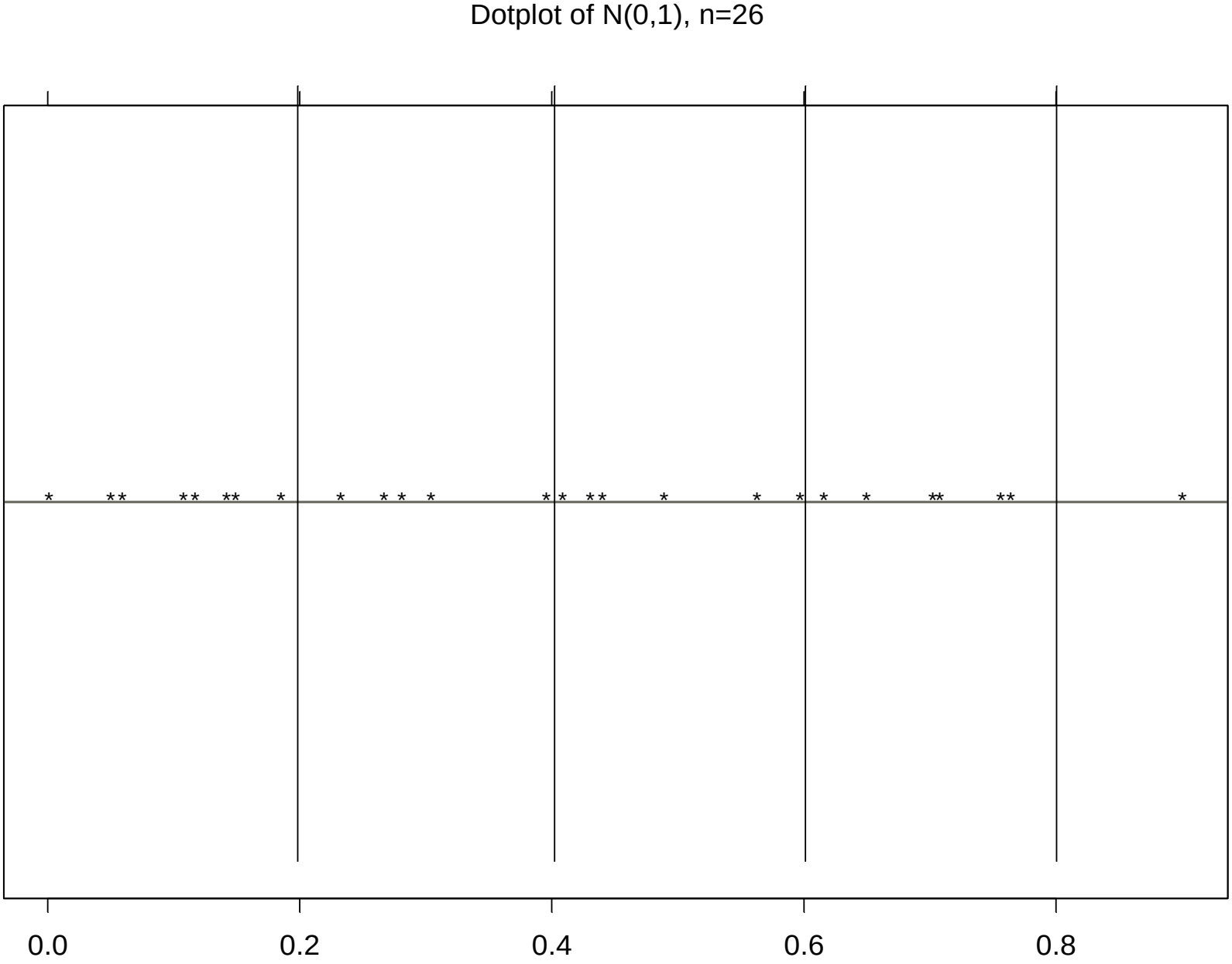
附註說明：

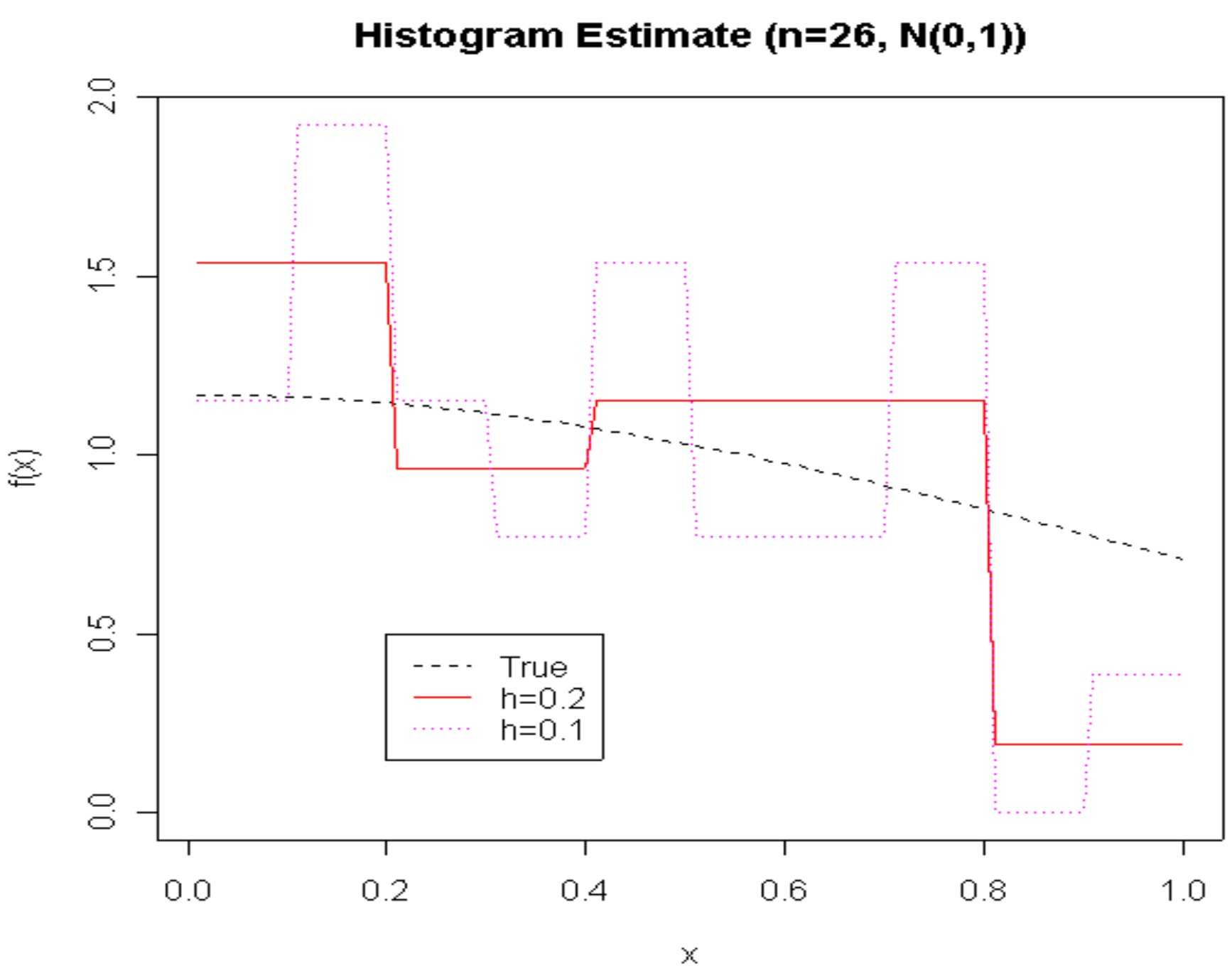
(1) 直方圖的密度估計值為階梯函數，且其形狀類似樣本CDF。

(2) 理論上，子區間的寬度 h 愈窄，密度函數估計值就愈能反映資料的特性；但實際上因為觀察個數有限，h 小到某個程度只可能包含一兩個觀察值，反而震盪幅度加大。（h 愈大直方圖就愈平滑。）

Q: 是否存在「最佳」的 h 值？

(換言之，是否有最佳的子區間個數？)


Dotplot of N(0,1), n=26
0.0
0.2
0.4
0.6
0.8


Histogram Estimate (n=26, N(0,1))
f(x)
x
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
True
h=0.2
h=0.1

- 直觀密度估計值
  (Naïve Density Estimator)

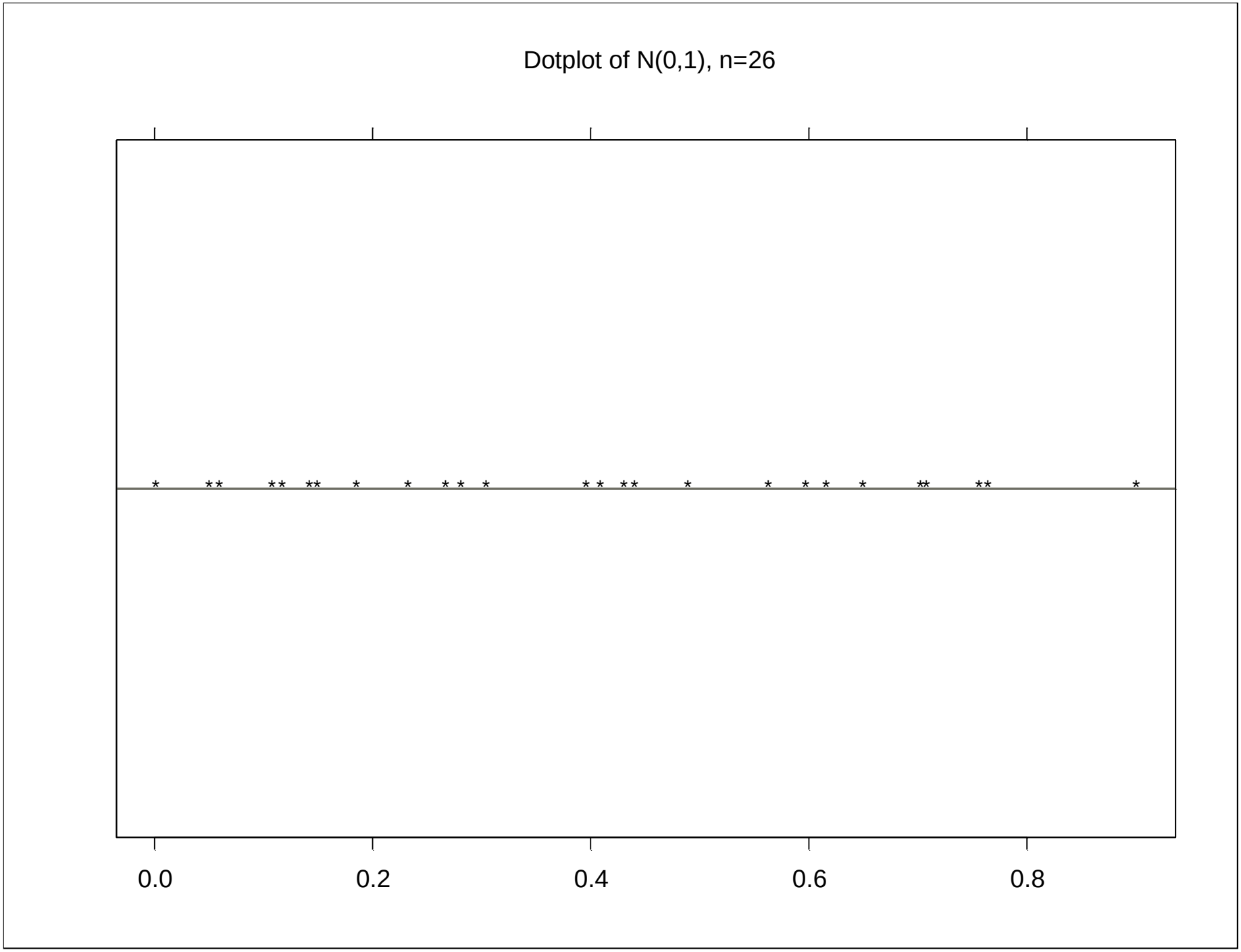
→Silverman (1986)建議將加權數以觀察值為中心，半徑 h 內皆有權數1：

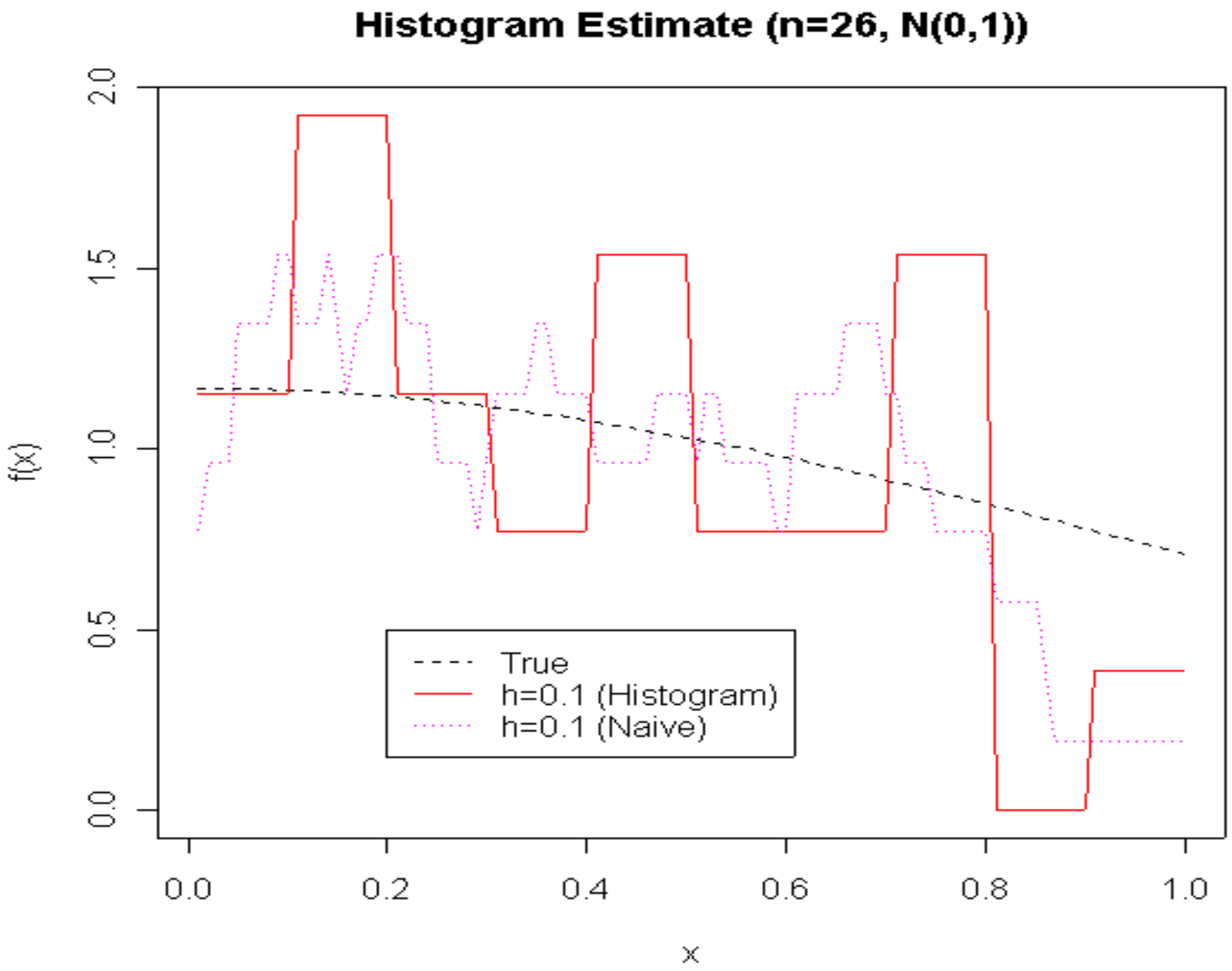
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x - x_i}{h}\right),$$

其中

$$w(x) = \begin{cases} \frac{1}{2}, & |x| < 1; \\ 0, & otherwise. \end{cases}$$

註：也稱這種密度估計為「寬度2h的移動平均直方圖」(*Moving-window histogram)*


Dotplot of N(0,1), n=26
0.0
0.2
0.4
0.6
0.8


Histogram Estimate (n=26, N(0,1))
f(x)
x
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
True
h=0.1 (Histogram)
h=0.1 (Naive)

- 核估計法(Kernel Estimator)

→因為權數與觀察值距離有關，可預期直觀密度估計值優於直方圖的估計值。然而由於加權函數為階梯函數，在接近觀察值時會有跳躍 (Jumps)，這種不平滑的現象可藉由調整加權函數，而達到平滑的要求。

→首先由Copas and Haberman(1983)及Ramlan-Hansen(1983)應用到死亡率的修勻。

→核修匀法的公式為：

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right),$$

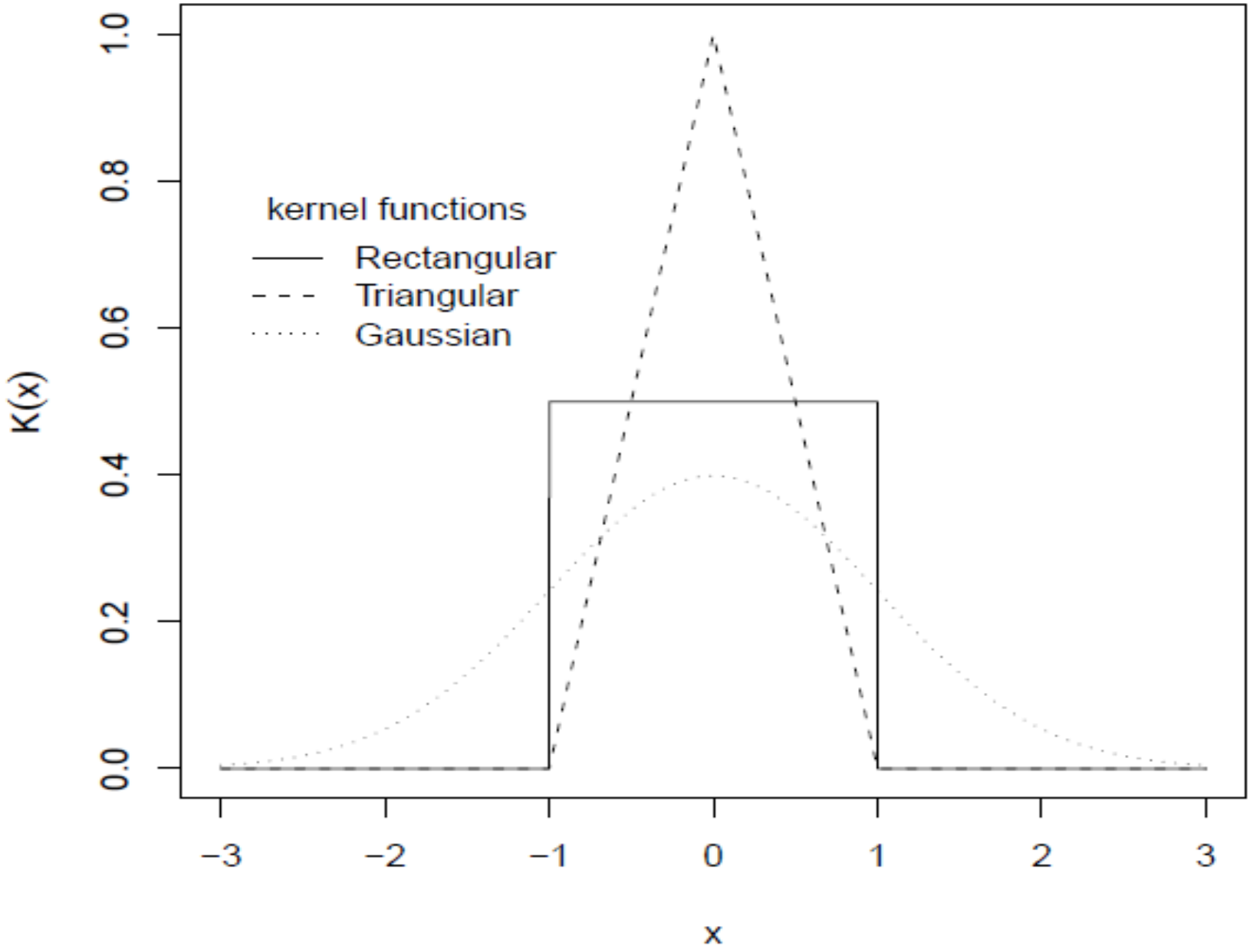
其中 $\int_{-\infty}^{\infty}K(t)dt=1$又稱為核函數(kernel of the estimator)。

→常見的核函數有下列幾種：

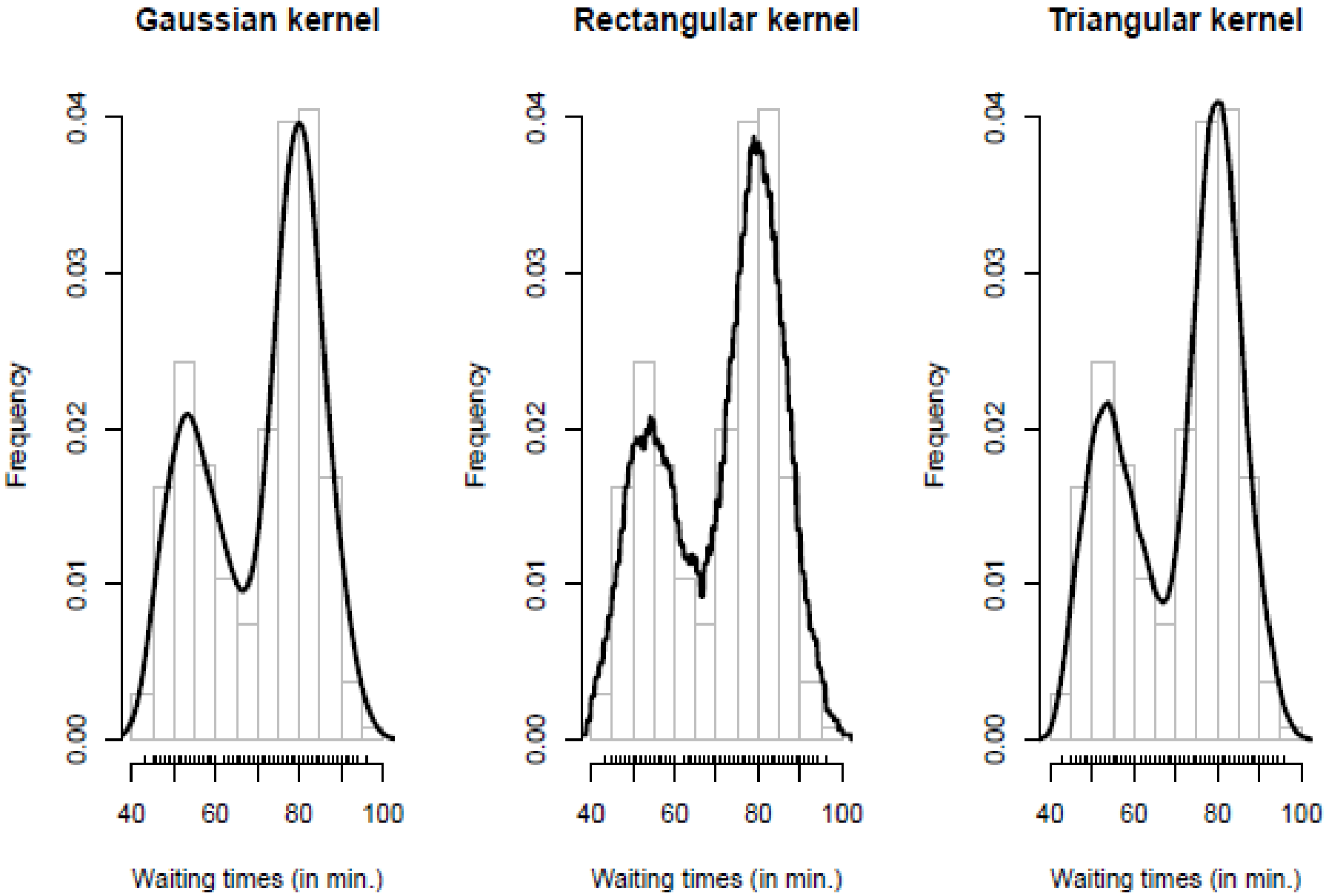
Guassian (i.e., normal), Cosine, Rectangular, Triangular, Lapalce.

$$K_N(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2},-\infty<x<\infty$$

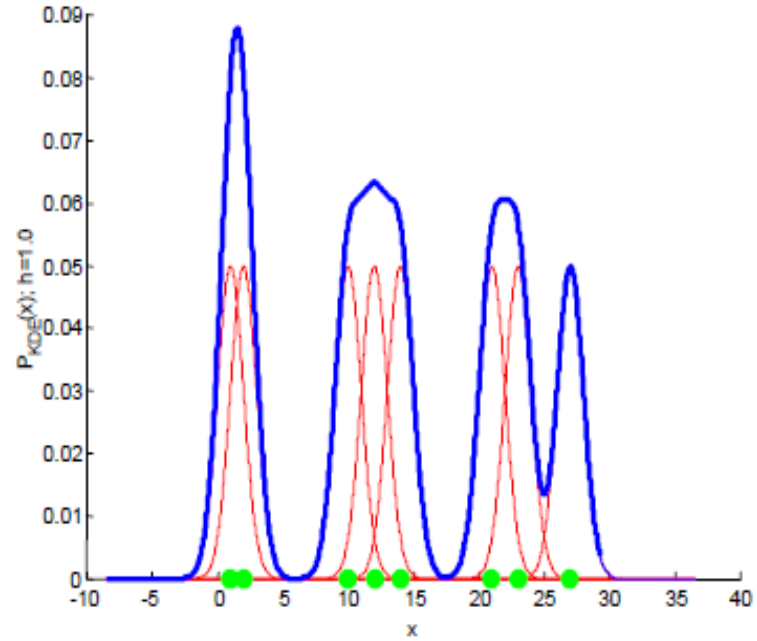
$$K_L(x)=\frac{1}{2}e^{-|x|},-\infty<x<\infty$$

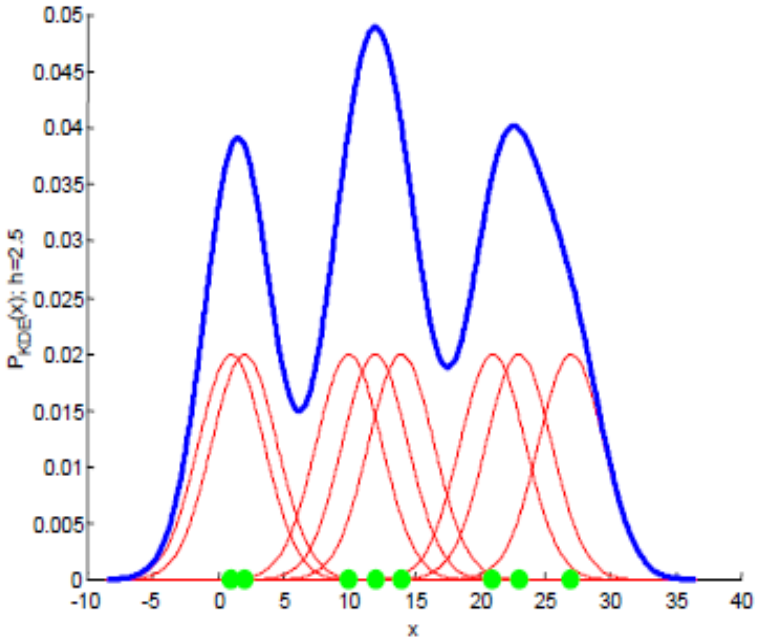

kernel functions
Rectangular
Triangular
Gaussian
K(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

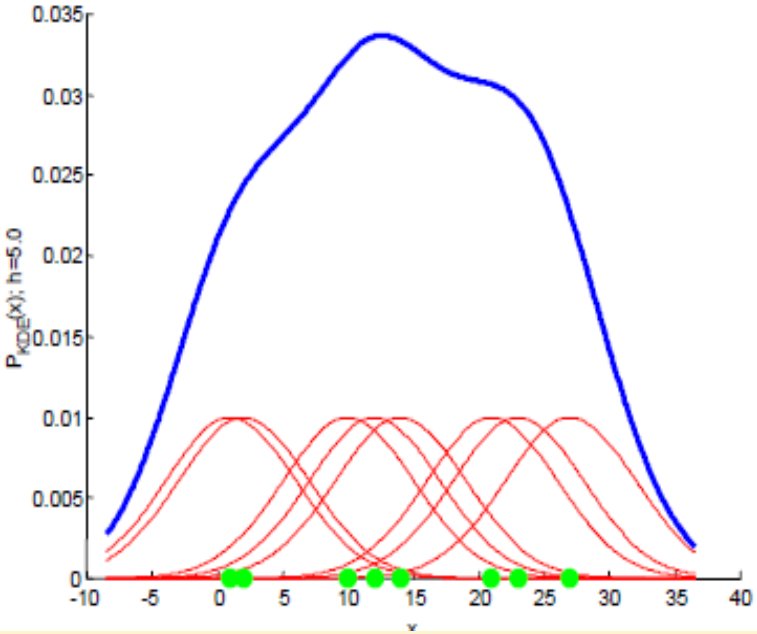
不同核函數的比較

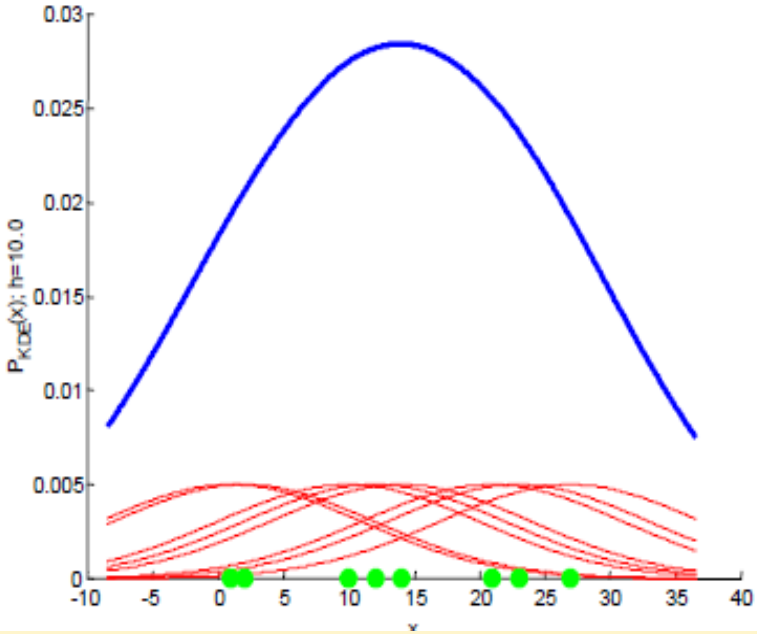

Gaussian kernel
Rectangular kernel
Triangular kernel
Frequency
0.00
0.01
0.02
0.03
0.04
40
60
80
100
Waiting times (in min.)

不同核函數的比較(續)

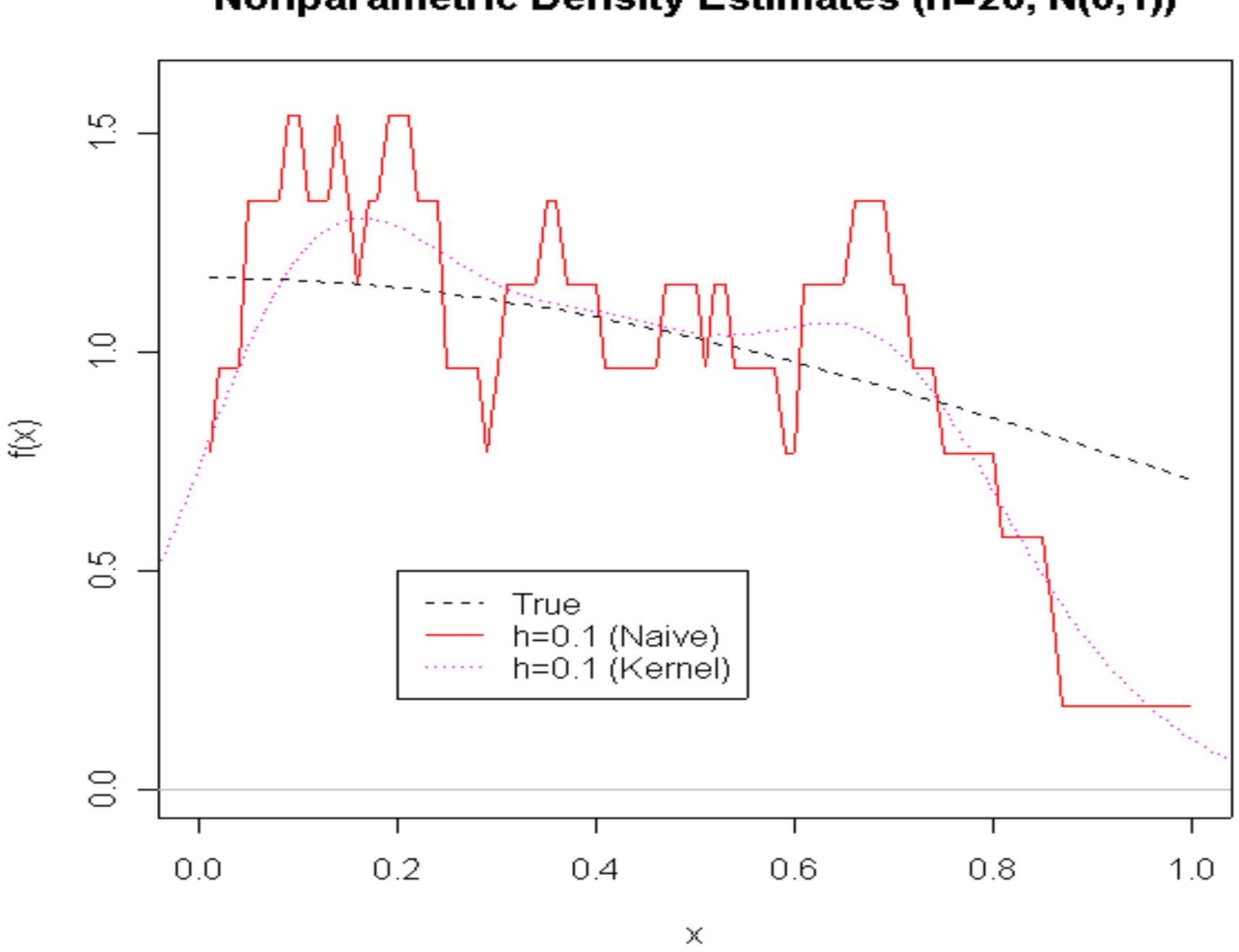








不同核修匀的環寬比較


Nonparametric Density Estimates (n=26, N(0,1))
f(x)
x
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
True
h=0.1 (Naive)
h=0.1 (Kernel)

→ 環寬(Bandwidth, h) 主宰各觀察值的加權比例。若 h 較小時，則僅有接近 x 的x_i值對函數估計有較大的貢獻。h 與Whittaker中 F + hS 的 h 角色類似，因此h的選擇通常比核函數更重要。

→死亡率的核修勻有以下兩種類型：

$$\begin{cases} \hat{q}_x^1 = [\sum_{i=1}^{n} \hat{q}_i K(\frac{x - x_i}{h})] / [\sum_{i=1}^{n} K(\frac{x - x_i}{h})], \\ \hat{q}_x^2 = [\sum_{i=1}^{n} d_i K(\frac{x - x_i}{h})] / [\sum_{i=1}^{n} e_i K(\frac{x - x_i}{h})], \end{cases}$$

→兩種死亡率的核修勻主要差異在於是否使用暴露數 e_i。如同Whittaker與MWA的差異，可預期只使用各年齡死亡率觀察值的核修勻，會受到離群值(如高齡死亡率)的影響。

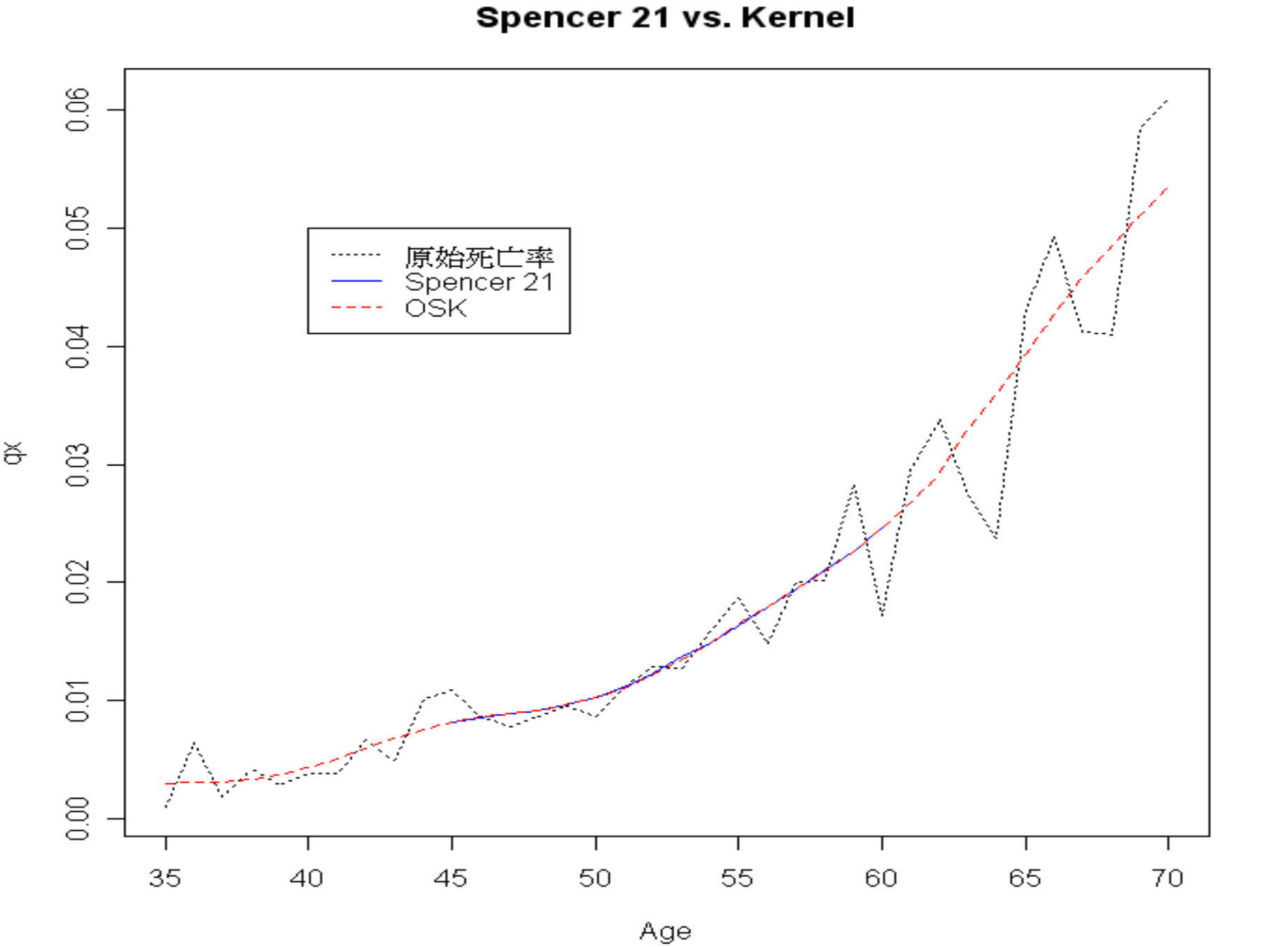
→以下將只介紹第二種核修勻法：

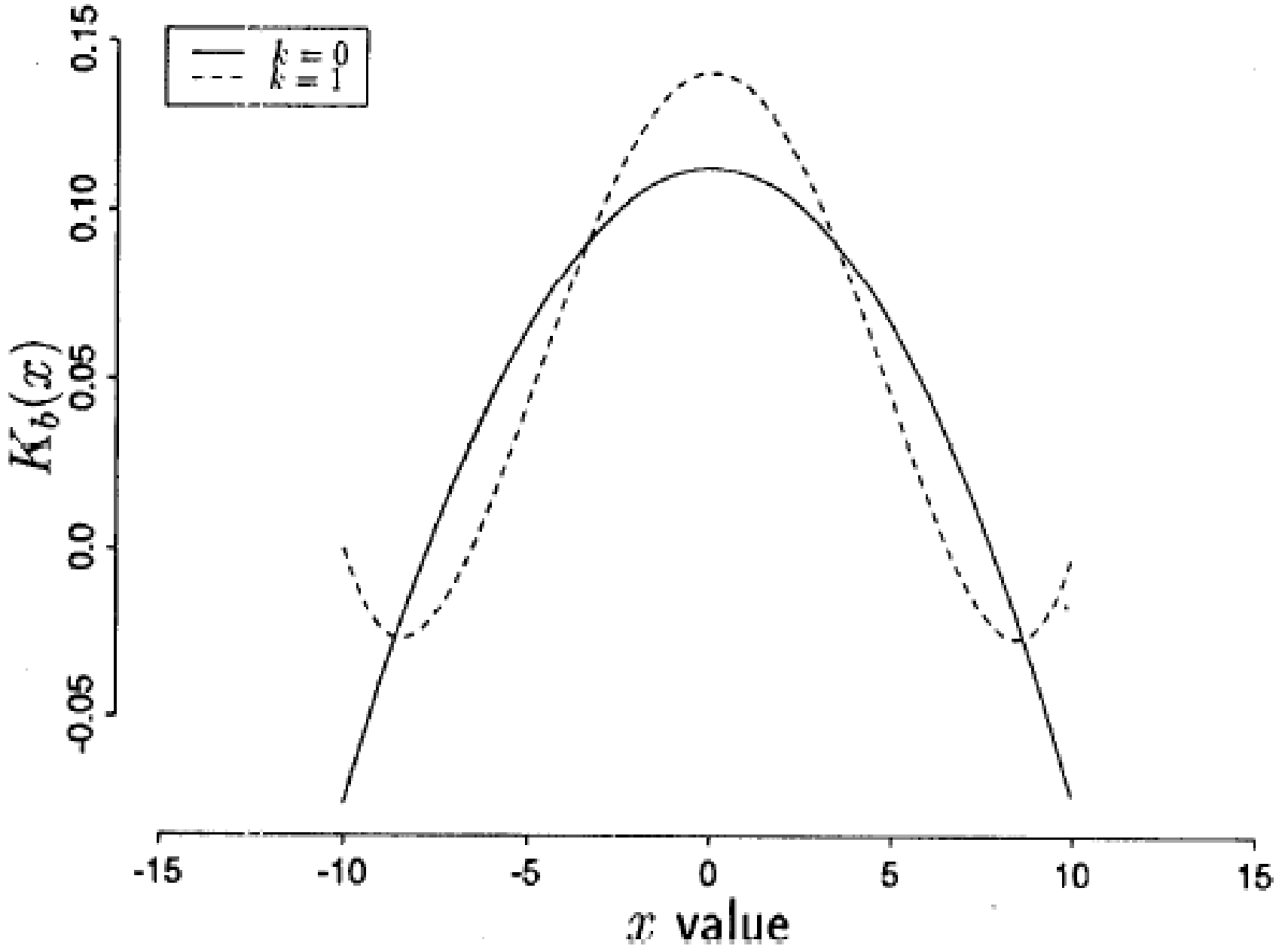
$$\hat{q}_x^2 = [\sum_{i=1}^{n} d_i K(\frac{x - x_i}{h})] / [\sum_{i=1}^{n} e_i K(\frac{x - x_i}{h})]$$

與Whittaker法類似，雖然可任何年齡死亡率的修勻值，但在兩個端點也會有較不平滑的現象。

範例一、Gavin, Haberman, and Verrall(1993)仿照MWA法中的作法，定義最適修匀核函數(The Optimal Smoothing Kernel, 簡稱OSK)是使MWA中R_z^2極小化的核函數。

→比較OSK及Spencer 21項公式。死亡率資料選自Benjamin and Pollard(1980)，死亡率由起自三十五歲至七十歲止，OSK的環寬為10，差分$z = 3$。明顯OSK與Spencer 21項公式的修匀結果重合，但OSK並未損失任何年齡的死亡率。


Spencer 21 vs. Kernel
原始死亡率
Spencer 21
OSK
qx
0.00
0.01
0.02
0.03
0.04
0.05
0.06
35
40
45
50
55
60
65
70
Age



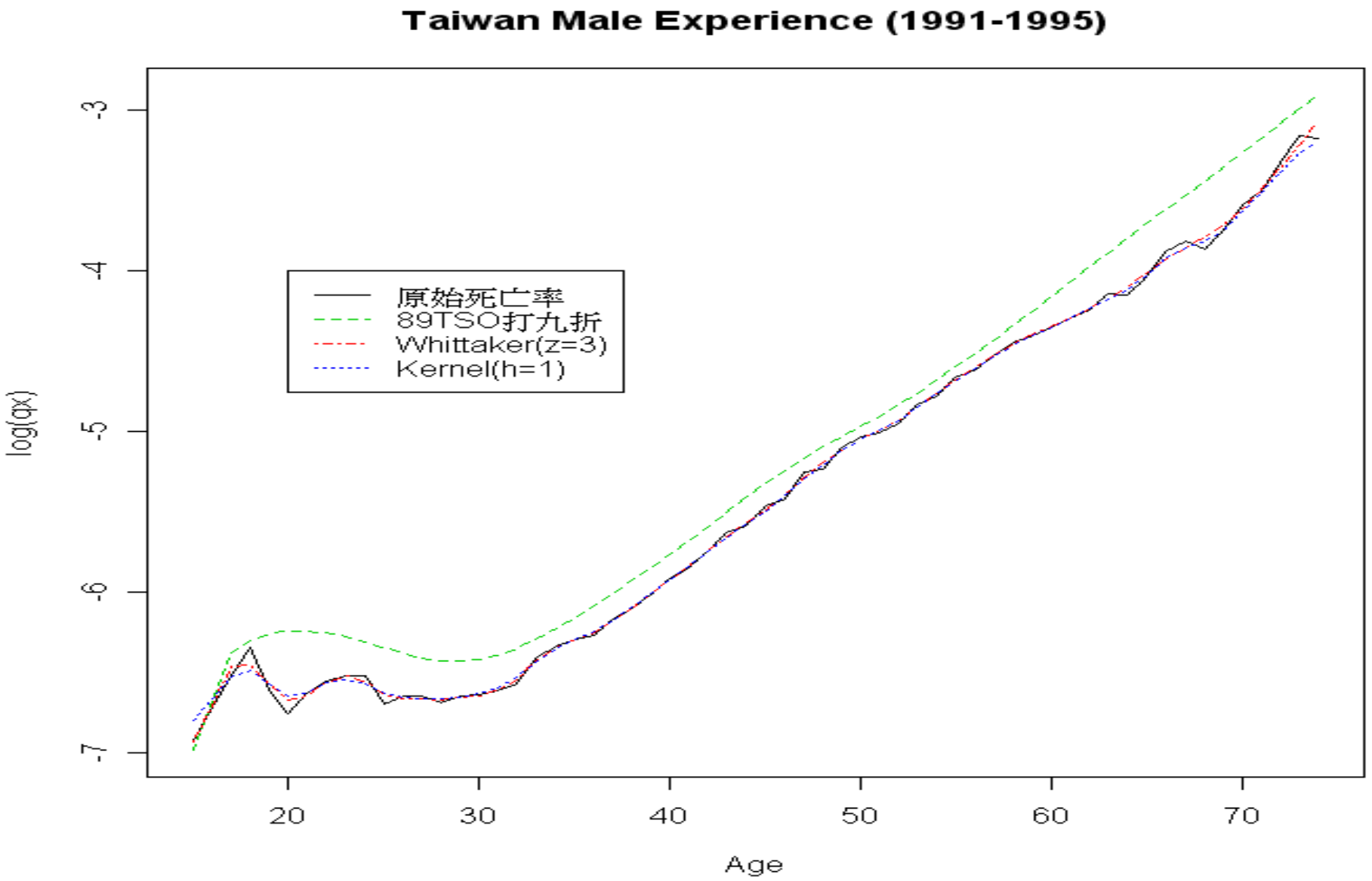
OSK修匀的加權係數(係數可為負值)

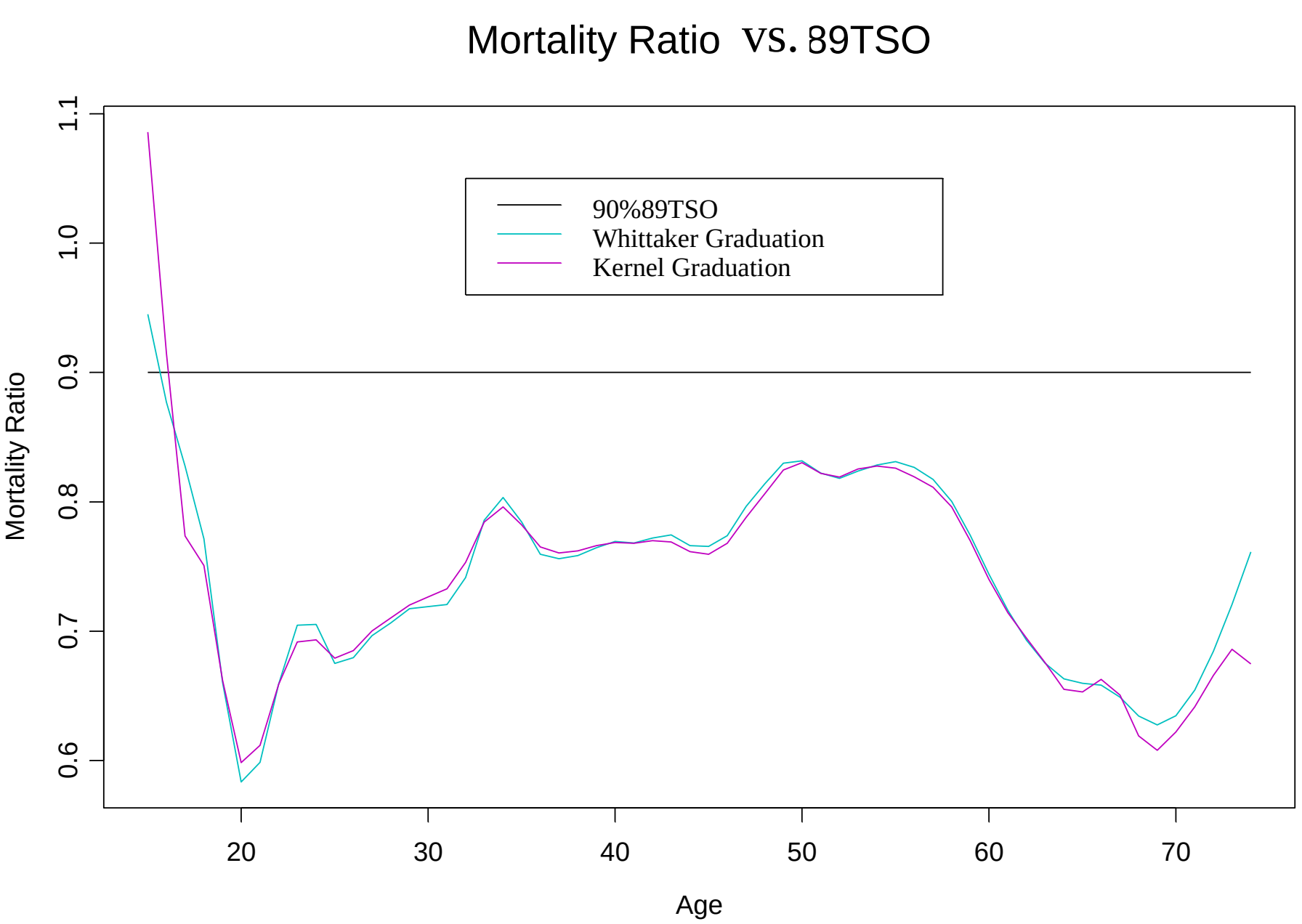
- 對於環寬 h的選取，Gavin, Haberman, and Verrall (1994)引用Stone (1974)的交叉驗證(Cross-Validation)，在使損失函數CV(h)極小化的考量下，找出最佳的環寬值 。

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(\hat{q}_i^{'} - \hat{q}_i^{(-j)})^2$$

→在CV(h)的考量下，OSK雖然可視為MWA的一般式推廣，但仍不如直接使用Normal kernel的核修匀法。換言之，以 $MinR_z^2$ 求得的加權平均，不如常態核函數。

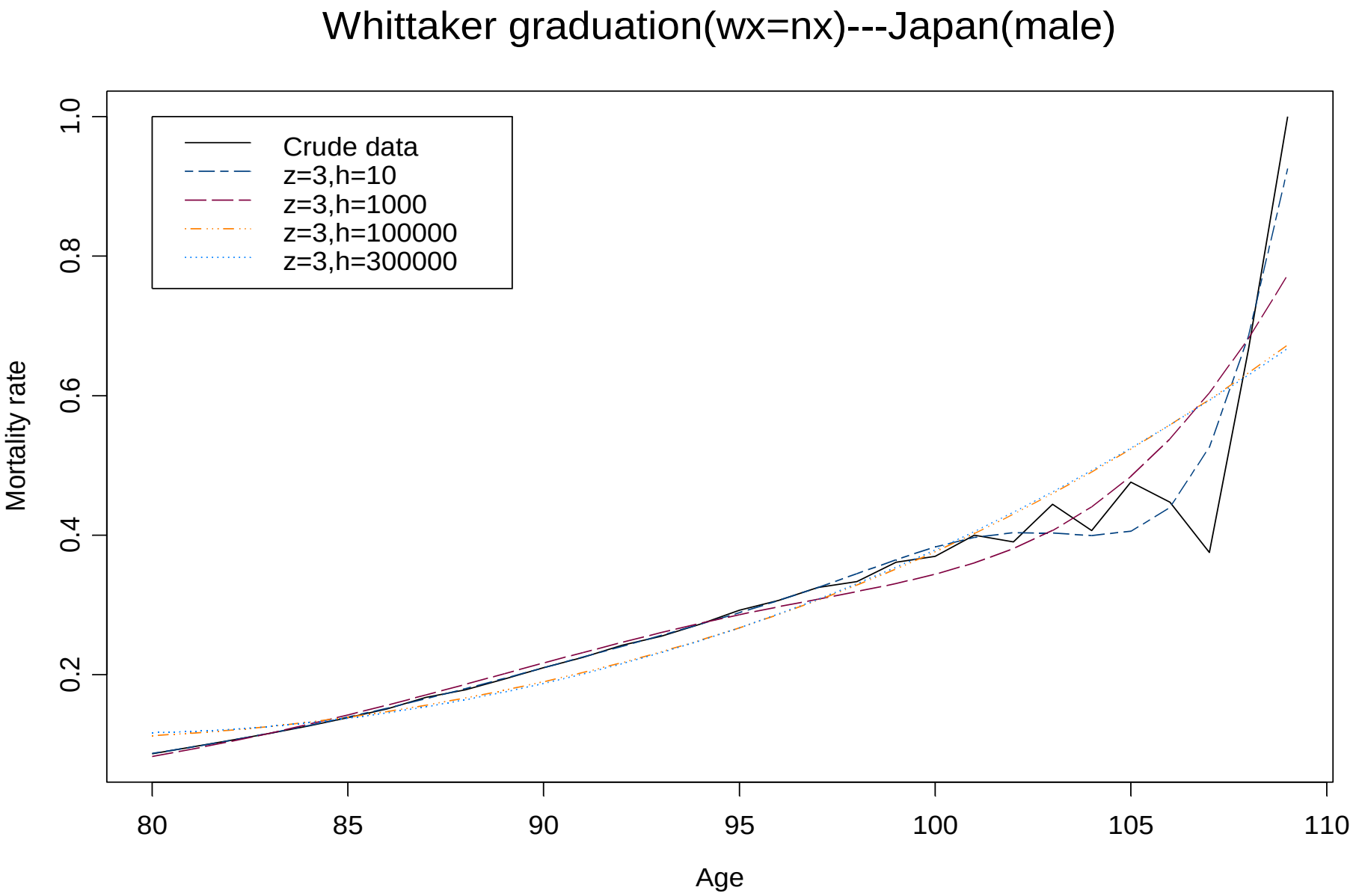
範例二、台灣壽險業八十至八十四觀察年度
單一年齡男性15至74歲粗死亡率

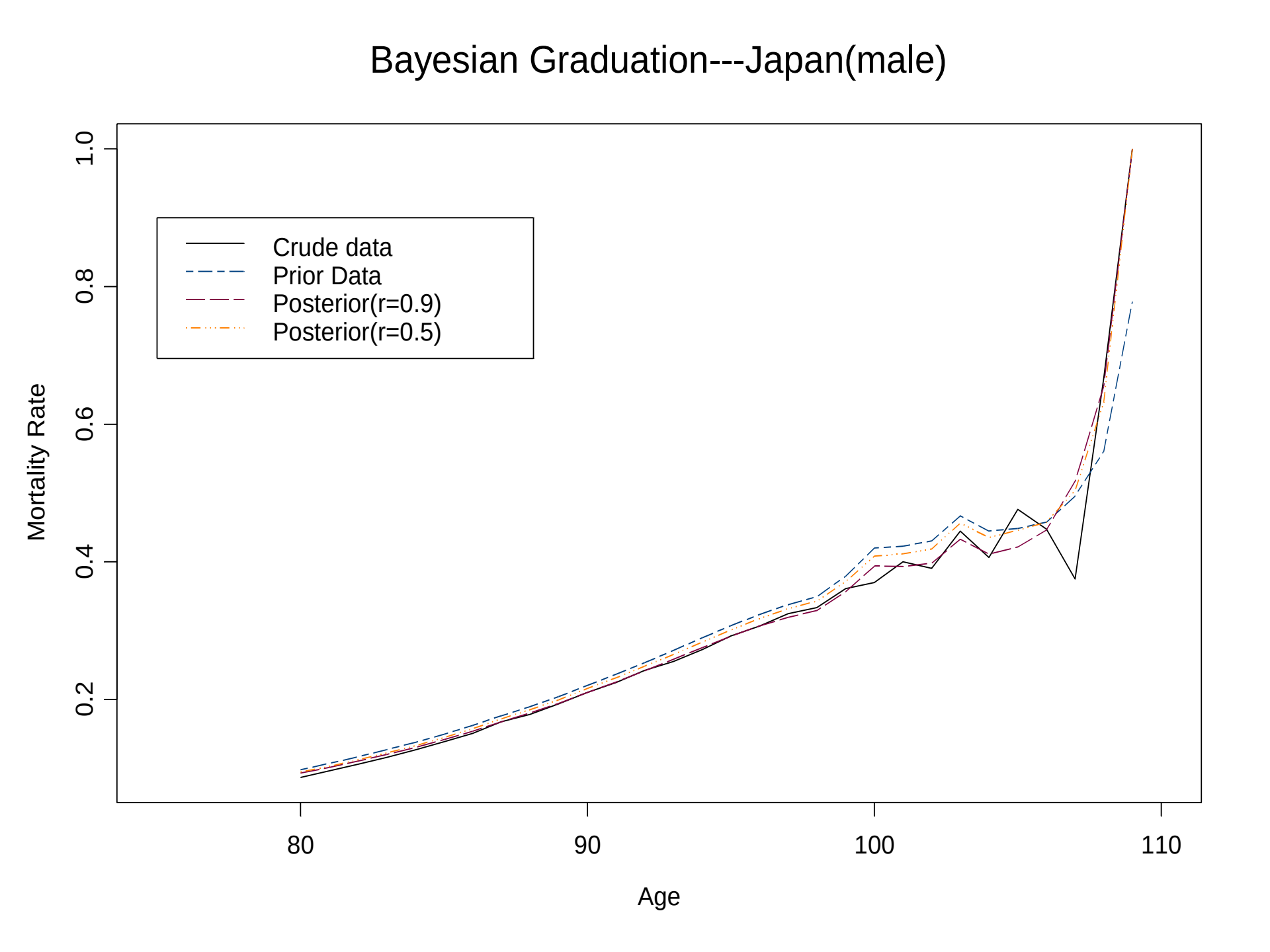


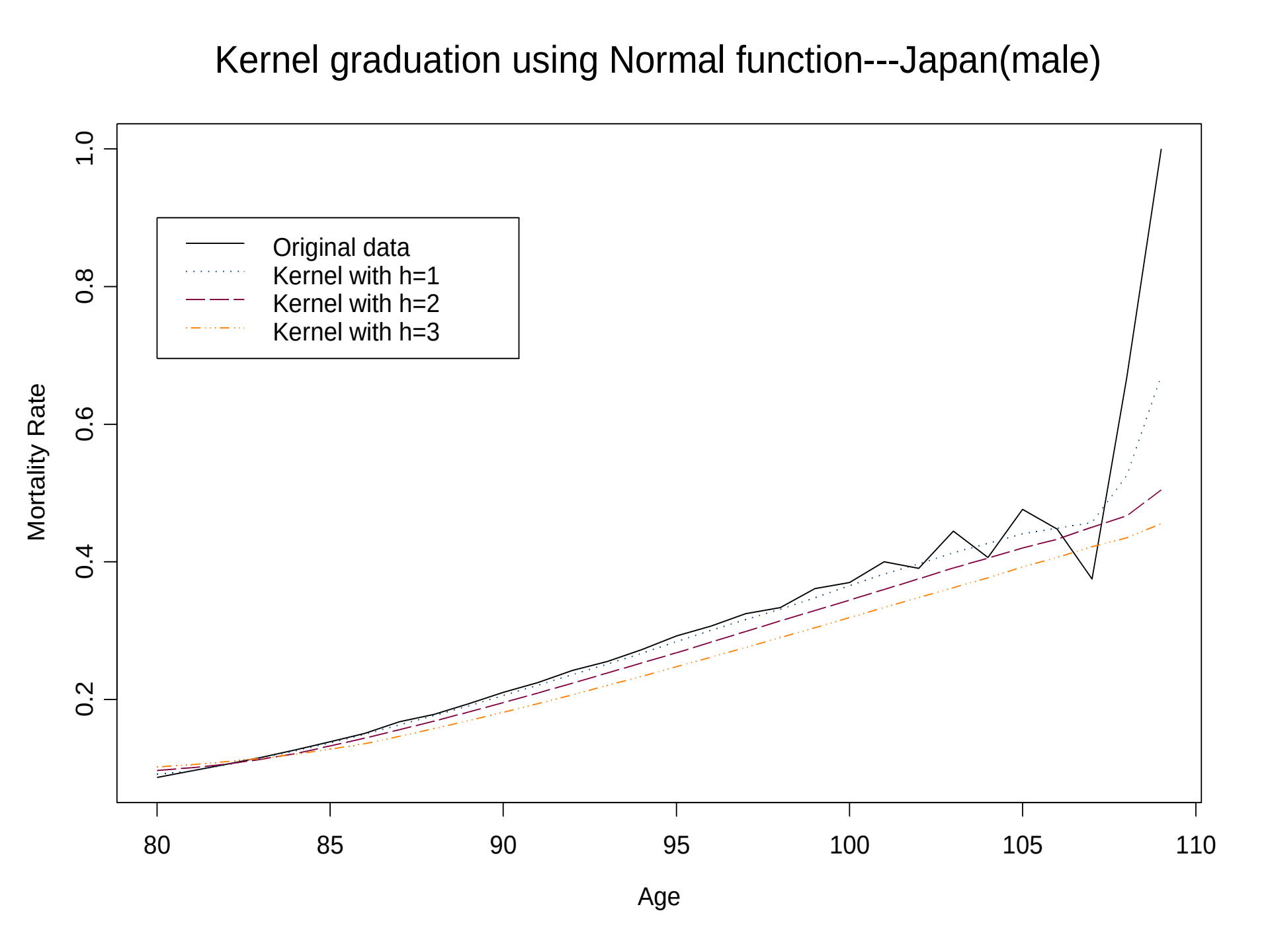


男性死亡率改善幅度在各年齡不盡相同

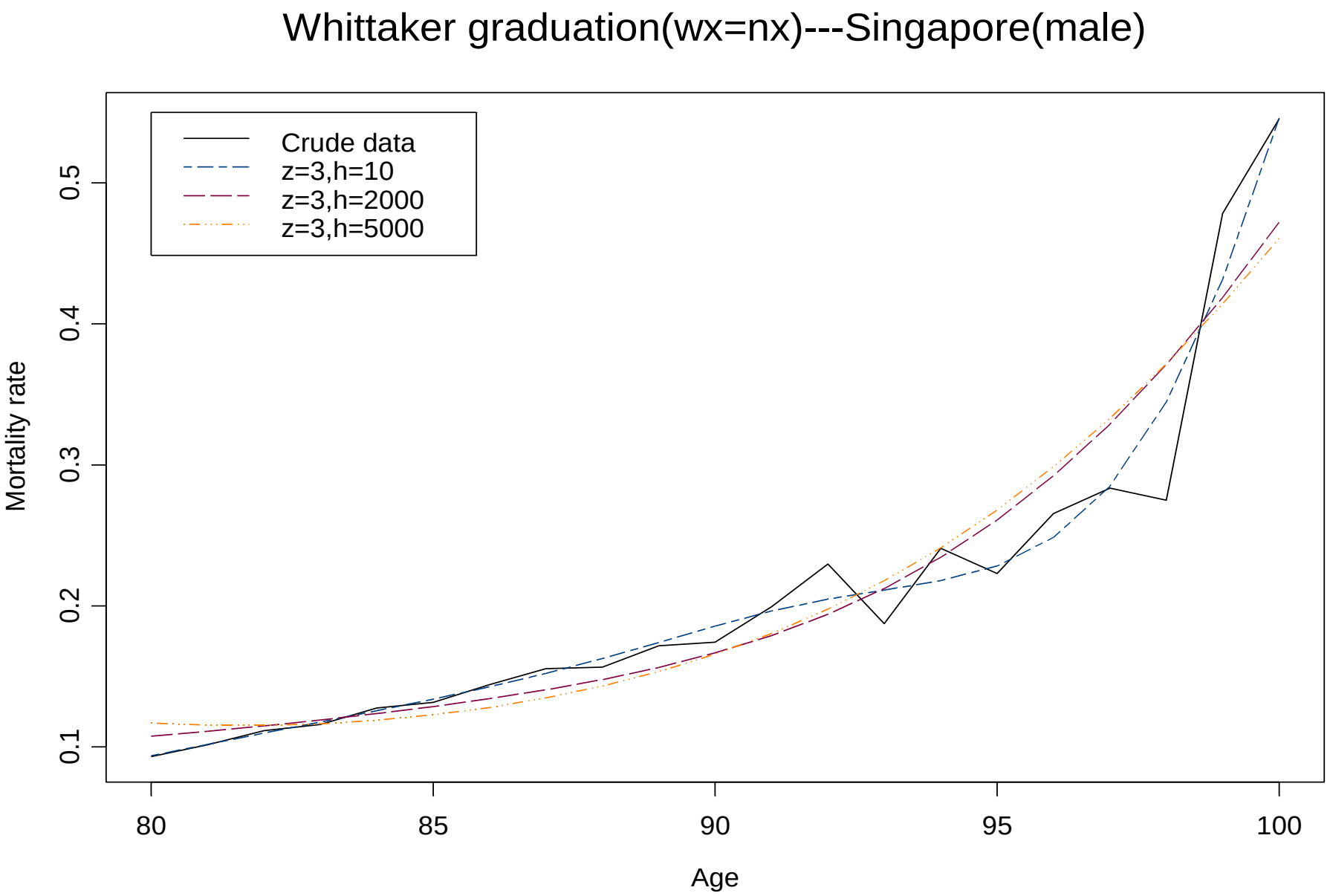
範例三、日本高齡男性死亡率(1980-1990)

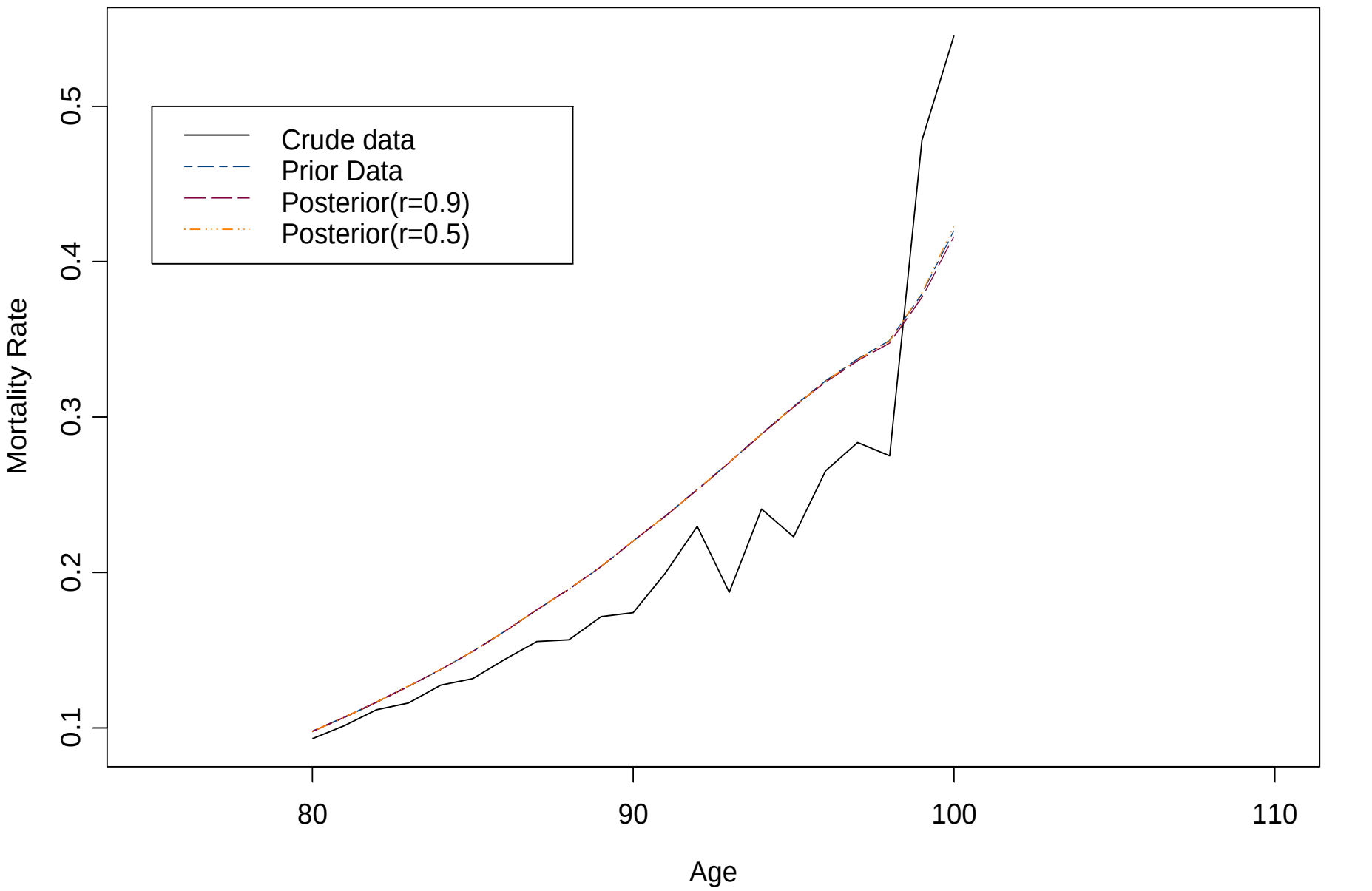


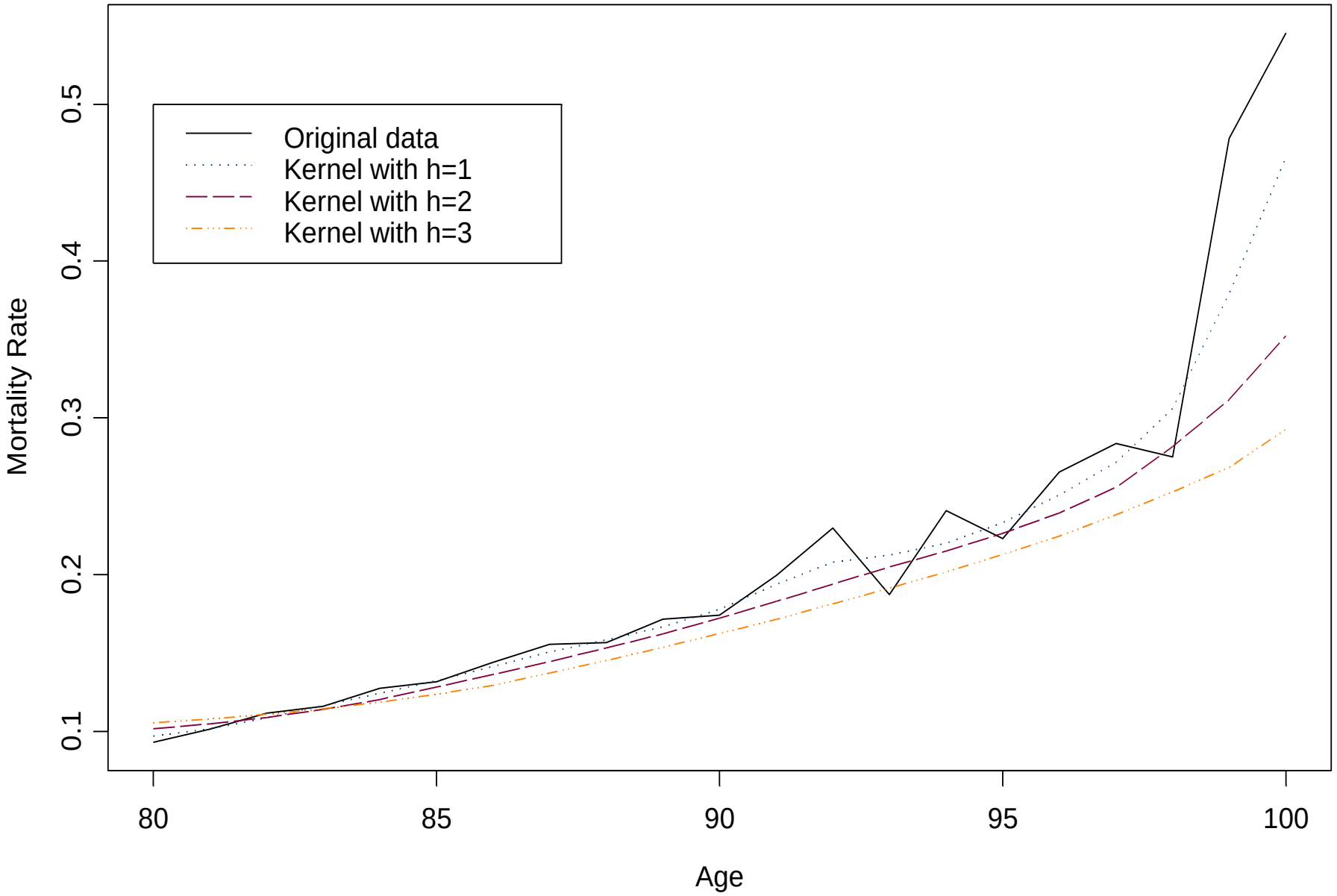

Bayesian Graduation---Japan(male)
Crude data
Prior Data
Posterior(r=0.9)
Posterior(r=0.5)
Mortality Rate
0.2
0.4
0.6
0.8
1.0
80
90
100
110
Age


Kernel graduation using Normal function---Japan(male)
Original data
Kernel with h=1
Kernel with h=2
Kernel with h=3
Mortality Rate
0.2
0.4
0.6
0.8
1.0
80
85
90
95
100
105
110
Age

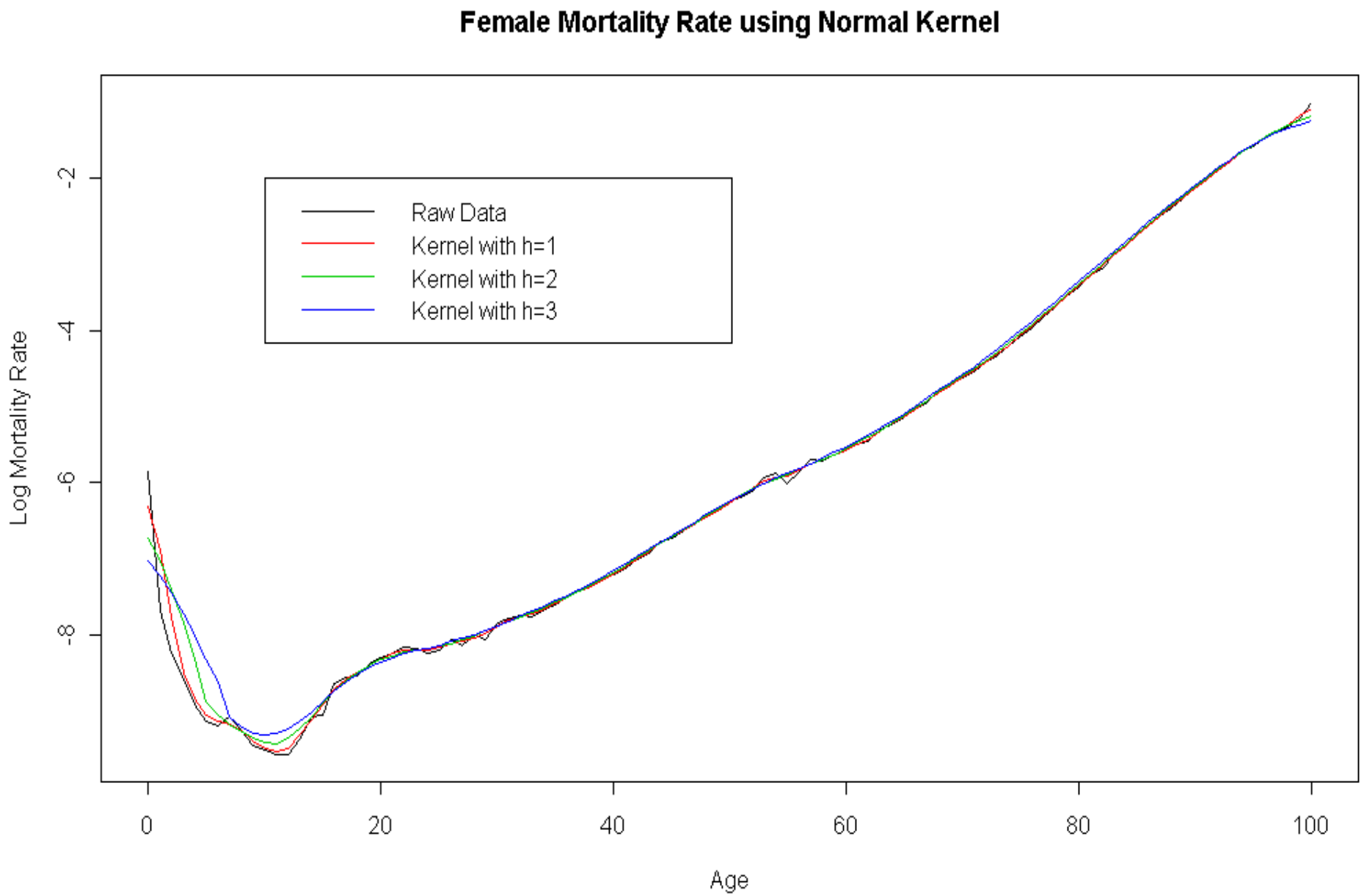
範例四、新加坡高齡男性死亡率(1980-1990)

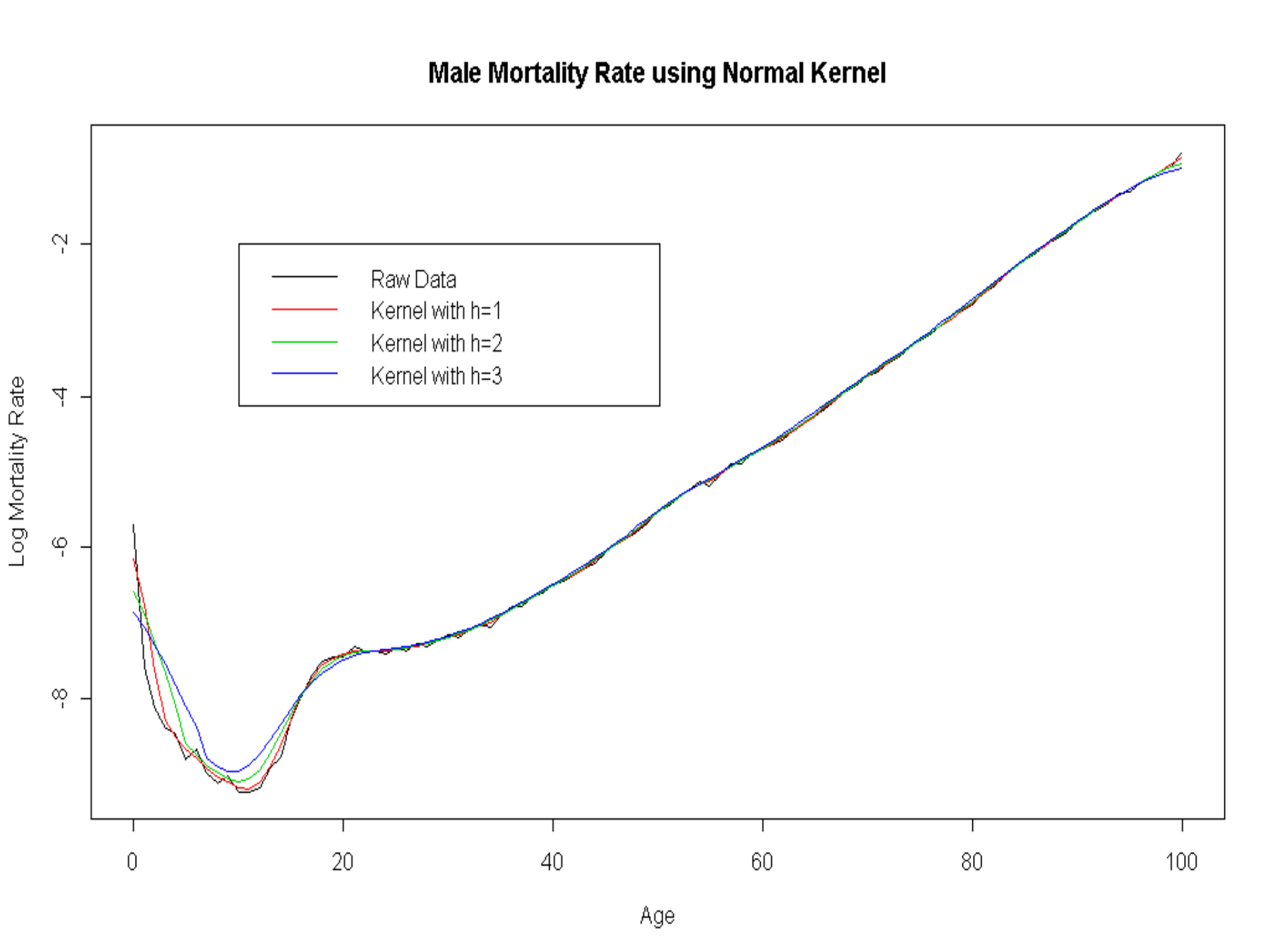



Bayesian Graduation---Singapore(male)
Crude data
Prior Data
Posterior(r=0.9)
Posterior(r=0.5)
Mortality Rate
0.1
0.2
0.3
0.4
0.5
80
90
100
110
Age


Kernel graduation using Normal function---Singapore(male)
Original data
Kernel with h=1
Kernel with h=2
Kernel with h=3
Mortality Rate
0.1
0.2
0.3
0.4
0.5
80
85
90
95
100
Age

範例五、日本2001年死亡率(HMD data)




Male Mortality Rate using Normal Kernel
Raw Data
Kernel with h=1
Kernel with h=2
Kernel with h=3
Log Mortality Rate
-2
-4
-6
-8
0
20
40
60
80
100
Age

- 最近鄰區估計值
  (Nearest-neighbor Estimator, NNE)

→核修勻法是以每個觀察值為中心，固定一個距離(「環寬」)，求出密度函數的估計值。可推廣類似想法，但以距離最近的觀察值求得估計值，也就是以待修勻的點為中心。

→例如：如果 x_0 緊鄰著 x ，則稱 x 是 x_0 的1-neighbor；若還有一個與 x_0 更接近的點，稱 x 是 x_0 的 2-neighbor。

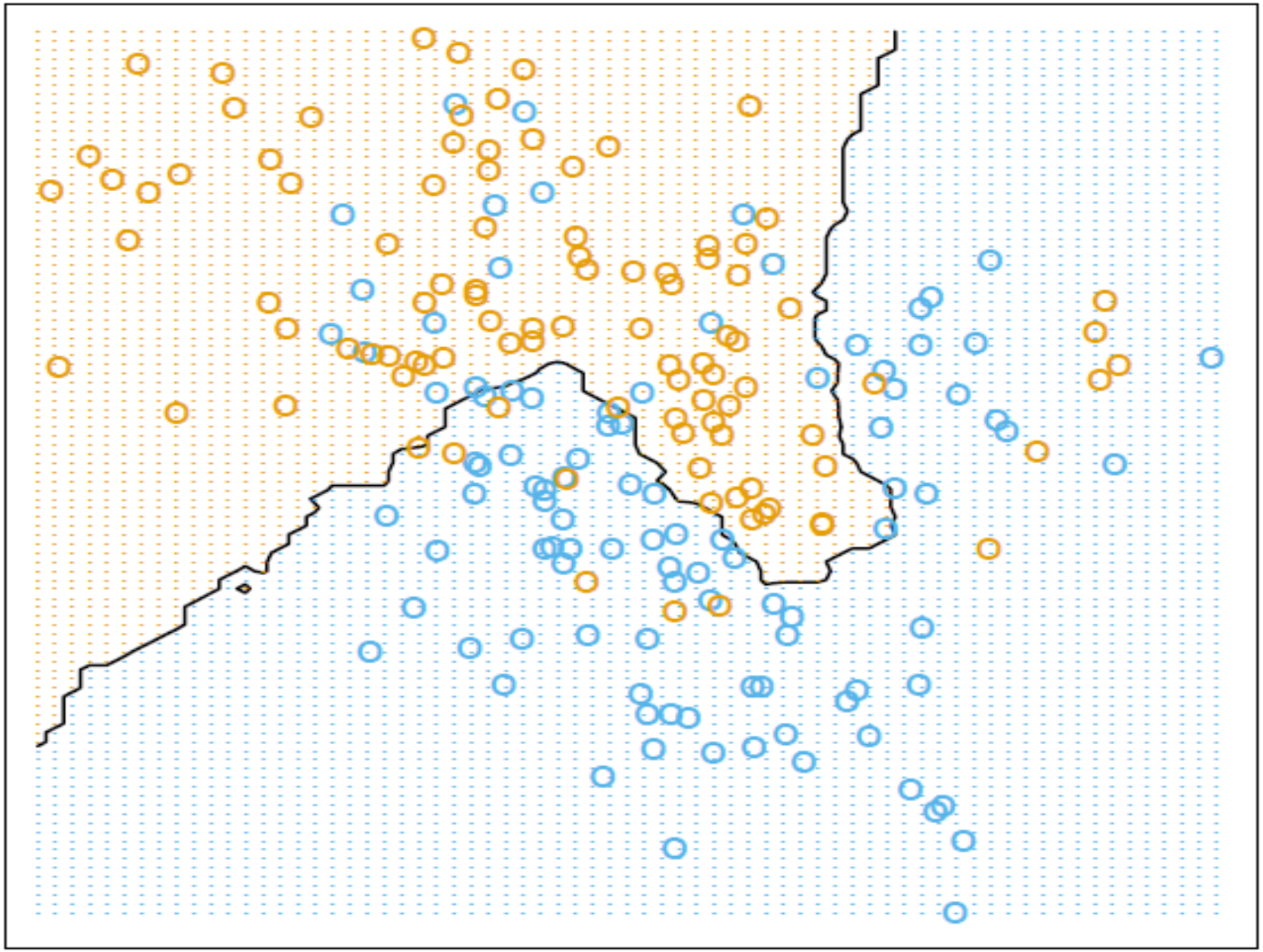
→ 以NNE的方式估計密度函數，選取最接近修勻點 x 的 k 個觀察值：

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_k(x)}K\left(\frac{x-x_i}{h_k(x)}\right),$$

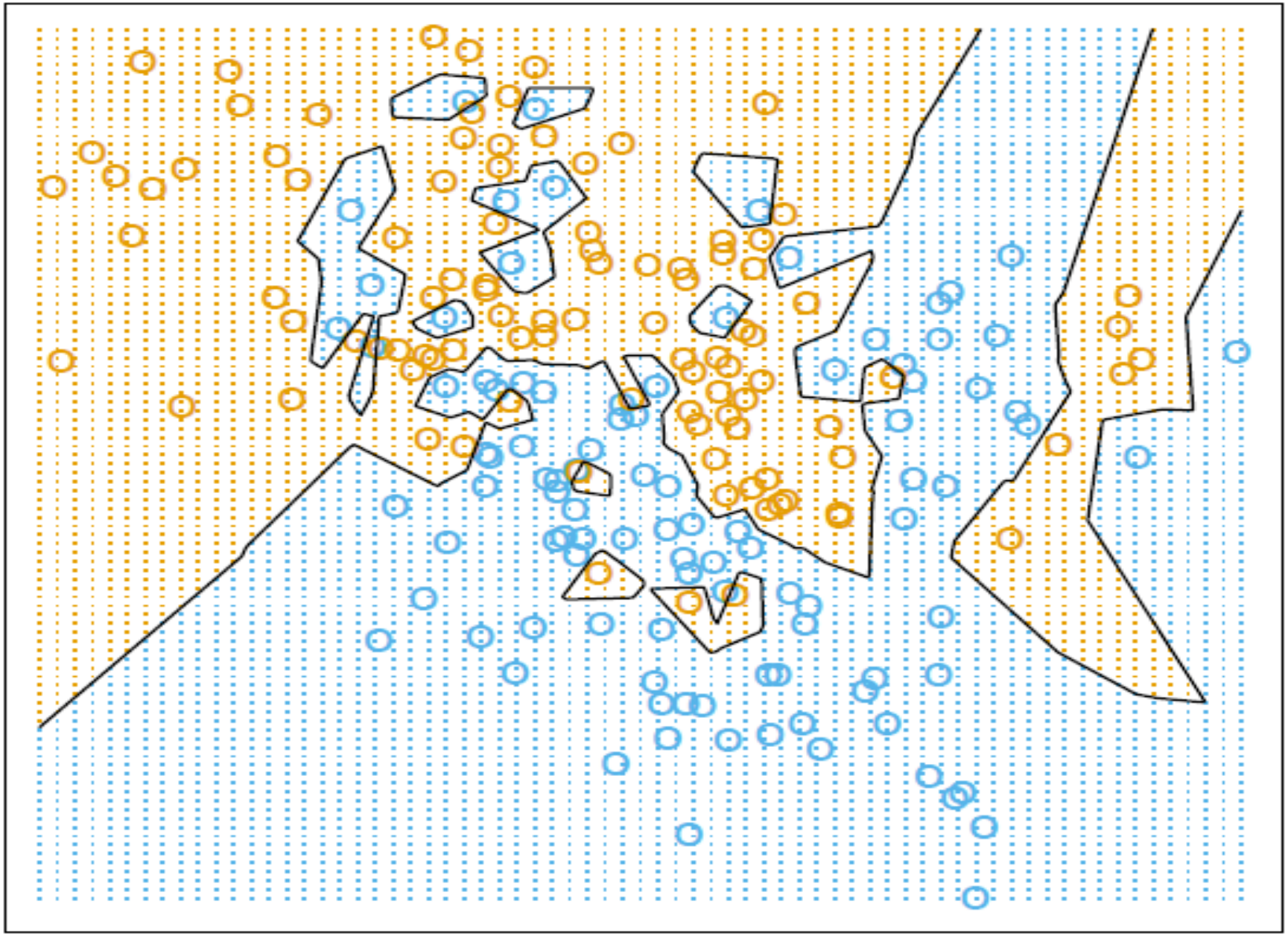
其中 $h_k(x)$ 以 x 為中心，可以包含最近的 k 個觀察值之最小半徑。

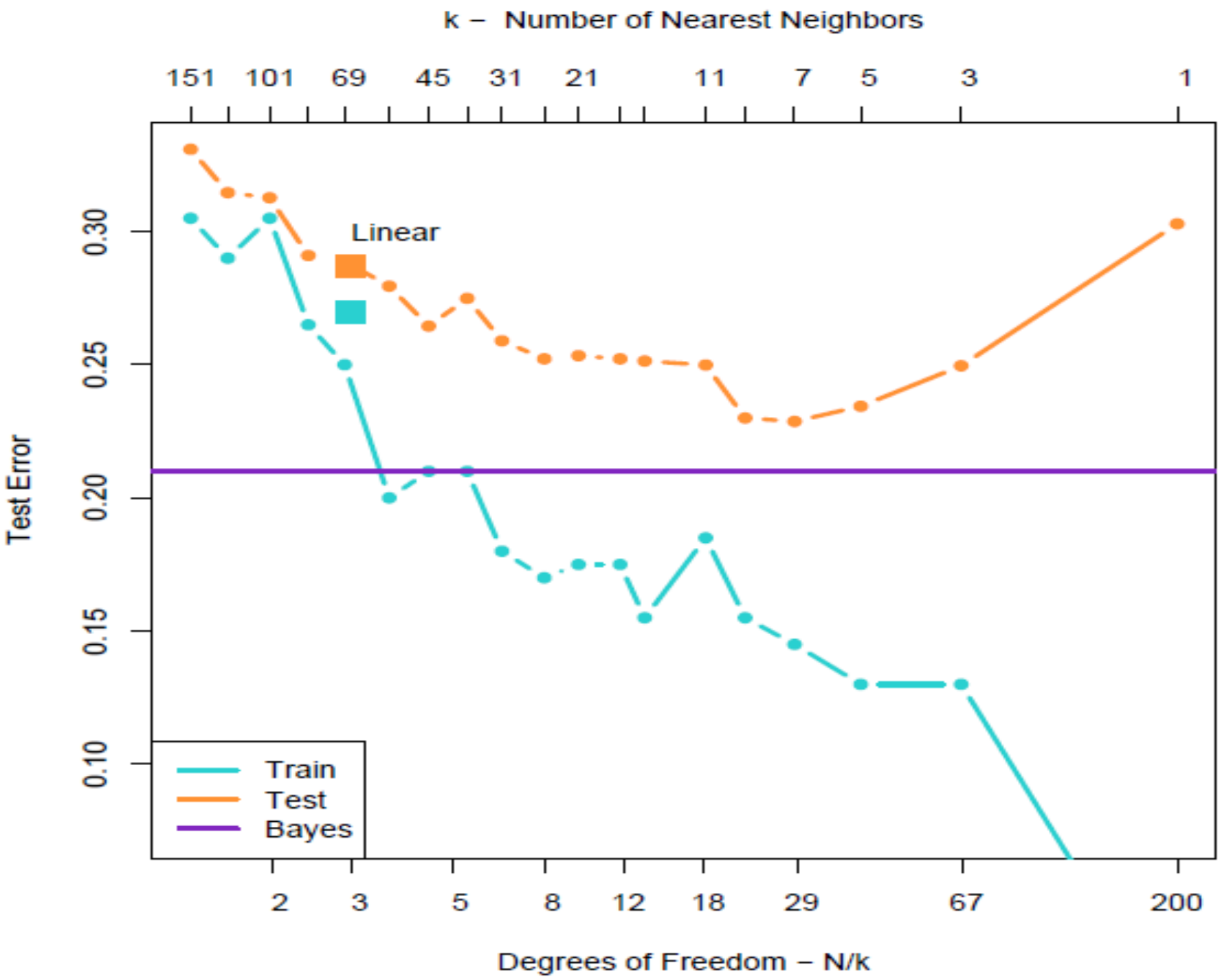
註：與核修勻法不同，NNE修勻的「環寬」不是定值，隨著 k 的選擇，修勻點、所有觀察值而改變。

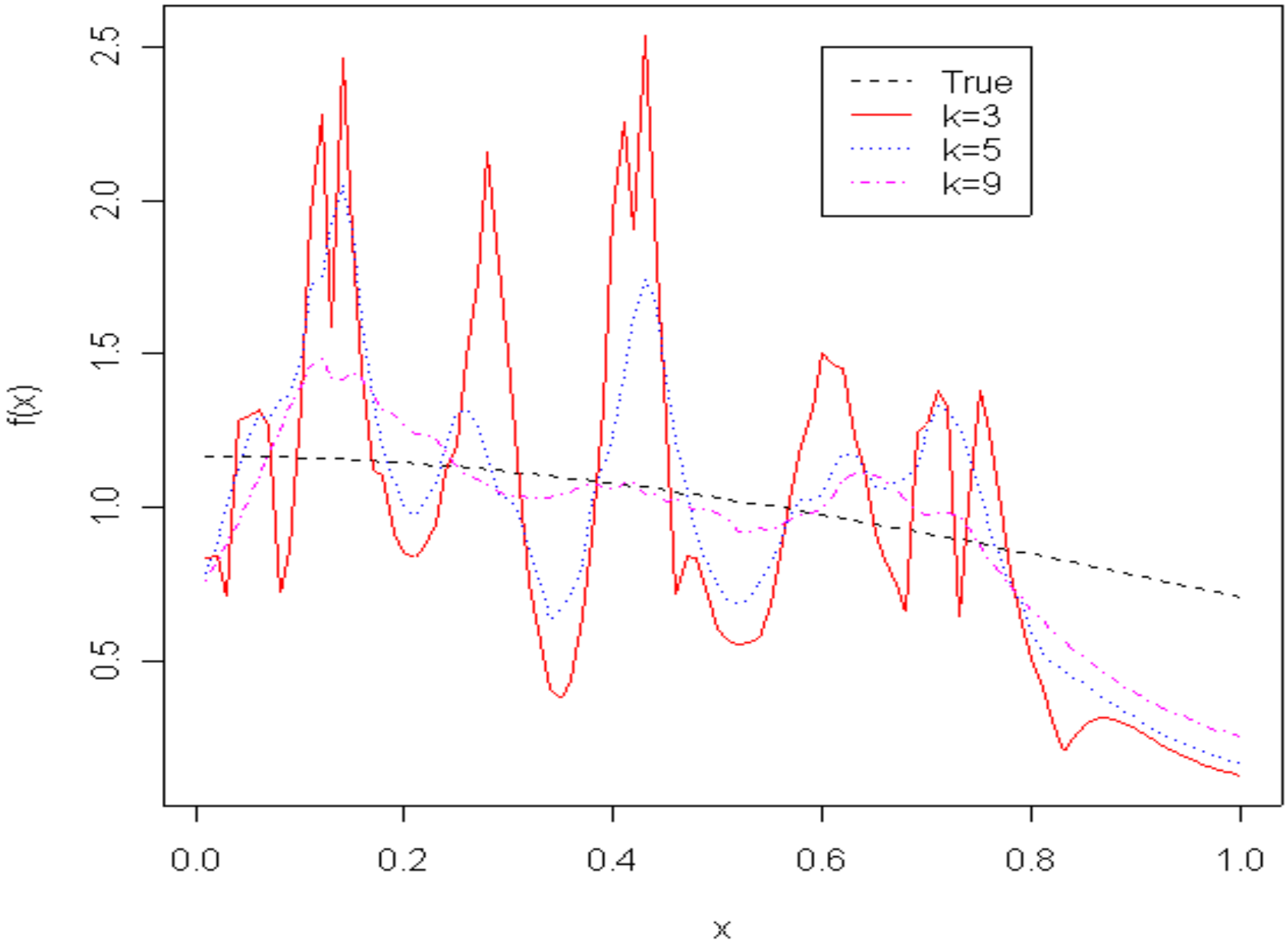
15-Nearest Neighbor Classifier



1−Nearest Neighbor Classifier




k – Number of Nearest Neighbors
151
101
69
45
31
21
11
7
5
3
1
Linear
Test Error
0.30
0.25
0.20
0.15
0.10
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
Degrees of Freedom – N/k


Nearest-neighbor Estimates (n=26, N(0,1))
True
k=3
k=5
k=9
f(x)
x
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0

- 移動平均(Running Means)

→移動平均類似核修匀法，藉由另一個解釋變數X計算目標函數Y的數值，像是計算某一年齡 x_j 的死亡率時，會將 x_j 附近的死亡率 y_j's 平均。

→ 對於 x_j 附近的選擇可用NNE的想法，也就是選擇所有死亡率y_j's，其中解釋變數 x_j 滿足 $|i-j| \le k$。這種選取方式可將左右兩邊各 k 個死亡率皆列入平均，可預期在幼齡及高齡兩端，死亡率應該也會較不平滑。

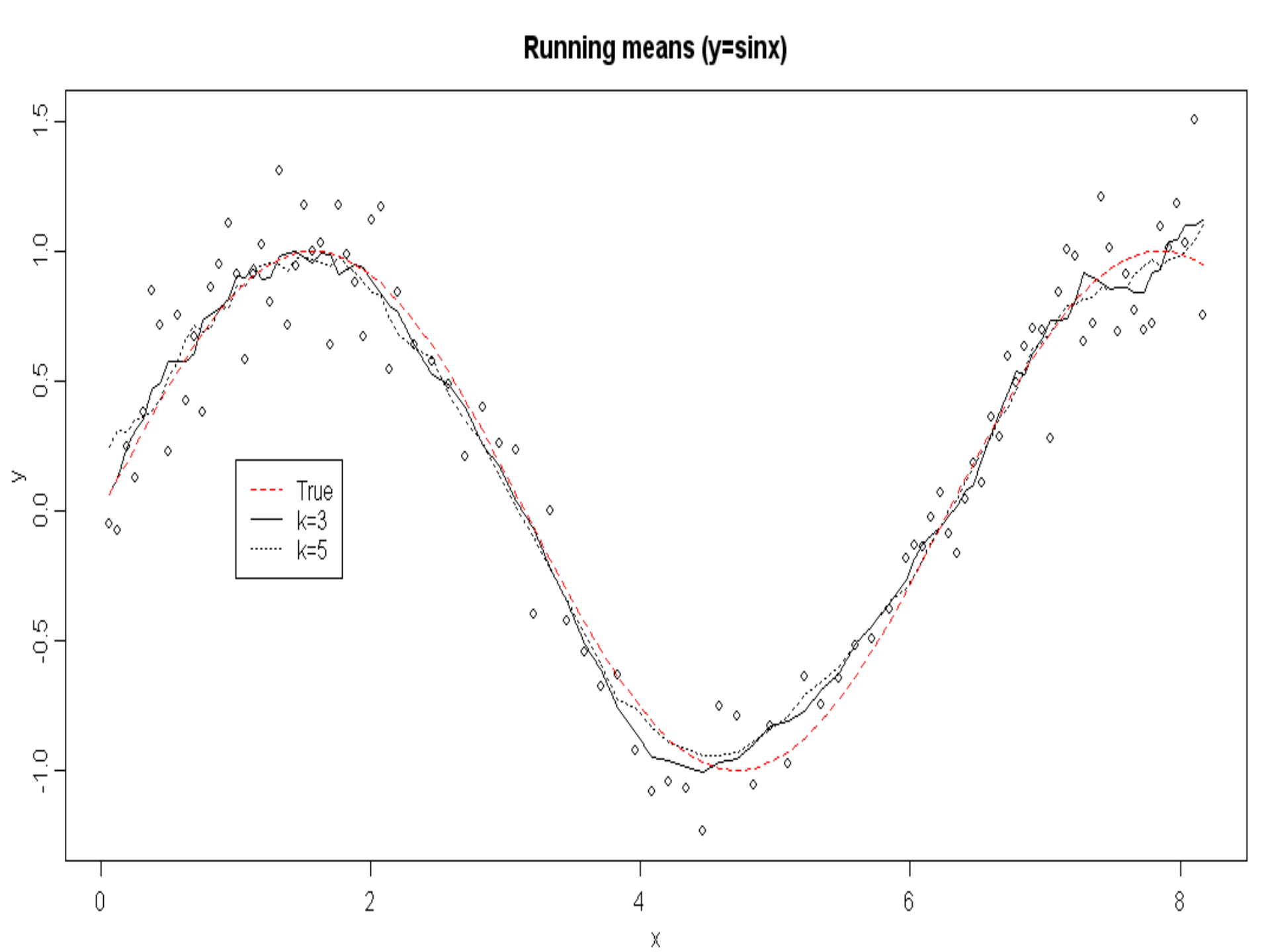
註：移動平均的參數 k 稱為伸展(S*pan*)，主宰修勻值的平滑程度。

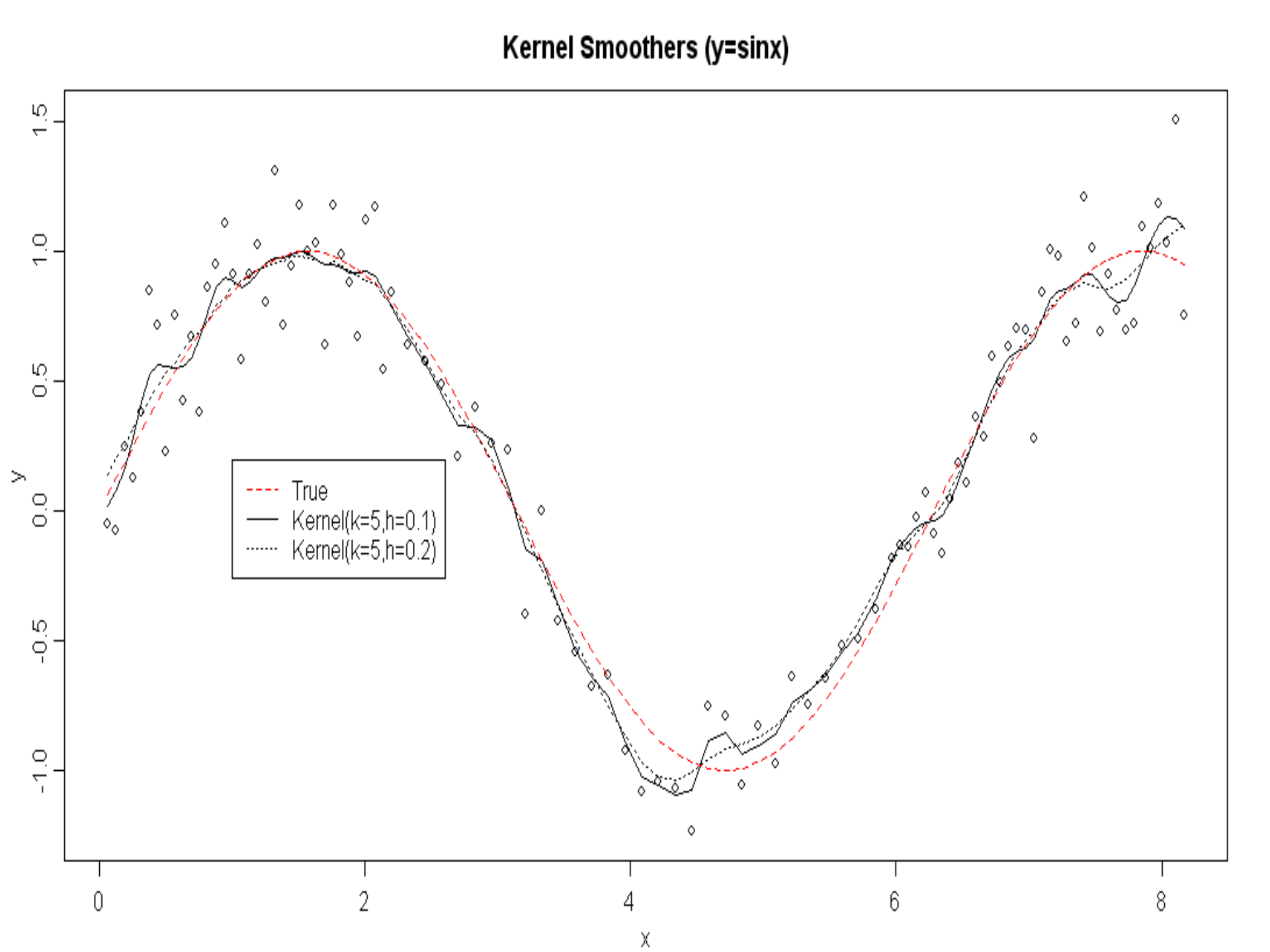
- 範例六、 假設資料服從下列分配：

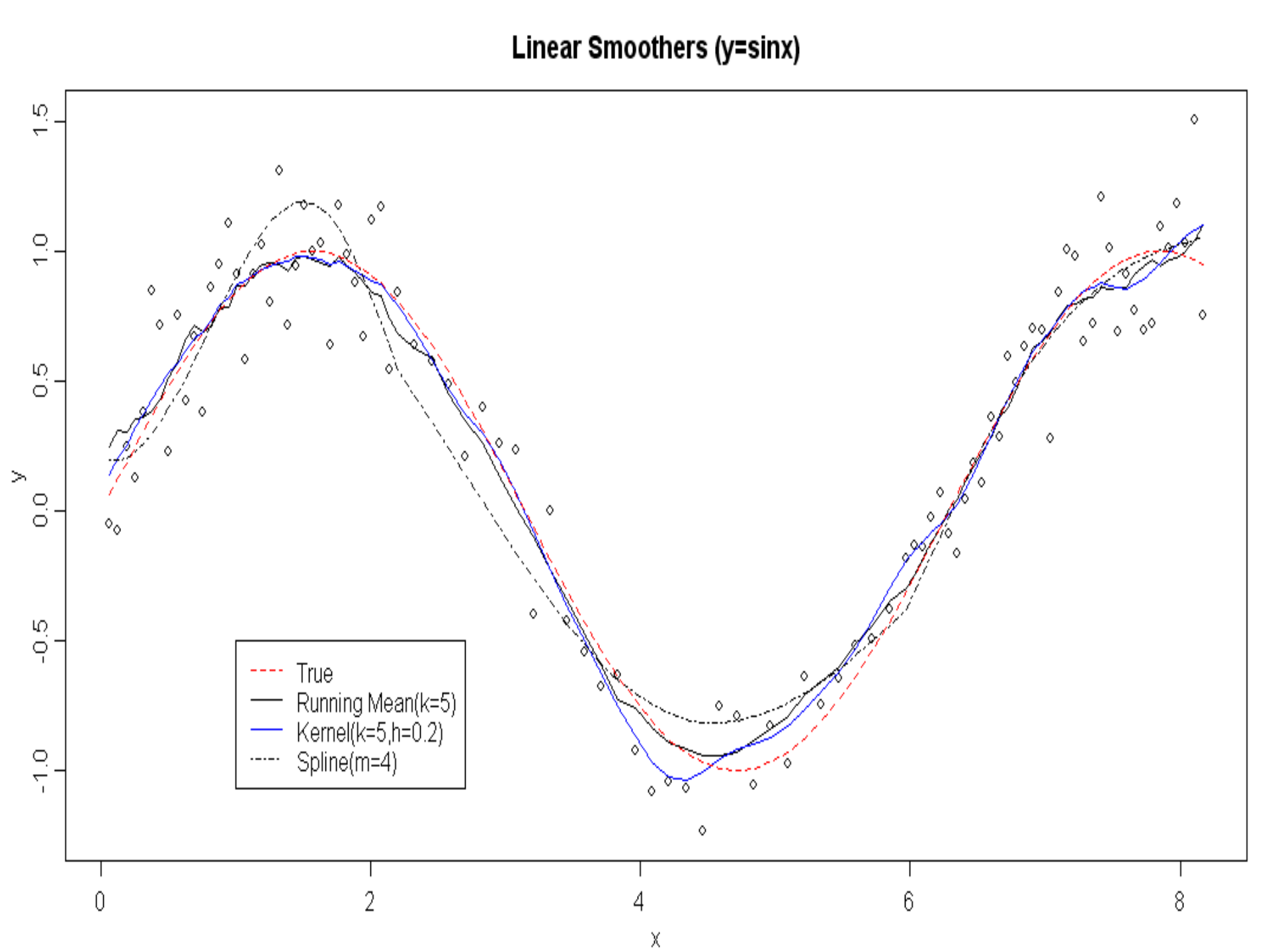
$$Y_i = \sin X_i + \varepsilon_i, \quad 0 \le X_i \le \pi,$$

其中誤差項 ε_i 服從常態分配，期望值 0、變異數 0.04。另外， X 在[0,0.3π]有15個點、 在[0.3 π,0.7π]有10個點、在[0.7π,π]有 15個點。

→我們將以這組資料比較移動平均、核修勻法、Spline法。


Running means (y=sinx)
True
k=3
k=5
y
x
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8


Kernel Smoothers (y=sinx)
True
Kernel(k=5,h=0.1)
Kernel(k=5,h=0.2)
y
x
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8


Linear Smoothers (y=sinx)
1.5
1.0
0.5
0.0
-0.5
-1.0
y
0
2
4
6
8
x
True
Running Mean(k=5)
Kernel(k=5,h=0.2)
Spline(m=4)