Μη Παραμετρική Εκτίμηση Συνάρτησης Πυκνότητας Πιθανότητας

- Ιστόγραμμα
- Εκτίμηση με Πυρήνες

Μη Παραμετρική Στατιστική

Παναγιώτης Παπασταμούλης
Επίκουρος Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

Παρασκευή, 15/05/2020

Περιεχόμενα

1. Εισαγωγή στις μεθόδους εξομάλυνσης
   - Ολοκληρωμένο μέσο τετραγωνικό σφάλμα
   - Cross validation
2. Ιστογραμμα
   - Ορισμός
   - Ιδιότητες ιστογράμματος
   - Επιλογή h μέσω IMSE
   - Η εντολή `hist()` στην `R`
3. Πυρήνες
   - Ο απλοϊκός εκτιμητής (boxcar kernel)
   - Πυρήνες
   - Επιλογή του h μέσω IMSE
   - Η συνάρτηση `density()` στην `R`
   - Επιλογή h μέσω cross-validated πιθανοφάνειας
   - Bootstrap διάστημα εμπιστοσύνης
4. Περαιτέρω σχόλια

Διάφοροι εκτιμητές συνάρτησης πυκνότητας

- Παραμετρικές μέθοδοι
  - Υποθέτουμε ότι $f(x) \in \mathcal{F} = \{f(\cdot;\theta); \theta \in \Theta\}$
  - Αρκεί να εκτιμηθεί το θ (πχ: ΕΜΠ, ΑΟΕΔ, ΕΜΡ κλπ)
- Μη παραμετρικές μέθοδοι
  - Ιστόγραμμα
  - Χρήση πυρήνα (kernels)
  - Άλλες μέθοδοι: splines, ορθογώνιες σειρές, χρήση πολυωνύμων κλπ
  - Εδώ υποθέτουμε ότι $f(x)$ «ομαλή»
  - Θεωρούμε κατάλληλες κλάσεις εκτιμητών που εξαρτώνται από κάποια «παράμετρο εξομάλυνσης» h
  - Αρκεί να επιλεχθεί η παράμετρος εξομάλυνσης
- Ημιπαραμετρική μέθοδος
  - Μείξεις κατανομών (mixtures of distributions)
  - Υποθέτουμε ότι η $f(x)$ εκφράζεται σαν σταθμισμένος μέσος όρος διάφορων παραμετρικών μοντέλων

- Στο μάθημα θα περιγράψουμε την εκτίμηση μέσω ιστογράμματος και με χρήση πυρήνα
- Στη διάθεσή μας έχουμε ένα τυχαίο δείγμα $X_1, \ldots, X_n$ από την (άγνωστη) $f(x)$
- Συμβολισμός: εκτιμητής πυκνότητας

$$\hat{f}_{h,n}(x)$$

με παράμετρο εξομάλυνσης h ο οποίος βασίζεται σε τυχαίο δείγμα μεγέθους n

- Στον συμβολισμό παραλείπουμε να τονίσουμε την εξάρτηση από τα δεδομένα $(X_1, \ldots, X_n)$:

$$\hat{f}_{h,n}(x) := \hat{f}_{h,n}(X_1, \ldots, X_n; x)$$

- $\mathrm{E}\hat{f}_{h,n}(x)$ μέση τιμή της $\hat{f}_{h,n}$ (συνάρτηση του x)
- $\mathrm{Var}\hat{f}_{h,n}(x)$ διασπορά της $\hat{f}_{h,n}$ (συνάρτηση του x)
- Για παράδειγμα, για σταθερό x:

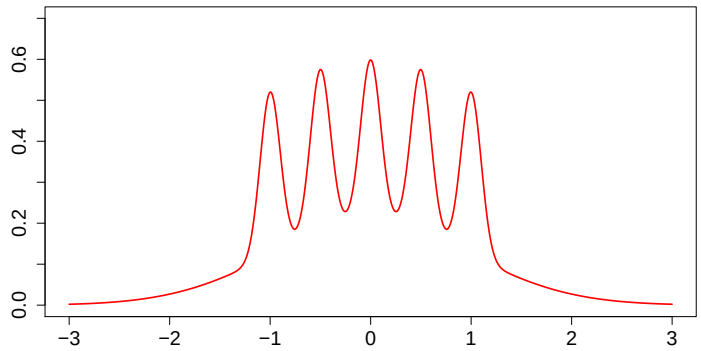
$$\begin{aligned}
\mathrm{E}\hat{f}_{h,n}(x) &= \mathrm{E}\hat{f}_{h,n}(X_1, \ldots, X_n; x) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{f}_{h,n}(x_1, \ldots, x_n; x) \prod_{i=1}^{n} f(x_i) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&=: \bar{f}_h(x) \quad (\text{θα δούμε γιατί δεν μπαίνει το } n \text{ ως δείκτης εδώ})
\end{aligned}$$

- Για σταθερά $x_1, \ldots, x_n$ και μεταβλητό X

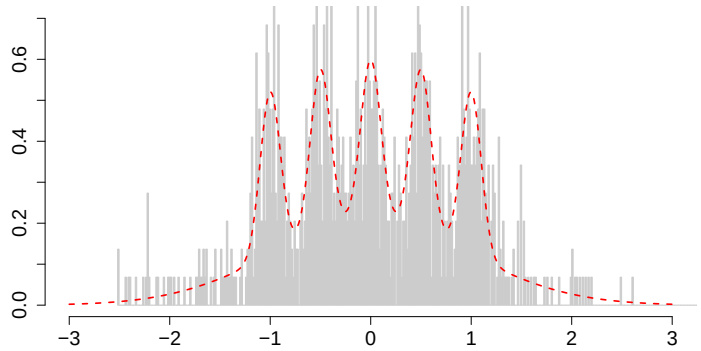
$$\begin{aligned}
\mathrm{E}\hat{f}_{h,n}(X) &:= \mathrm{E}\hat{f}_{h,n}(x_1, \ldots, x_n; X) \\
&= \int_{-\infty}^{\infty} \hat{f}_{h,n}(x) f(x) \mathrm{d}x.
\end{aligned}$$

Εισαγωγή στην εκτίμηση συνάρτησης πυκνότητας

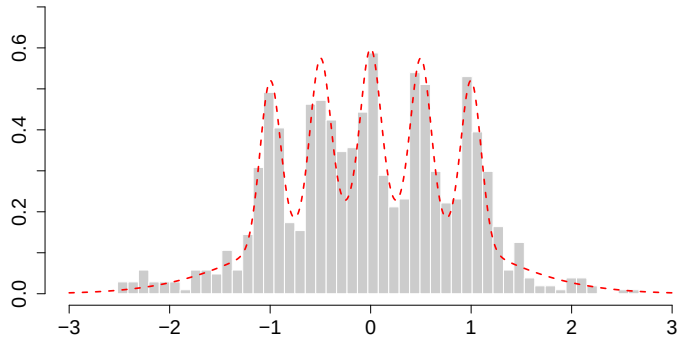
- Παράδειγμα: $n = 1000$ παρατηρήσεις από $f(x)$
- Εκτίμηση $f(x)$ μέσω **ιστογράμματος** $\hat{f}_{h,n}(x)$ με πλάτος κελιών h
- Το πλάτος κελιών (ισοδύναμα: το πλήθος των κελιών) αποτελεί **παράμετρο εξομάλυνσης** (smoothing parameter)



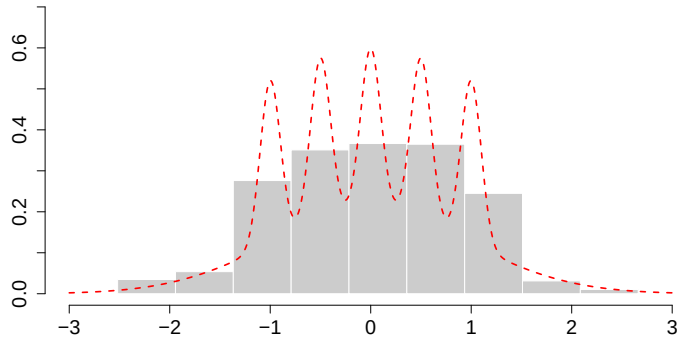
πραγματική $f(x)$



$\hat{f}_{0.015,n}$: πολλά κελιά (undersmoothed)



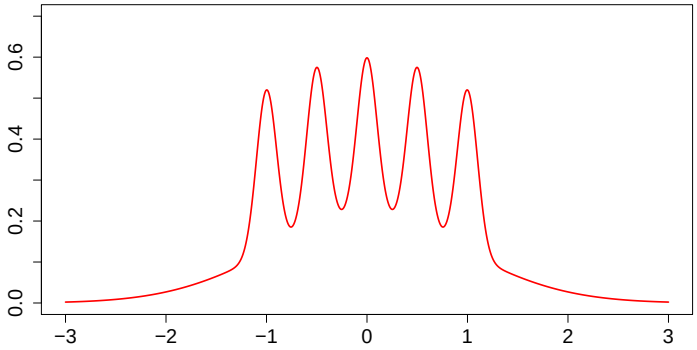
$\hat{f}_{0.104,n}$: όσα πρέπει



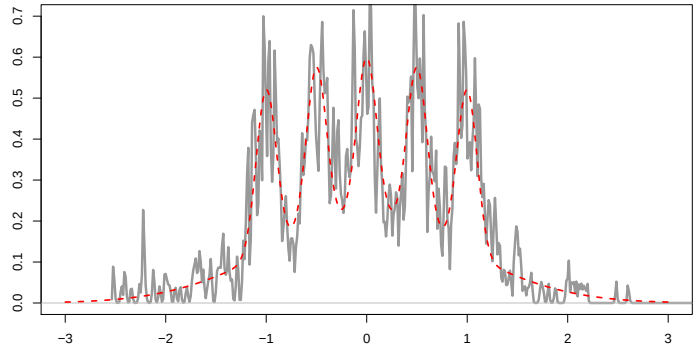
$\hat{f}_{0.575,n}$: λίγα κελιά (oversmoothed)

Εισαγωγή στην εκτίμηση συνάρτησης πυκνότητας

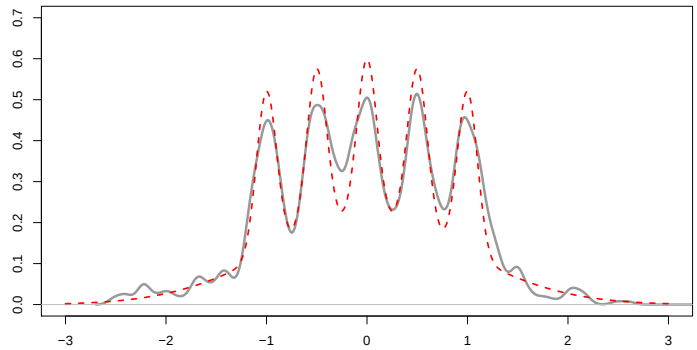
- Παράδειγμα: $n = 1000$ παρατηρήσεις από $f(x)$
- Εκτίμηση $f(x)$ με χρήση (κανονικού) **πυρήνα** $\hat{f}_{h,n}(x)$ με bandwidth h



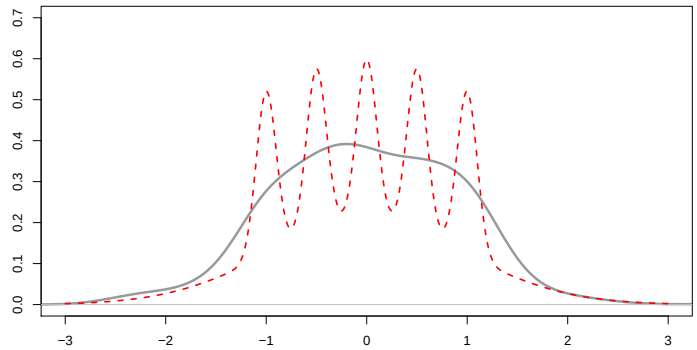
πραγματική $f(x)$



$\hat{f}_{0.006,n}$: μικρό h (undersmoothed)



$\hat{f}_{0.059,n}$: όσο πρέπει



$\hat{f}_{0.295,n}$: μεγάλο h (oversmoothed)

Παρατηρήσεις

- Κρίσιμη η επιλογή της παραμέτρου εξομάλυνσης
  - Στο ιστόγραμμα είναι το εύρος των κελιών
  - Στον εκτιμητή με χρήση πυρήνα είναι η διασπορά του πυρήνα
- Μεγάλο h: κρύβονται υπάρχουσες δομές και άρα **αύξηση της μεροληψίας** (oversmoothing)
- Μικρό h: εμφανίζονται τυχαίες αποκλίσεις ως υπάρχουσες δομές, που οφείλεται στην **μεγάλη διασπορά** του εκτιμητή (undersmoothing)
- Στόχος: επιλογή h (ούτε πολύ «μικρό» ούτε πολύ «μεγάλο») ώστε να ισοσταθμίζεται η διασπορά και η μεροληψία του εκτιμητή.
- Κατάλληλο κριτήριο: ολοκληρωμένο μέσο τετραγωνικό σφάλμα (Integrated Mean Square Error)
- Τεχνική για την εκτίμησή του: **cross-validation** (διεπικύρωση)

Κριτήριο σφάλματος

- Μεροληψία
$$\text{bias}(\hat{f}_{h,n}(x)) = \text{E}\hat{f}_{h,n}(x) - f(x)$$
- Διασπορά
$$\text{Var}(\hat{f}_{h,n}(x)) = \text{E}\left\{\hat{f}_{h,n}(x) - \text{E}\hat{f}_{h,n}(x)\right\}^2$$
- Μέσο τετραγωνικό σφάλμα (Mean Square Error)
$$\begin{aligned}\text{MSE}(\hat{f}_{h,n}(x)) &= \text{E}\left\{\hat{f}_{h,n}(x) - f(x)\right\}^2 \\ &= \text{bias}^2(\hat{f}_{h,n}(x)) + \text{Var}(\hat{f}_{h,n}(x))\end{aligned}$$
- Το MSE όμως αφορά μία συγκεκριμένη τιμή του x
- Πρέπει να λαβουμε υπόψη όλες τις δυνατές τιμές του x
- Ολοκληρωμένο μέσο τετραγωνικό σφάλμα (Mean Integrated Square Error)

Ορισμός (Ολοκληρωμένο μέσο τετραγωνικό σφάλμα)

Το ολοκληρωμένο μέσο τετραγωνικό σφάλμα της $\hat{f}_{h,n}(x)$ που χρησιμοποιείται για την εκτίμηση της $f(x)$ είναι

$$\begin{aligned} \mathrm{IMSE}(\hat{f}_{h,n}(x)) &= \int_{-\infty}^{\infty} \mathrm{MSE}(\hat{f}_{h,n}(x))\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \mathrm{bias}^2(\hat{f}_{h,n}(x))\mathrm{d}x + \int_{-\infty}^{\infty} \mathrm{Var}(\hat{f}_{h,n}(x))\mathrm{d}x \end{aligned} \tag{1}$$

- Το ολοκληρωμένο μέσο τετραγωνικό σφάλμα λαμβάνει υπόψη το MSE για κάθε x
- Ολοκληρώνει ως προς όλες τις δυνατές τιμές ώστε να υπολογιστεί το συνολικό σφάλμα

Ελαχιστοποίηση του IMSE

Από την (1) έχουμε

$$\begin{aligned}
\mathrm{IMSE}(\hat{f}_{h,n}(x)) &= \int_{-\infty}^{\infty} \mathrm{E}\left\{\hat{f}_{h,n}(x) - f(x)\right\}^2 \mathrm{d}x \\
&= \mathrm{E}\left[\int_{-\infty}^{\infty} \left\{\hat{f}_{h,n}(x) - f(x)\right\}^2 \mathrm{d}x\right] \\
&= \mathrm{E}\left[\int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - 2\int_{-\infty}^{\infty} \hat{f}_{h,n}(x) f(x)\mathrm{d}x + \int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right]
\end{aligned}$$

Ο τελευταίος όρος δεν εξαρτάται από το h, οπότε αρκεί να ελαχιστοποιήσουμε την

$$J(h) := \mathrm{E}\left[\int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - 2\int_{-\infty}^{\infty} \hat{f}_{h,n}(x) f(x)\mathrm{d}x\right] \tag{2}$$

Πώς θα εκτιμήσουμε το $J(h)$;

- Είναι σαφές ότι το $J(h)$ είναι άγνωστο
- Οπότε πρέπει πρώτα να εκτιμηθεί και μετά να ελαχιστοποιηθεί ως προς h
- Μία πρώτη ιδέα είναι να χρησιμοποιήσουμε τον εκτιμητή αντικατάστασης

$$\tilde{J}(h) = \int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,n}(x_i) \tag{3}$$

- Δεν είναι καλός εκτιμητής
- Ο λόγος είναι ότι γίνεται διπλή χρήση των δεδομένων:
  1. Για να εκτιμηθεί το μοντέλο
  2. Για να εκτιμηθεί το μέσο σφάλμα

  $\Rightarrow$ overfitting
- Ο $\tilde{J}(h)$ είναι μεροληπτικός και πάντα θα υποδεικνύει ότι το μέσο σφάλμα ελαχιστοποιείται για $h \approx 0$

Αμερόληπτος εκτιμητής του $J(h)$

- Η Leave-One-Out Cross Validation εκτίμηση του $J(h)$ είναι

$$\hat{J}(h) = \int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,n-1}^{(-i)}(x_i) \tag{4}$$

- Η $\hat{f}_{h,n-1}^{(-i)}(x_i)$ είναι η εκτίμηση της πυκνότητας που προκύπτει αφαιρώντας την i παρατήρηση του δείγματος, $i = 1, \ldots, n$
- Ο $\hat{J}(h)$ είναι αμερόληπτος εκτιμητής του $J(h)$

$$\mathrm{E}\hat{J}(h) = J(h).$$

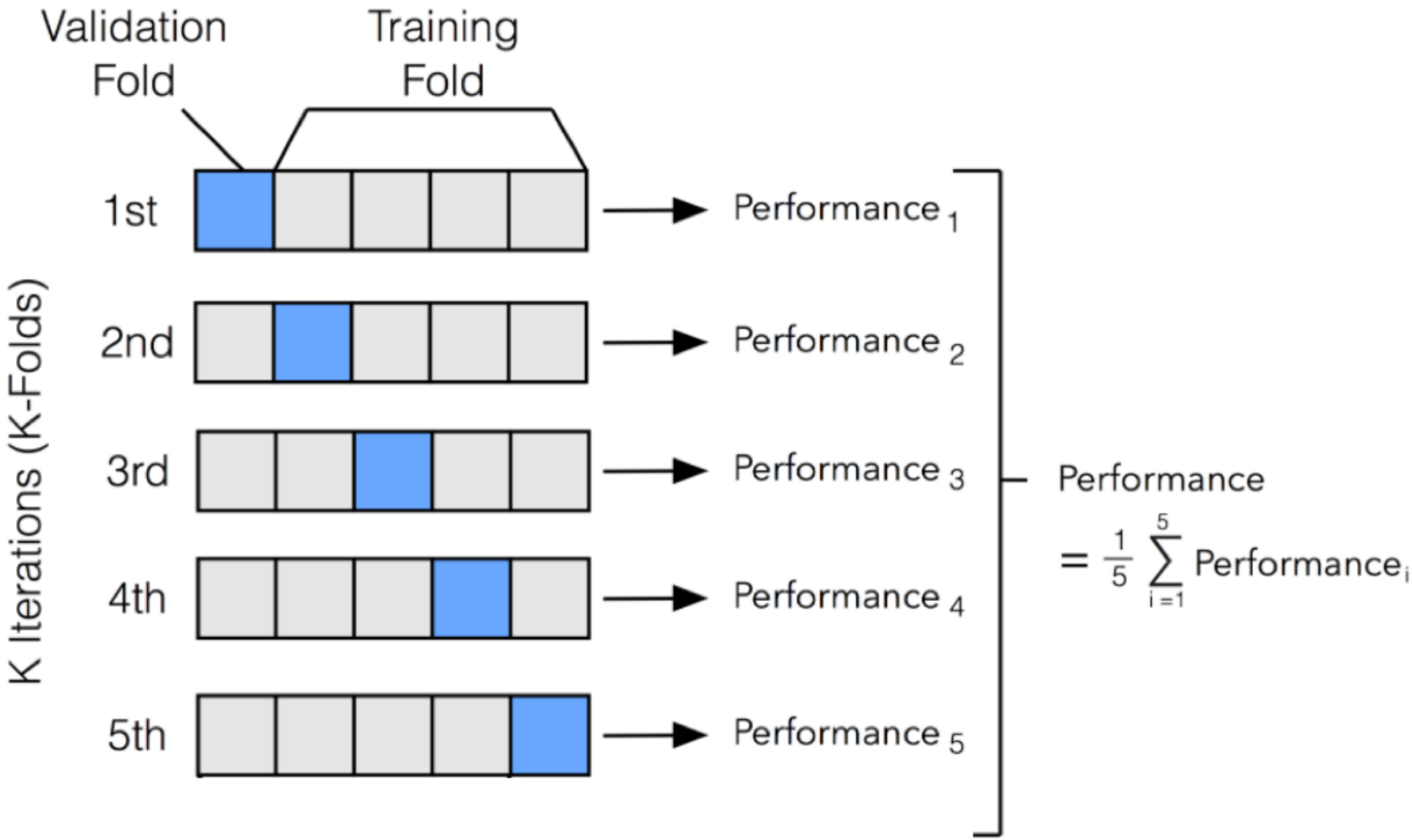
Cross-validation

- Η ιδέα είναι να εκτιμήσουμε το $\int_{-\infty}^{\infty} \hat{f}_{h,n}(x) f(x) \mathrm{d}x$ με cross-validation
- Τεχνική επαναδειγματοληψίας
- Βασίζεται στην ιδέα χωρισμού του δείγματος σε δύο τμήματα:
  - ένα για εκτίμηση του μοντέλου
  - ένα για αξιολόγηση του μοντέλου
- Το πρόβλημα με την παραπάνω τεχνική είναι ότι σπανίως έχουμε επαρκή δεδομένα για να το κάνουμε χωρίς σημαντική απώλεια πληροφορίας

Cross-validation

- Η μέθοδος cross-validation εφαρμόζει την ιδέα αυτή χωρίζοντας το δείγμα σε K ομάδες
- Χρησιμοποιεί $K - 1$ ομάδες για την εκτίμηση
- Το μοντέλο αξιολογείται μέσω της ομάδας που έμεινε εκτός
- Η διαδικασία επαναλαμβάνεται για κάθε μία ομάδα που μένει εκτός (K φορές)
- Το αποτέλεσμα προκύπτει σαν μέσος όρος όλων αυτών των επαναλήψεων

K-fold cross validation[1]



[1] εικόνα από http://ethen8181.github.io

Πόσες ομάδες;

- Τυπικές επιλογές είναι $K = 5, 10$ και n
- Όταν $K = n$ η τεχνική ονομάζεται Leave-One-Out Cross Validation (LOOCV)
  - Σε αυτήν την περίπτωση αφήνουμε 1 παρατήρηση εκτός και το μοντέλο εκτιμάται βάσει $n - 1$ παρατηρήσεων
  - Η παρατήρηση που έμεινε εκτός χρησιμοποιείται για την αξιολόγηση του μοντέλου
  - Επαναλαμβάνουμε αυτήν την διαδικασία n φορές, κάθε φορά αφήνοντας και μία διαφορετική παρατήρηση εκτός
  - Υπολογίζουμε τον μέσο όρο όλων των n εκτιμήσεων
- Υπό αυτήν την οπτική η τεχνική LOOCV «μοιάζει» με το jackknife, αλλά δεν είναι το ίδιο:
  - Το jackknife χρησιμοποιείται για εκτίμηση ενός μοντέλου
  - Το cross validation χρησιμοποιείται για να εξετάσουμε πόσο καλό είναι ένα εκτιμηθέν μοντέλο

Εκτίμηση $f(x)$ μέσω ιστογράμματος

Ιστόγραμμα

- Έστω $X_1, \ldots, X_n$ τ.δ από κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(x)$.
- Έστω ότι τα παρατηρηθέντα δεδομένα ανήκουν εντός του διαστήματος $[a, b]$
- Έστω $m \in \mathbb{Z}_+$ και η διαμέριση του $[a, b]$ σε m κελιά $B_1, \ldots, B_m$
- Το μήκος κάθε κελιού ισούται με

$$h = \frac{b-a}{m}$$

- Έστω

$$n_j = \{\#X_i : X_i \in B_j\} \tag{5}$$

$$\hat{p}_j = \frac{n_j}{n} \tag{6}$$

$$p_j = \int_{B_j} f(x)\mathrm{d}x \tag{7}$$

για $j = 1, \ldots, m$.

Ορισμός

Η συνάρτηση

$$\hat{f}_{h,n}(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} \mathrm{I}\{x \in B_j\}$$

$$\hat{f}_{h,n}(x) = \frac{\hat{p}_j}{h}, \quad x \in B_j \tag{8}$$

λέγεται εκτιμητής πυκνότητας με χρήση ιστογράμματος με πλάτος κελιών (παράμετρος εξομάλυνσης) $h > 0$.

Θεώρημα

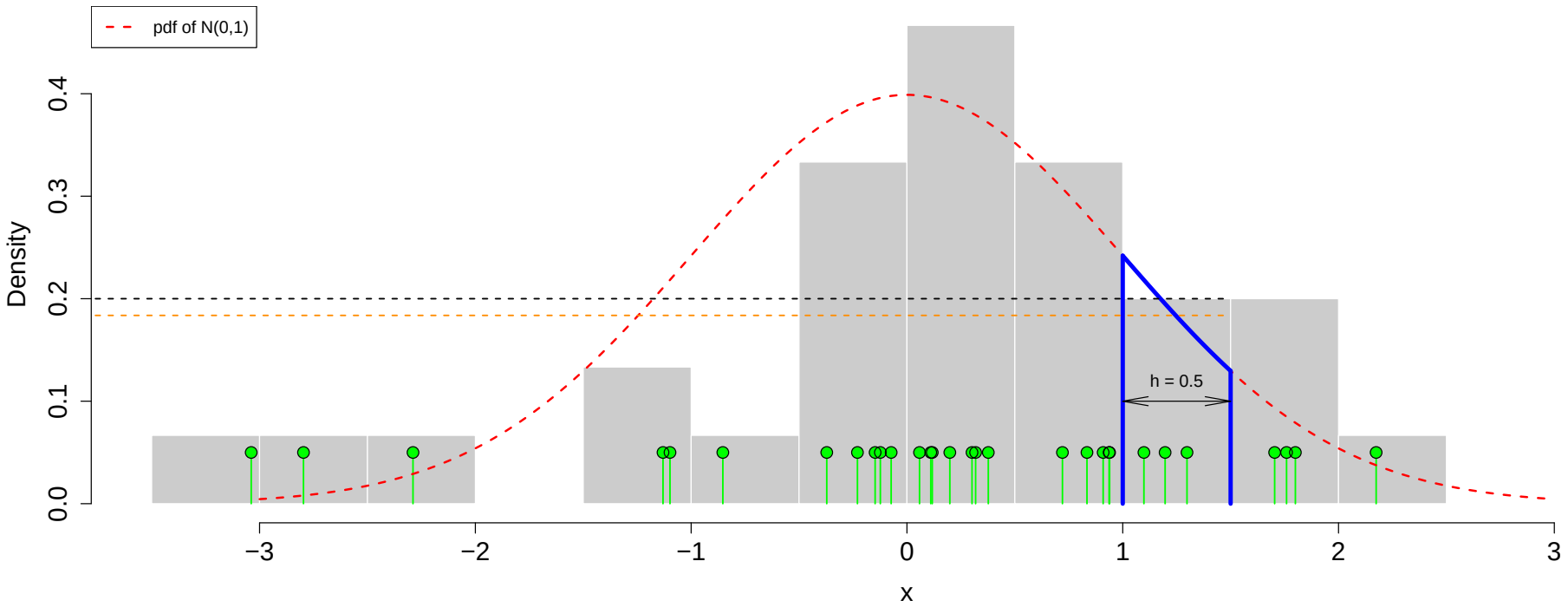
Για σταθερά h και x ισχύει ότι

$$\mathrm{E}\hat{f}_{h,n}(x) = \frac{p_j}{h}$$
$$\mathrm{Var}\hat{f}_{h,n}(x) = \frac{p_j(1-p_j)}{nh^2},$$

όπου $p_j = \int_{B_j} f(x)\mathrm{d}x$.

- Η απόδειξη προκύπτει άμεσα διότι: $n_j \sim \mathcal{B}(n, p_j)$
- Το ιστόγραμμα είναι *αμερόληπτος εκτιμητής της μέσης πυκνότητας* σε κάθε κελί
- Αυτό δεν είναι το ίδιο με έναν αμερόληπτο εκτιμητή της f (εκτός αν f σταθερή στα B_j)
- Αν f δεν είναι σταθερή στο B_j, η $\hat{f}_{h,n}$ είναι μεροληπτική.

Παράδειγμα: $n = 30$ παρατηρήσεις από $\mathcal{N}(0,1)$



- $B_1 = [-3.5, -3), \ldots, B_{12} = [2, 2.5)$ (πλάτος: $h = 0.5$)
- Για $x \in B_{10}$: $\hat{f}_{h,n}(x) = \dfrac{\hat{p}_{10}}{h} = \dfrac{n_{10}/n}{h} = \dfrac{3/30}{0.5} = 0.2$
- $p_{10} = \mathrm{P}(X \in B_{10}) = \int_1^{1.5} f(x)\mathrm{d}x \approx 0.092$
- Μέση πυκνότητα στο B_{10}: $p_{10}/h = 0.184$

Ιδιότητες ιστογράμματος

Υπό κάποιες συνθήκες ομαλότητας στην $f(x)$ μπορεί να δειχθεί ότι

$$\text{bias}\hat{f}_{h,n}(x) \approx \frac{1}{2}f'(x)[h - 2(x - b_{j-1})]$$

$$\text{Var}\hat{f}_{h,n}(x) \approx \frac{f(x)}{nh}$$

$$\text{IMSE}\hat{f}_{h,n}(x) \approx \frac{h^2}{12}\int_{-\infty}^{\infty} f'^2(x)\mathrm{d}x + \frac{1}{nh}$$

- Τα παραπάνω αποτελέσματα ισχύουν προσεγγιστικά
- Υπάρχουν κάποιες επί πλέον μικρότερες ποσότητες[2] στα δεξιά μέλη που αγνούνται για λόγους ευκολίας
- Παρατηρήστε ότι η μεροληψία της $\hat{f}_{h,n}$ δεν εξαρτάται από το n
- Καθώς αυξάνει το h μεγαλώνει η μεροληψία
- Καθώς μικραίνει το h μεγαλώνει η διασπορά

[2] δες σελίδες 74-75 σημειώσεων Ιωαννίδη

Συνέπεια ιστογράμματος

Το ιστόγραμμα είναι ασυμπτωτικά συνεπής εκτιμητής της $f(x)$

Θεώρημα

Έστω ότι $f(x)$ συνεχής στο x και $|f'(x)| < M$. Τότε για $h \to 0$ και $nh \to \infty$ καθώς $n \to \infty$ ισχύει ότι

$$\hat{f}_{h,n}(x) \xrightarrow{\mathrm{P}} f(x).$$

- Η συνθήκη $h \to 0$ εξασφαλίζει ότι η μεροληψία της $\hat{f}_{h,n}$ τείνει στο 0
- Η συνθήκη $nh \to \infty$ εξασφαλίζει ότι η διασπορά της $\hat{f}_{h,n}$ τείνει στο 0

Επιλογή του h στο ιστόγραμμα

Πρόταση

Η cross-validation εκτίμηση (4) *του* $J(h)$ (2) *για ιστόγραμμα με πλάτος κελιών* h *είναι ίση με*

$$\widehat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p}_j^2. \tag{9}$$

Απόδειξη της (9)

Αρχικά, να παρατηρήσουμε ότι για το ιστόγραμμα ισχύει ότι

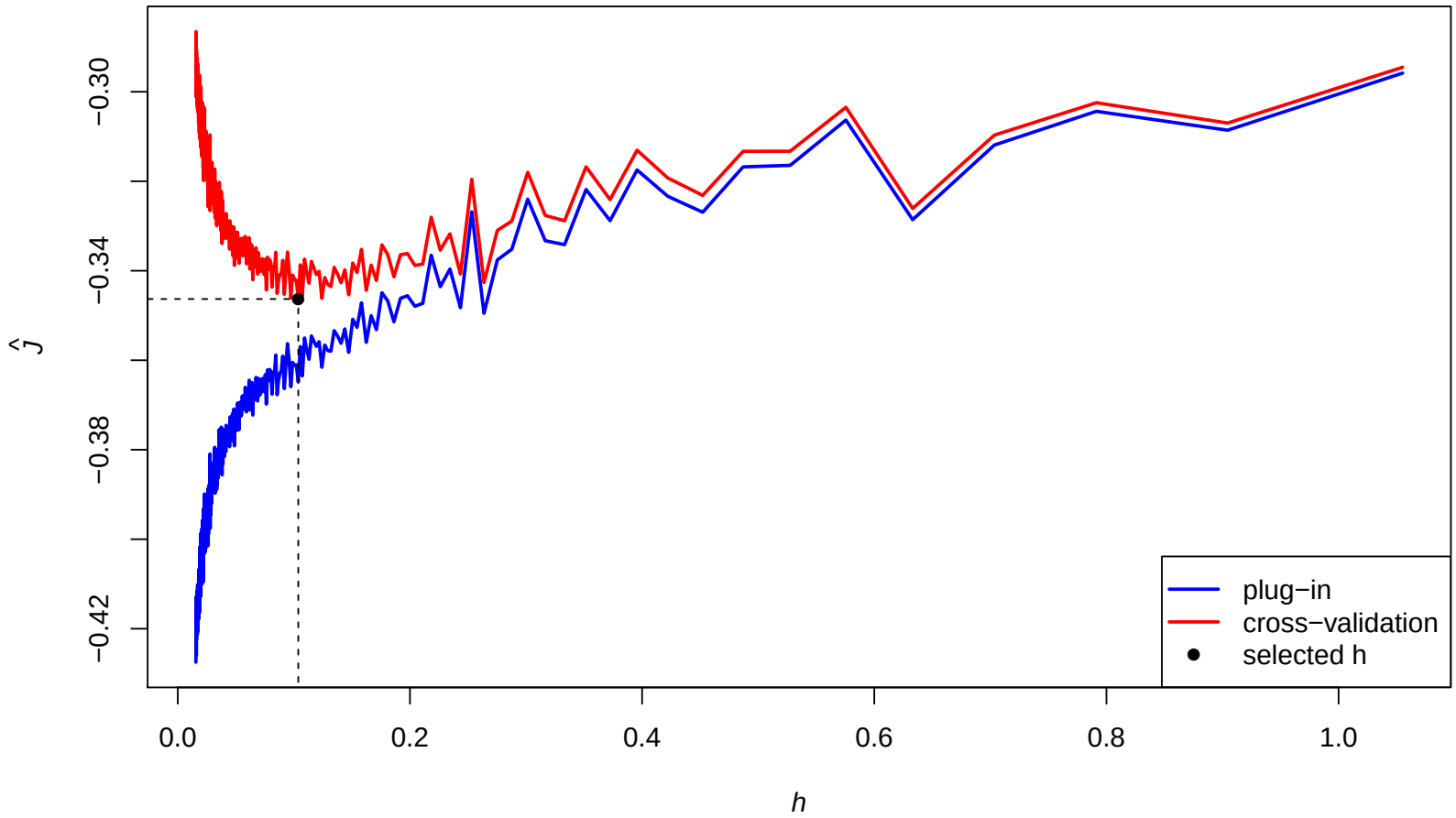
$$\begin{aligned}
\hat{f}^{(-i)}_{h,n-1}(x_i) &= \frac{n_{j_i} - 1}{h(n-1)} \\
&= \frac{n}{n-1}\frac{n_{j_i} - 1}{hn} \\
&= \frac{n}{n-1}\left(\frac{n_{j_i}}{hn} - \frac{1}{hn}\right) \\
&= \frac{n}{n-1}\left(\frac{\hat{p}_{j_i}}{h} - \frac{1}{hn}\right) \qquad (10)
\end{aligned}$$

όπου $j_i := \{j = 1, \ldots, m : x_i \in B_j\}$.

Απόδειξη της (9)

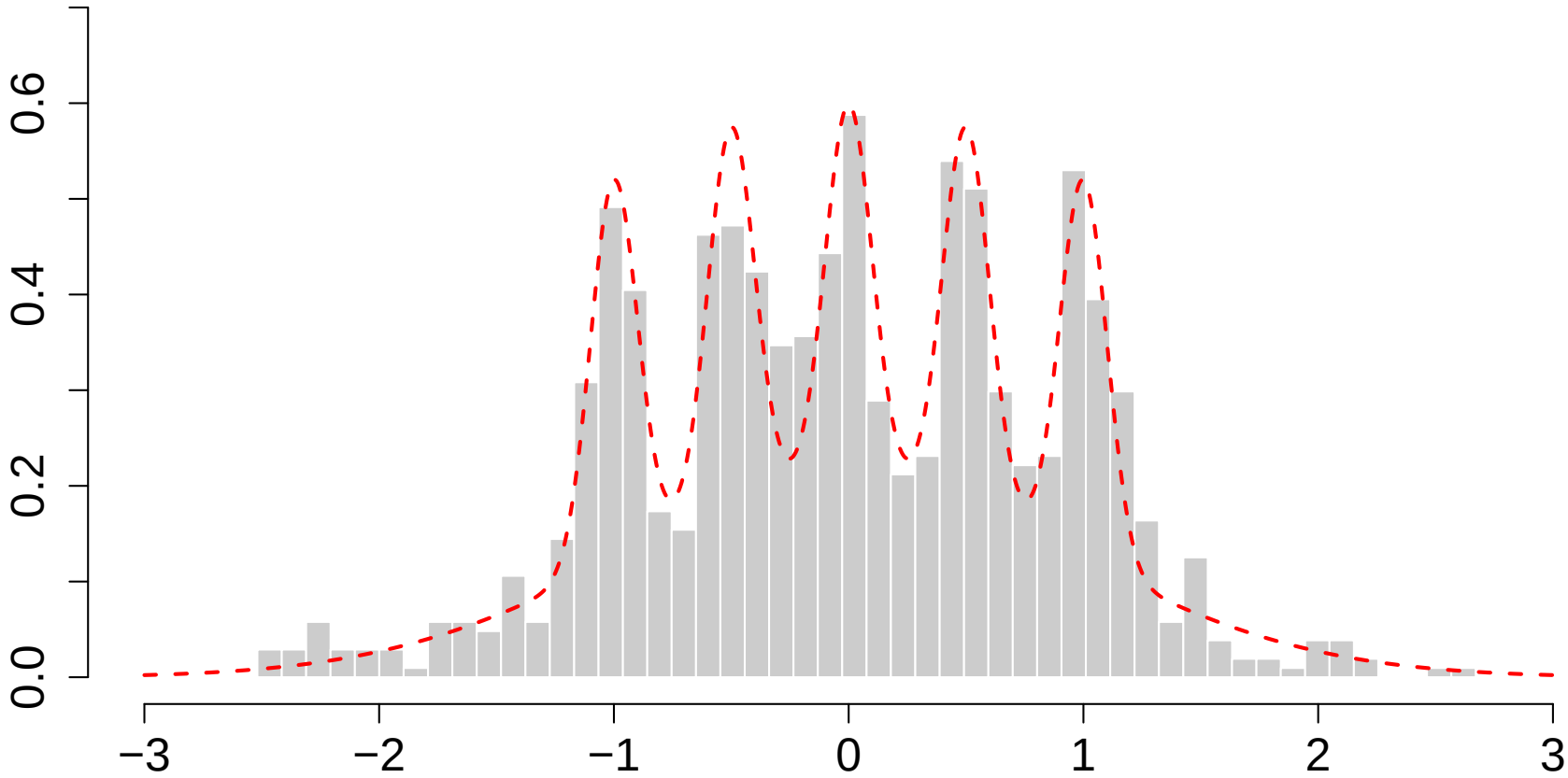
$$
\begin{aligned}
\widehat{J}(h) &= \int \hat{f}_{h,n}^2(x)\mathrm{d}x - 2\sum_{i=1}^{n}\frac{1}{n}\hat{f}_{h,n-1}^{(-i)}(x_i) \\
&= \sum_{j=1}^{m}\int_{B_j}\hat{f}_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{h,n-1}^{(-i)}(x_i) \\
&= \sum_{j=1}^{m}\int_{B_j}\frac{\hat{p}_j^2}{h^2}\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n}\frac{n}{n-1}\left(\frac{\hat{p}_{j_i}}{h} - \frac{1}{hn}\right) \\
&= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{i=1}^{n}\hat{p}_{j_i} \\
&= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{j=1}^{m}\sum_{i:x_i\in B_j}\hat{p}_j \\
&= \frac{2}{h(n-1)} + \frac{1}{h}\sum_{j=1}^{m}\hat{p}_j^2 - \frac{2}{h(n-1)}\sum_{j=1}^{m}n\hat{p}_j^2 \\
&= \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)}\sum_{j=1}^{m}\hat{p}_j^2.
\end{aligned}
$$

Παράδειγμα: Bart Simpson ($n = 1000$)



- Ελάχιστοποίηση εκτιμηθέντος μέσου σφάλματος (9) μέσω cross-validation: $\hat{J}(h) \approx -0.346$ για $h \approx 0.104$
- Παρατηρήστε ότι για τον εκτιμητή αντικατάστασης (3) το βέλτιστο $h \to 0$ (αναμενόμενο)

Παράδειγμα: Bart Simpson ($n = 1000$)



- Εκτίμηση $f(x)$ με ιστόγραμμα $\hat{f}_{h,n}(x)$ με παράμετρο εξομάλυνσης $h = 0.104$
- Πρόκειται για το βέλτιστο εύρος κελιών βάσει του ολοκληρωμένου μέσου τετραγωνικού σφάλματος

Η εντολή `hist()` στην `R`

- Οι τιμές `f$density` επιστρέφουν τις τιμές $\frac{\hat{p}_j}{h}$
- Προσοχή: το πλήθος των κελιών δεν επιλέγεται μέσω του κριτηρίου που περιγράψαμε
- Μέσω του ορίσματος `breaks = ...` μπορούμε να ορίσουμε τα κελιά

```
> x <- rnorm(20)
> h <- hist(x, freq = F)
> h$density
[1] 0.1 0.4 0.7 0.3 0.4 0.1
> h <- hist(x, breaks = c(-2,-1,0,1,2), freq = F)
> h$density
[1] 0.05 0.55 0.35 0.05
```

Εκτίμηση συνάρτηση πυκνότητας με χρήση πυρήνα

Εισαγωγή

- Έστω συνεχής τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας $f(x)$ και συνάρτηση κατανομής $F(x)$, $x \in \mathbb{R}$.
- Εξ ορισμού

$$f(x_0) = \lim_{h \to 0} \frac{\mathrm{P}(x_0 - h < X < x_0 + h)}{2h} \tag{11}$$

- Διότι

$$\begin{aligned}
\lim_{h \to 0} \frac{\mathrm{P}(x_0 - h < X < x_0 + h)}{2h} &= \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\
&= \frac{1}{2} \left(\lim_{h \to 0} \frac{F(x_0 + h)}{h} - \lim_{h \to 0} \frac{F(x_0 - h)}{h} \right) \\
&= \frac{1}{2} \left(\lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{h} - \lim_{h \to 0} \frac{F(x_0 - h) - F(x_0)}{h} \right) \\
&\overset{\text{Θέτω } \ell = -h}{=} \frac{1}{2} \left(\lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{h} + \lim_{\ell \to 0} \frac{F(x_0 + \ell) - F(x_0)}{\ell} \right) \\
&= \frac{1}{2} \left(F'(x_0) + F'(x_0) \right) = f(x_0).
\end{aligned}$$

Εισαγωγή

- Δειγματικό «ανάλογο» της (11) δοθέντος τυχαίου δείγματος μεγέθους n

$$\widehat{f}_{n,h}(x_0) = \frac{1}{n} \frac{\#\text{ παρατηρήσεων στο διάστημα } (x_0 - h, x_0 + h)}{2h} \tag{12}$$

  όπου $h > 0$ είναι παράμετρος εξομάλυνσης (smoothing parameter).
- Ο παραπάνω εκτιμητής ονομάζεται απλοϊκός (naive)
- Επειδή μοιάζει και με κουτάκι που διανύει τον οριζόντιο άξονα, ένα άλλο όνομα είναι boxcar

Απλοϊκός εκτιμητής (boxcar, naive)

- Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα από κάποια κατανομή με συνάρτηση πυκνότητας πιθανότητας $f(x)$.

Ορισμός

Η συνάρτηση

$$\widehat{f}_{h,n}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}\{|X_i - x| \leqslant h\}}{2h}, \quad x \in \mathbb{R} \tag{13}$$

ονομάζεται απλοϊκός εκτιμητής με παράμετρο εξομάλυνσης $h > 0$ της συνάρτησης πυκνότητας $f(x)$.

Εναλλακτική έκφραση

- Ο απλοϊκός εκτιμητής (13) γράφεται ισοδύναμα

$$\widehat{f}_{h,n}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R} \tag{14}$$

όπου

$$K(u) := \frac{1}{2}\mathrm{I}\{|u| \leqslant 1\}$$

- Παρατηρήστε ότι η $K(u)$ είναι η συνάρτηση πυκνότητας της $\mathcal{U}(-1, 1)$
- Η $K(u)$ είναι ένα παράδειγμα αυτού που θα ορίσουμε στη συνέχεια ως πυρήνα (kernel).
- Η $K(u)$ ονομάζεται **απλοϊκός πυρήνας** (boxcar kernel).
- Η (14) λέγεται **εκτιμητής πυκνότητας απλοϊκού πυρήνα με παράμετρο εξομάλυνσης** h (boxcar kernel density estimator with bandwidth h).
- Θα δούμε ότι υπάρχουν και άλλες επιλογές για την μορφή του πυρήνα.

Παράδειγμα 1

- Έστω τυχαίο δείγμα

$$\boldsymbol{x} = (-0.07, -1.36, -0.75, -0.12, 0.56, -0.94, 0.08, 1.09, 2.16, -0.82)$$

- Να εκτιμηθεί η $f(-1)$ με τον απλοϊκό εκτιμητή και παράμετρο εξομάλυνσης $h = 0.5$ και $h = 1$
- Έχουμε $n = 10$ και $x = 1$

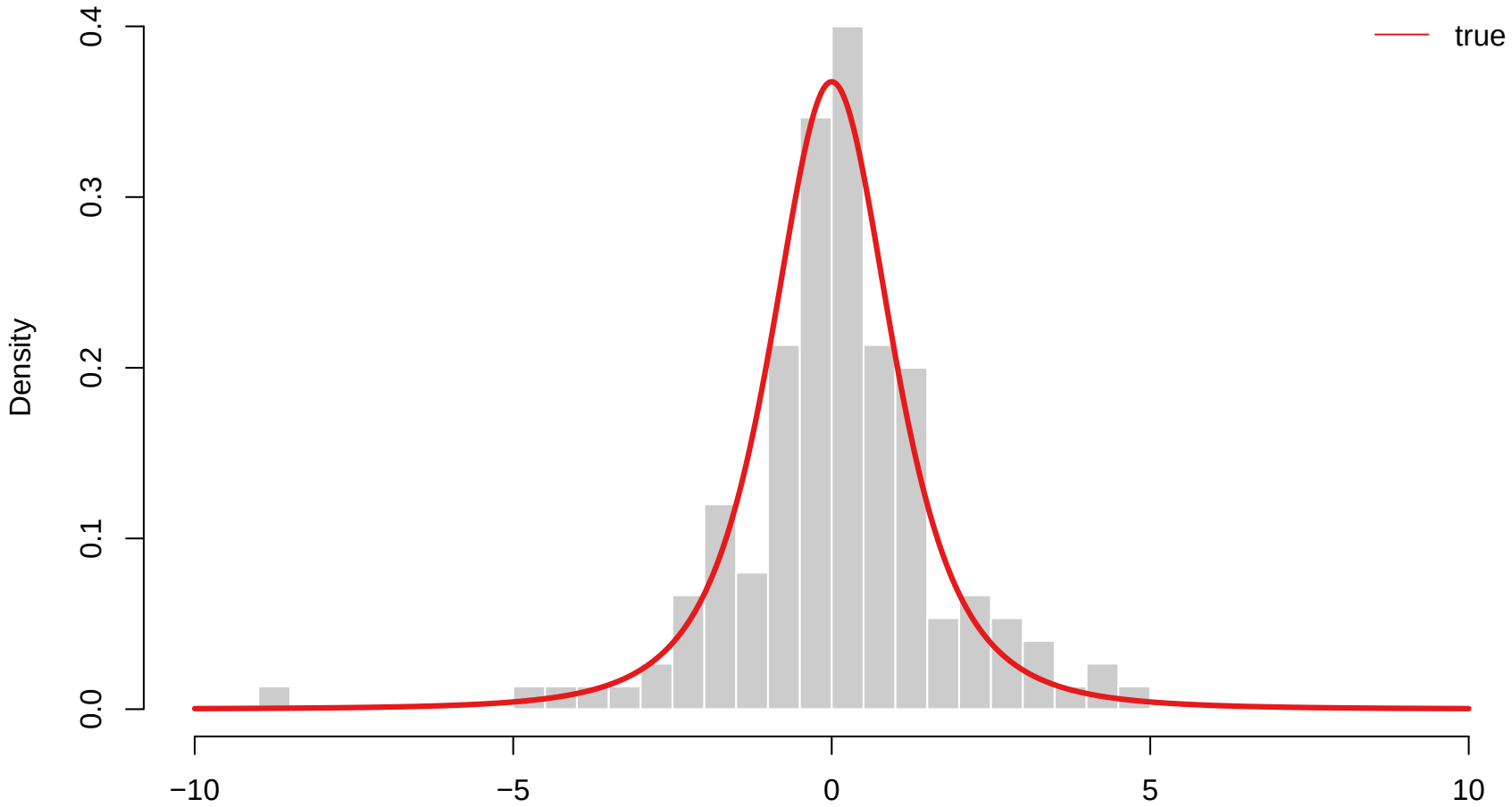
$$\widehat{f}_{10,h}(1) = \frac{1}{10} \sum_{i=1}^{10} \frac{\mathrm{I}\{|x_i - 1| \leqslant h\}}{2h}$$

- Για $h = 0.5$: $\sum_{i=1}^{10} \mathrm{I}\{|x_i - 1| \leqslant 0.5\} = 2 \Rightarrow \widehat{f}_{10,0.5}(1) = \frac{1}{10} \frac{2}{2 \times 0.5} = 0.2$
- Για $h = 1$: $\sum_{i=1}^{10} \mathrm{I}\{|x_i - 1| \leqslant 1\} = 3 \Rightarrow \widehat{f}_{10,1}(1) = \frac{1}{10} \frac{3}{2 \times 1} = 0.15$

Παράδειγμα απλοϊκών εκτιμητών: 2

$n = 150$ παρατηρήσεις από t_3

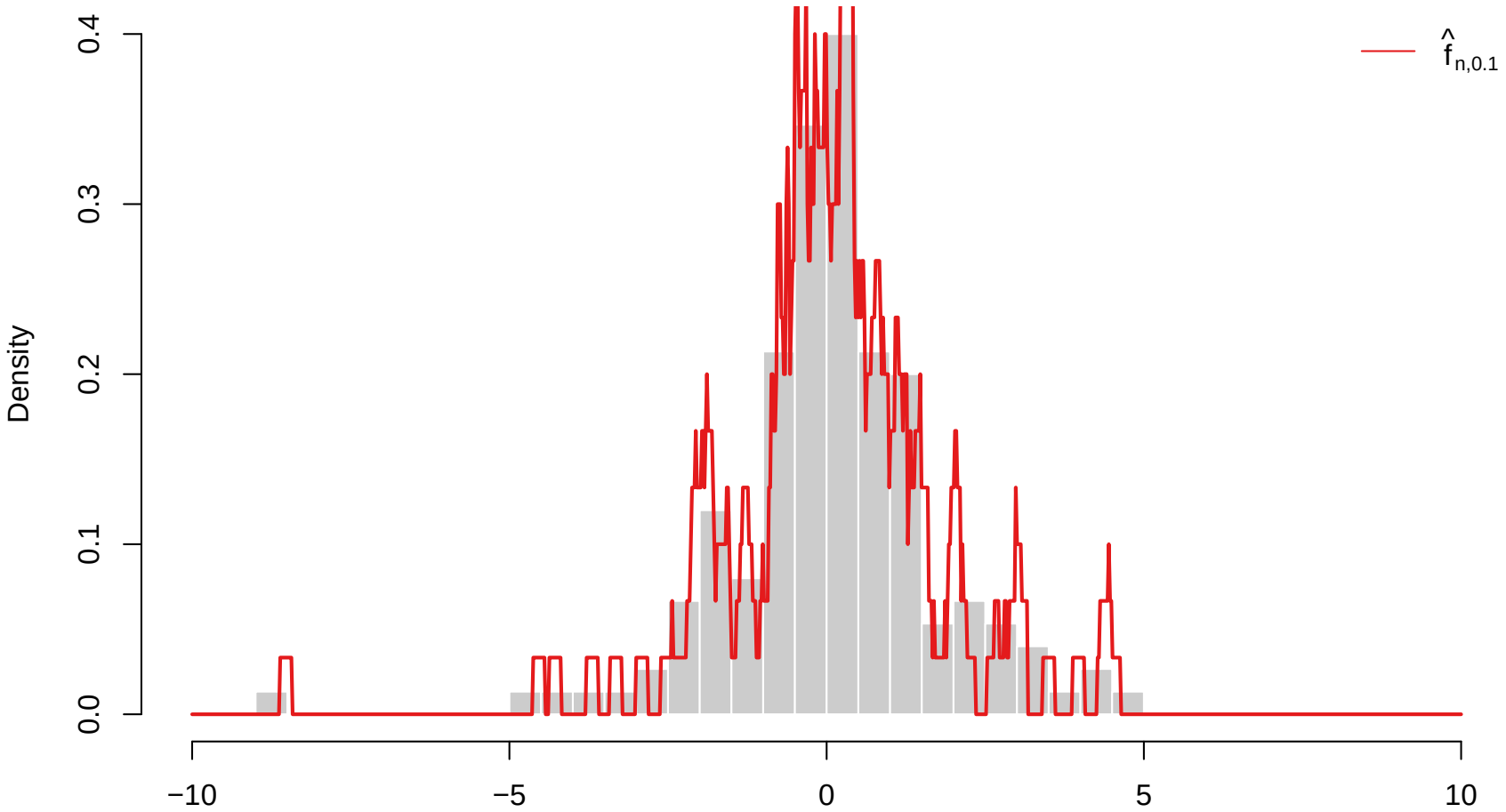
πραγματική συνάρτηση πυκνότητας



Παράδειγμα απλοϊκών εκτιμητών: 2

$n = 150$ παρατηρήσεις από t_3

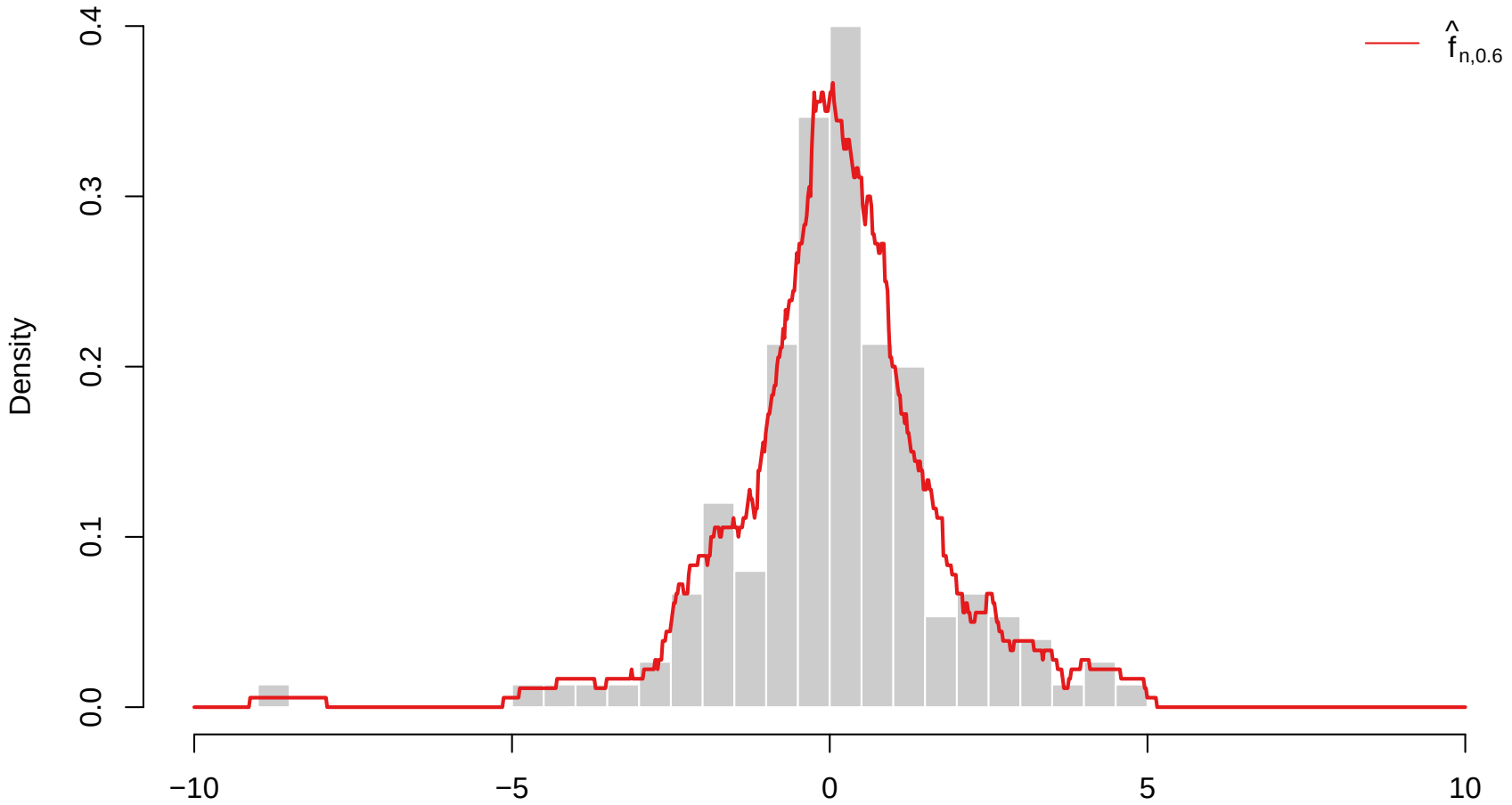
boxcar με παράμετρο εξομάλυνσης $h = 0.1$



Παράδειγμα απλοϊκών εκτιμητών: 2

$n = 150$ παρατηρήσεις από t_3

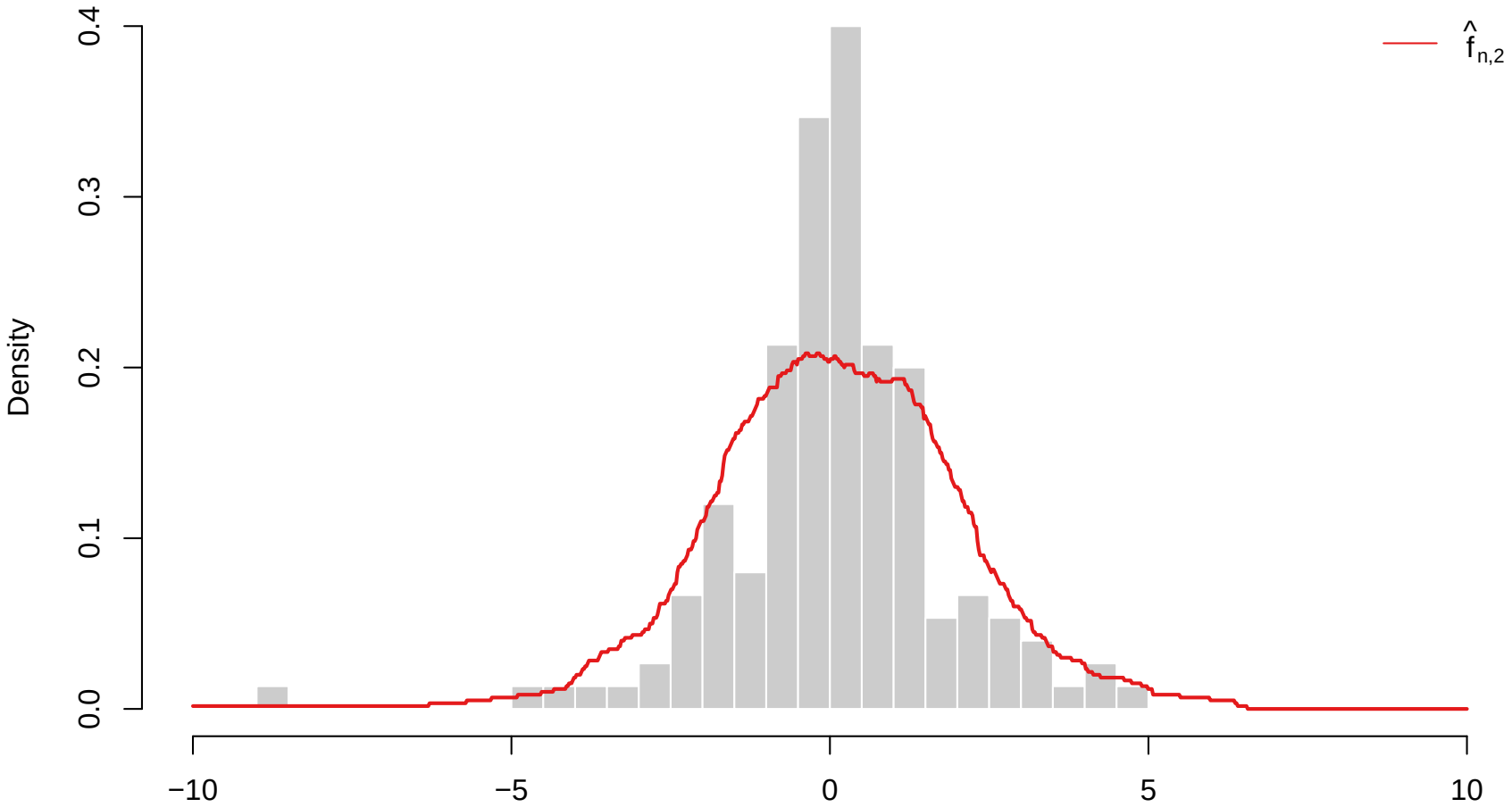
boxcar με παράμετρο εξομάλυνσης $h = 0.6$



Παράδειγμα απλοϊκών εκτιμητών: 2

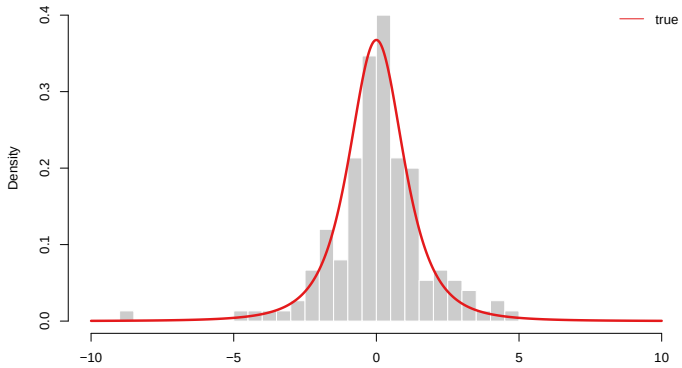
$n = 150$ παρατηρήσεις από t_3

boxcar με παράμετρο εξομάλυνσης $h = 2$

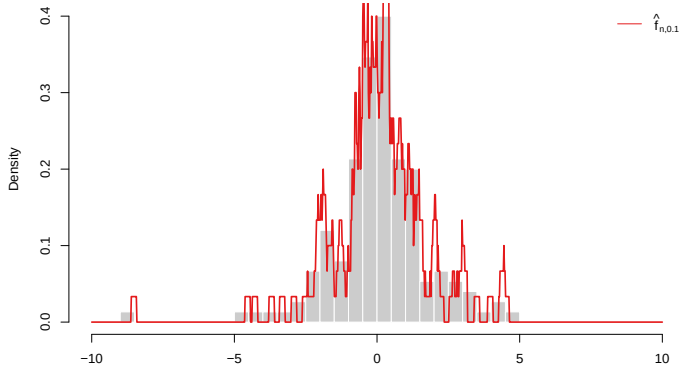


Παράδειγμα απλοϊκών εκτιμητών: 2

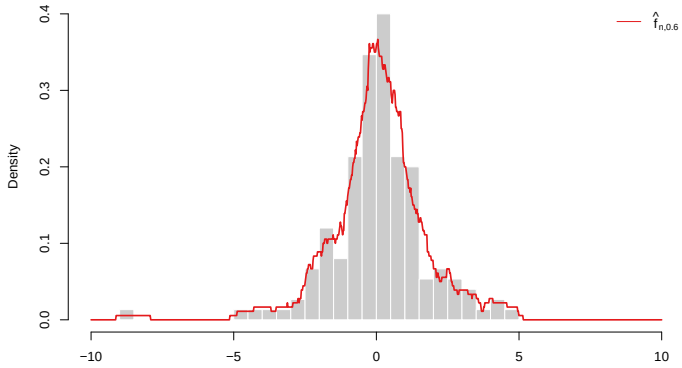
$n = 150$ παρατηρήσεις από t_3



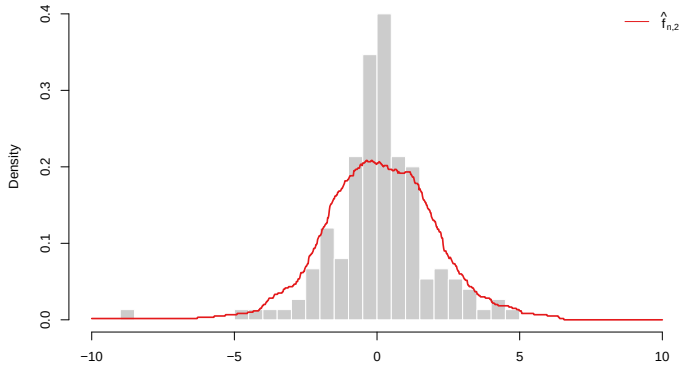
t_3 (πραγματική f)



$h = 0.1$ (undersmoothed)



$h = 0.6$ (όχι και άσχημα)



$h = 2$ (oversmoothed)

Bias-variance trade-off

- «Μικρό» $h \Rightarrow$
  μικρή μεροληψία αλλά μεγάλη διασπορά (**undersmoothing**)
- «Μεγάλο» $h \Rightarrow$:
  μικρή διασπορά αλλά μεγάλη μεροληψία (**oversmoothing**)
- Πρόβλημα: επιλογή παραμέτρου εξομάλυνσης h
- Στην πράξη θέλουμε να επιλέξουμε μία ενδιάμεση τιμή (ούτε πολύ «μικρή» ούτε πολύ «μεγάλη») ώστε να ισοσταθμίζεται η διασπορά και η μεροληψία του εκτιμητή.

Ορισμός (Πυρήνας)

Μία πραγματική συνάρτηση $K(u)$ με τις ιδιότητες

1. $K(u) \geqslant 0$, *για κάθε* $u \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$
3. $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$
4. $\sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u)\mathrm{d}u > 0$

ονομάζεται ***πυρήνας*** *(kernel).*

- Παρατήρηση: ένας πυρήνας είναι μία συνάρτηση πυκνότητας πιθανότητας (ιδιότητες $1 + 2$) η οποία έχει
  - μέση τιμή 0 (ιδιότητα 3)
  - (πεπερασμένη) διασπορά σ_K^2 (ιδιότητα 4).

Παραδείγματα πυρήνων

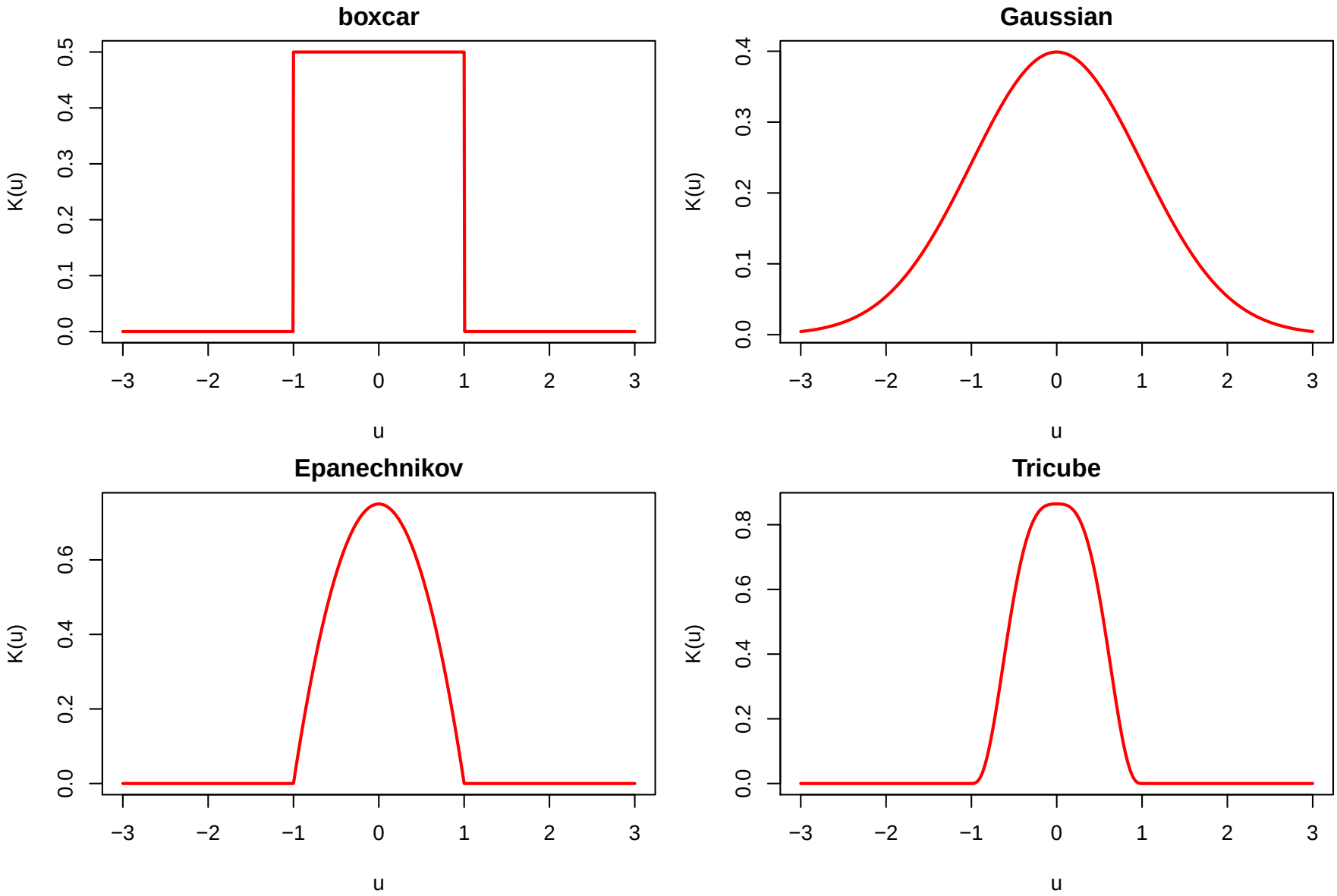
boxcar kernel: $K(u) = \frac{1}{2}\mathrm{I}\{|u| \leqslant 1\}$

Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\mathrm{I}\{x \in \mathbb{R}\}$

Epanechnikov kernel: $K(u) = \frac{3}{4}\left(1 - u^2\right)\mathrm{I}\{|u| \leqslant 1\}$

tricube kernel: $K(u) = \frac{70}{81}\left(1 - |x|^3\right)^3\mathrm{I}\{|u| \leqslant 1\}$

Παραδείγματα πυρήνων



Εκτίμηση πυκνότητας με χρήση πυρήνα

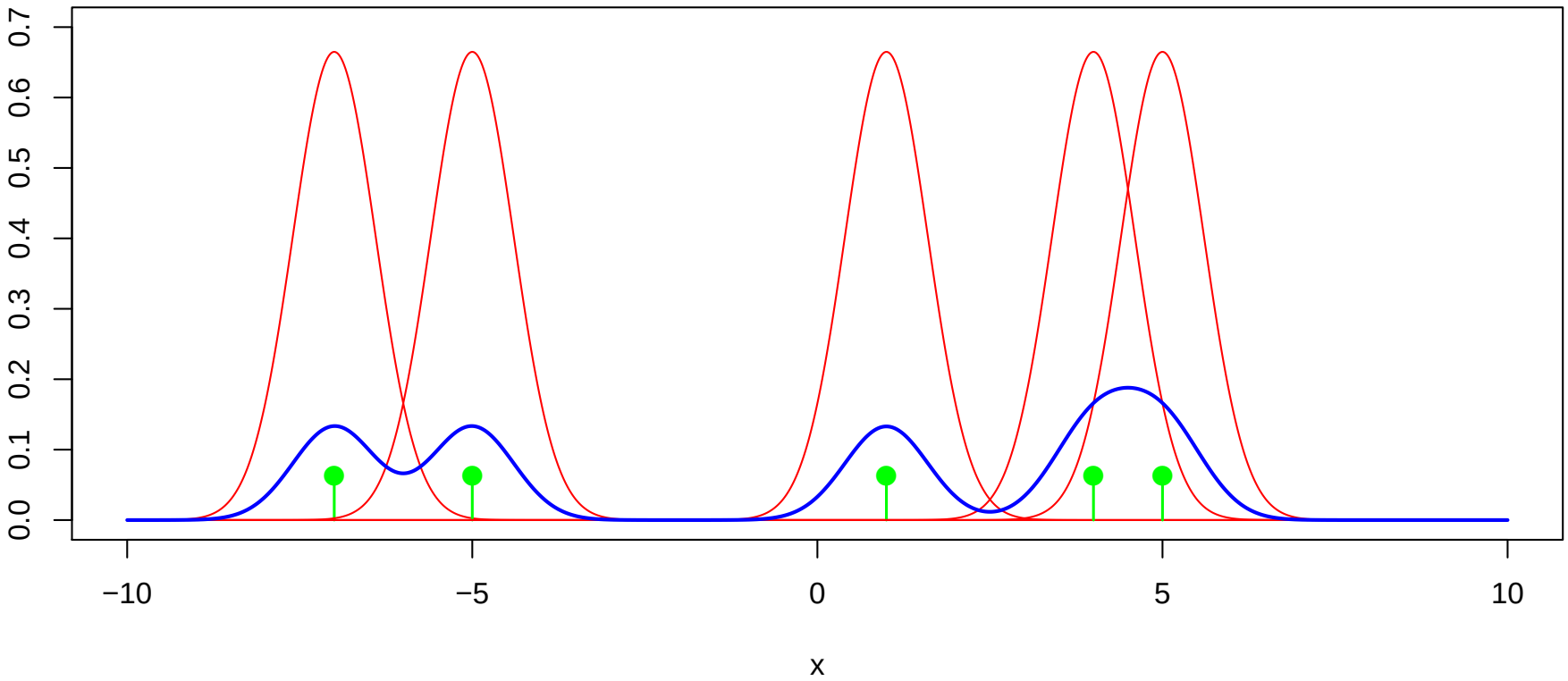
Ορισμός

Η εκτίμηση μιας συνάρτησης πυκνότητας πιθανότητας με τη χρήση ενός πυρήνα $K(u)$ δίνεται ως

$$\widehat{f}_{h,n}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}. \tag{15}$$

Η παράμετρος $h > 0$ λέγεται παράμετρος εξομάλυνσης (bandwidth).

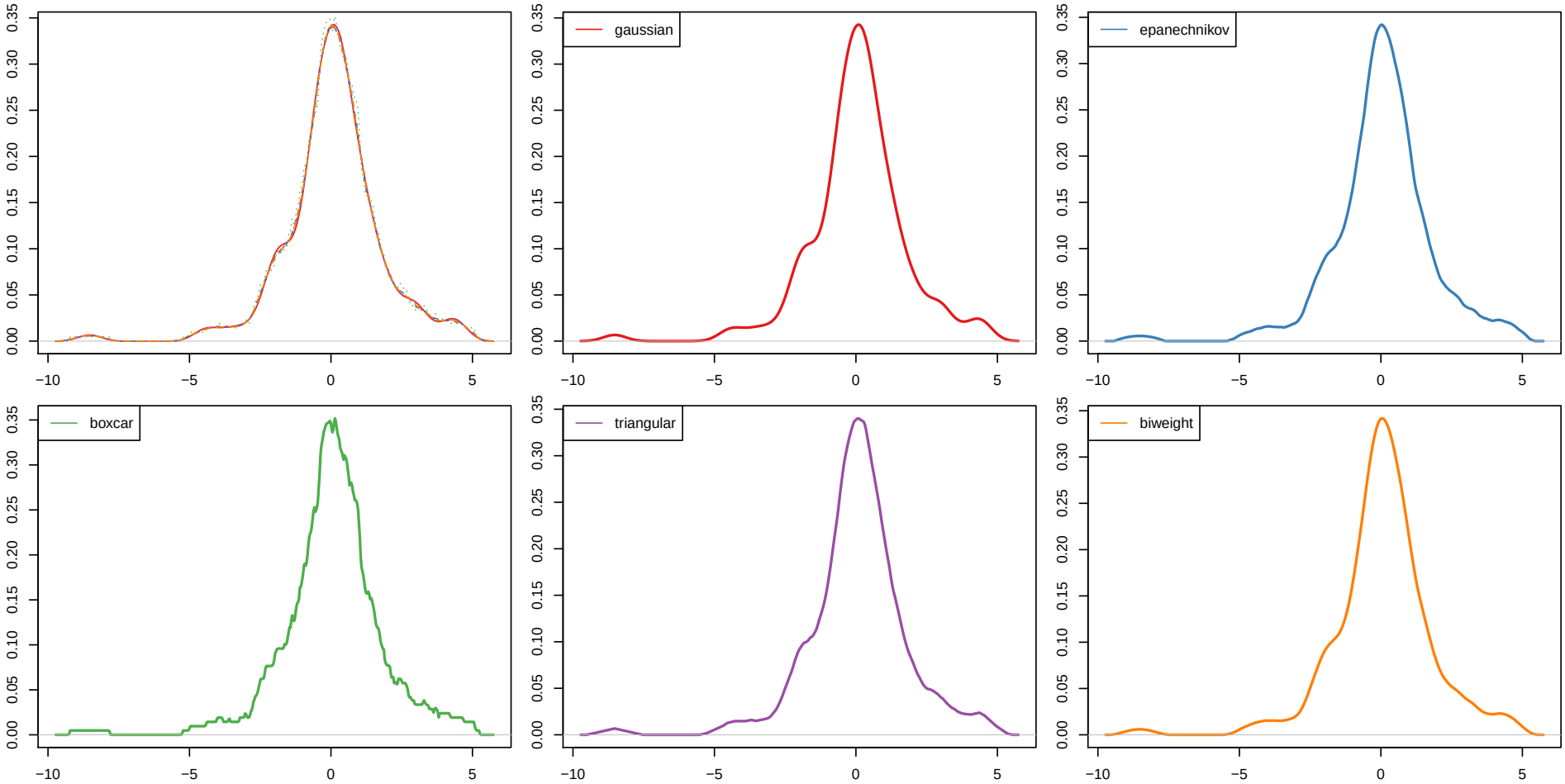
Παράδειγμα



- Παρατηρηθέντα δεδομένα ($n = 5$).
- Ένας εκτιμητής πυκνότητας $\widehat{f}_{h,n}(x)$ με χρήση Gaussian πυρήνα.
- Για κάθε σημείο x, η $\widehat{f}_{h,n}(x)$ είναι ο μέσος όρος των πυρήνων με κεντρικό σημείο τα παρατηρηθέντα X_i.

Η επιλογή του πυρήνα δεν παίζει μεγάλο ρόλο

- Θεωρητικά ο πυρήνας Epanechnikov είναι ο βέλτιστος, αλλά
- Η διαφορά μεταξύ διαφορετικών επιλογών είναι αμελητέα
- Αυτό που είναι σημαντικό είναι η επιλογή του h



Ιδιότητες εκτιμητή με χρήση πυρήνα

Από τον ορισμό (15) προκύπτει ότι ο εκτιμητής με χρήση πυρήνα είναι όντως συνάρτηση πυκνότητας πιθανότητας

- Προφανώς $\widehat{f}_{h,n}(x) \geqslant 0$ για κάθε $x \in \mathbb{R}$
- Ισχυεί ότι

$$\begin{aligned}
\int_{-\infty}^{\infty} \widehat{f}_{h,n}(x)\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\mathrm{d}x \\
&= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-X_i}{h}\right)\mathrm{d}x \quad \left(u = \frac{x-X_i}{h}\right) \\
&= \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} K(u)\mathrm{d}u \quad \text{(ιδιότητα 2)} \\
&= \frac{n}{n} \\
&= 1.
\end{aligned}$$

Μεροληψία και διασπορά

Πρόταση

Έστω ότι η f είναι τρεις φορές παραγωγίσιμη, με $f^{(3)}(x) < M < \infty$ και ότι K είναι συμμετρικός πυρήνας με διασπορά σ_K^2. Τότε

$$|\text{bias}\hat{f}_{h,n}(x)| \approx \frac{1}{2}h^2 f''(x)\sigma_K^2$$

$$\text{Var}\hat{f}_{h,n}(x) \approx \frac{f(x)\int K^2(u)\text{d}u}{nh}$$

- Τα παραπάνω αποτελέσματα ισχύουν προσεγγιστικά
- Υπάρχουν κάποιες επί πλέον μικρότερες ποσότητες[3] στα δεξιά μέλη που αγνοούνται για λόγους ευκολίας
- Παρατηρήστε ότι η μεροληψία της $\hat{f}_{h,n}$ δεν εξαρτάται από το n
- Καθώς αυξάνει το h μεγαλώνει η μεροληψία
- Καθώς μικραίνει το h μεγαλώνει η διασπορά

[3] δες σελίδες 82-83 σημειώσεων Ιωαννίδη

Συνέπεια εκτιμητή με χρήση πυρήνα

Ο εκτιμητής με χρήση πυρήνα είναι ασυμπτωτικά συνεπής εκτιμητής της $f(x)$

Θεώρημα

Έστω ότι $f(x)$ συνεχής στο x και $|f'(x)| < M$. Τότε για $h \to 0$ και $nh \to \infty$ καθώς $n \to \infty$ ισχύει ότι

$$\hat{f}_{h,n}(x) \xrightarrow{\mathrm{P}} f(x).$$

Επιλογή του bandwidth (h)

- Χρησιμοποιούμε το ολοκληρωμένο μέσο τετραγωνικό σφάλμα (1) για την επιλογή του h
- Παραλείποντας ό,τι δεν εξαρτάται από h ελαχιστοποιούμε την (2)

$$J(h) := \mathrm{E}\left[\int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - 2\int_{-\infty}^{\infty} \hat{f}_{h,n}(x)f(x)\mathrm{d}x\right]$$

- Αμερόληπτος εκτιμητής της $J(h)$ μέσω cross-validation (4)

$$\begin{aligned}\hat{J}(h) &= \int_{-\infty}^{\infty} \hat{f}_{h,n}^2(x)\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,n-1}^{(-i)}(x_i) \\ &\approx \frac{1}{hn^2}\sum_i\sum_j K^\star\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0)\end{aligned}$$

  όπου $K^\star(x) = K^{(2)}(x) - 2K(x)$ και $K^{(2)}(z) = \int K(z-y)K(y)\mathrm{d}y$.
- Όταν K είναι ο κανονικός πυρήνας $\mathcal{N}(0,1)$, τότε η $K^{(2)}(z)$ είναι η συνάρτηση πυκνότητας της $\mathcal{N}(0,2)$.

Η συνάρτηση `density()` στην `R`

Μέσω `R` χρησιμοποιούμε την εντολή

```
density(x, bw = ..., kernel = ...)
```

όπου

- `x`: τα δεδομένα
- `bw`: καθορίζει το bandwidth (h) και μπορεί να είναι
  - είτε κάποια θετική τιμή (προκαθορισμένο h)
  - είτε κάποιος χαρακτήρας (`"nrd"`, `"ucv"`, `"bcv"`, ...) που καθορίζει τον τρόπο επιλογής του h βάσει διαθέσιμων κριτηρίων
- `kernel`: ο τύπος του πυρήνα με δυνατές επιλογές `"gaussian"`, `"epanechnikov"`, `"rectangular"`, ...

Επιλογή του h στην R μέσω IMSE

- Η ελαχιστοποίηση του IMSE μέσω της $\hat{J}(h)$ είναι απαιτητική
- Υπάρχουν τεχνικές για γρήγορο υπολογισμό της $\hat{J}(h)$ (fast Fourier transform)
- Μέσω R χρησιμοποιούμε την εντολή
  `density(x, bw = "ucv", kernel = ...)`
- όπου
  - `x`: τα δεδομένα
  - `kernel`: ο τύπος του πυρήνα με δυνατές επιλογές `"gaussian"`, `"epanechnikov"`, `"rectangular"`, ...

Παράδειγμα: Bart Simpson

- Εκτίμηση πυκνότητας με χρήση κανονικού πυρήνα και bandwidth $h = 1$

```
> f <- density(x, bw=0.1,kernel='gaussian')
> print(f)

Call:
density.default(x = x, bw = 0.1, kernel = "gaussian")

Data: x (1000 obs.); Bandwidth 'bw' = 0.1

       x                   y
 Min.   :-2.8184   Min.   :0.0000001
 1st Qu.:-1.0857   1st Qu.:0.0047480
 Median : 0.6471   Median :0.0384578
 Mean   : 0.6471   Mean   :0.1441377
 3rd Qu.: 2.3798   3rd Qu.:0.3106494
 Max.   : 4.1126   Max.   :0.4575043
```

Παράδειγμα: Bart Simpson

- Εκτίμηση πυκνότητας με χρήση κανονικού πυρήνα και bandwidth επιλεγμένου βάσει του IMSE

```
> f<- density(x, bw='ucv',kernel='gaussian')
> print(f)

Call:
density.default(x = x, bw = "ucv", kernel = "gaussian")

Data: x (1000 obs.); Bandwidth 'bw' = 0.05891

       x                    y
 Min.   :-2.6952   Min.   :0.000000
 1st Qu.:-1.0240   1st Qu.:0.006677
 Median : 0.6471   Median :0.049624
 Mean   : 0.6471   Mean   :0.149454
 3rd Qu.: 2.3182   3rd Qu.:0.300083
 Max.   : 3.9893   Max.   :0.514404
```

- $h \approx 0.059$

Άλλες επιλογές της `density()`

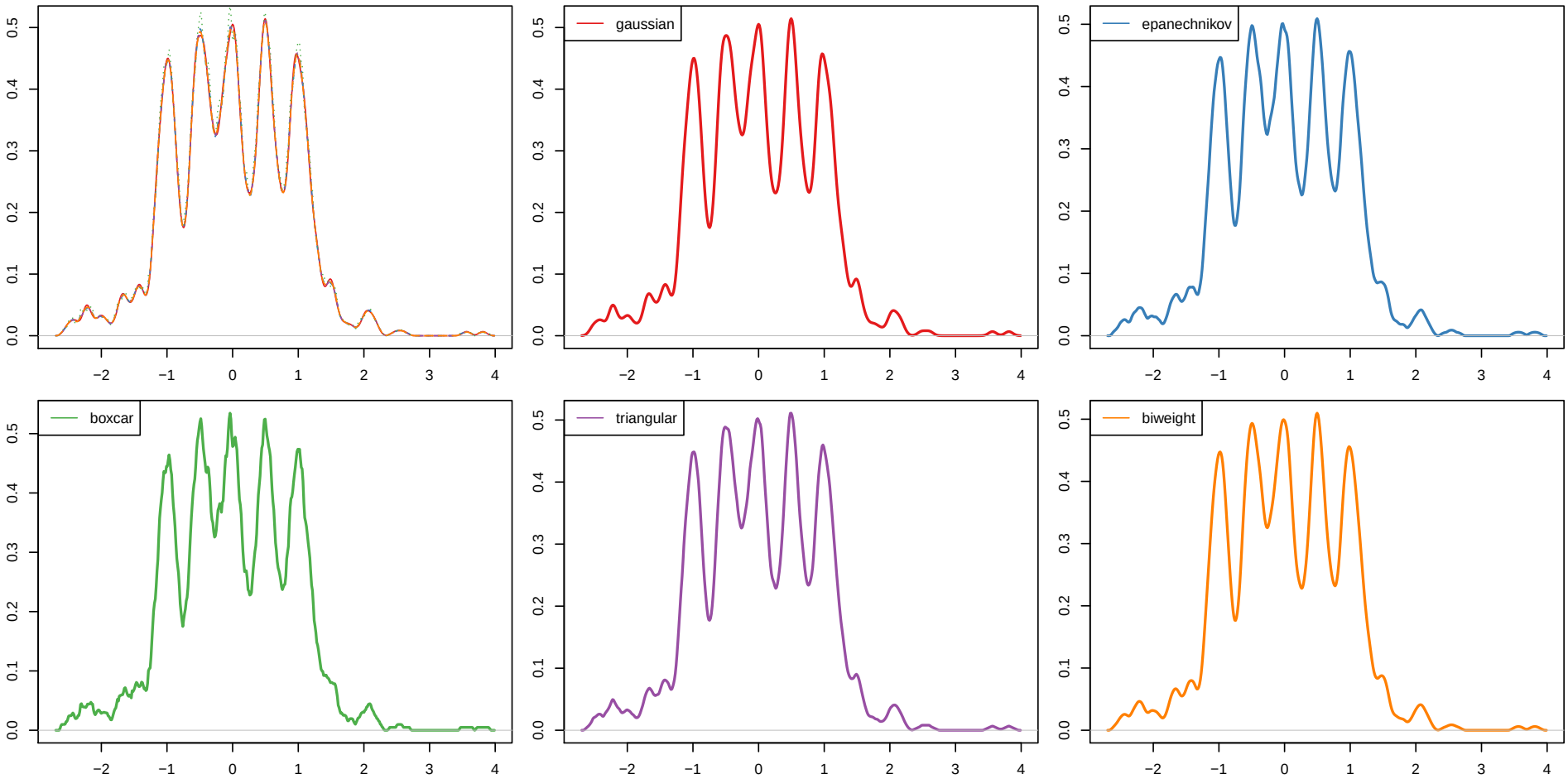
- Εξ ορισμού, η εκτιμηθείσα πυκνότητα θα υπολογιστεί σε $M = 512$ σημεία τα οποία ισαπέχουν και καλύπτουν το εύρος του παρατηρηθέντος δείγματος
- Οι τιμές αυτές επιστρέφονται μέσω των εντολών `f$x`
- Οι αντίστοιχες εκτιμήσεις της πυκνότητας επιστρέφονται μέσω της εντολής `f$y`
- Το πλήθος αυτών των σημείων μπορεί να αλλάξει μέσω του ορίσματος `n`, πχ

```
f <- density(x, bw='ucv',kernel='gaussian', n = 1024)
```

- Η συνάρτηση έχει μέθοδο `plot()` για την γραφική αναπαράστασή της:

```
plot(f)
```

IMSE: Bart Simpson data ($n = 1000$)



Επιλογή του h μέσω cross-validated πιθανοφάνειας

- Εναλλακτική τεχνική για επιλογή του h
- Κάθε τιμή του h αντιστοιχεί σε ένα μοντέλο που περιγράφει τα δεδομένα
- Για δοθέν h, η εκτιμηθείσα πιθανοφάνεια είναι $L(h) = \prod_{i=1}^{n} \hat{f}_{h,n}(x_i)$
- Μεγιστοποιείται για $h = 0$ άρα δεν δουλεύει για επιλογή του h
- Έστω $\hat{f}^{(-i)}_{h,n-1}(x)$ ο εκτιμητής της πυκνότητας στο x αν αφαιρέσουμε την i-οστή παρατήρηση, $i = 1, \ldots, n$

Επιλογή του h μέσω cross-validated πιθανοφάνειας

Ορισμός

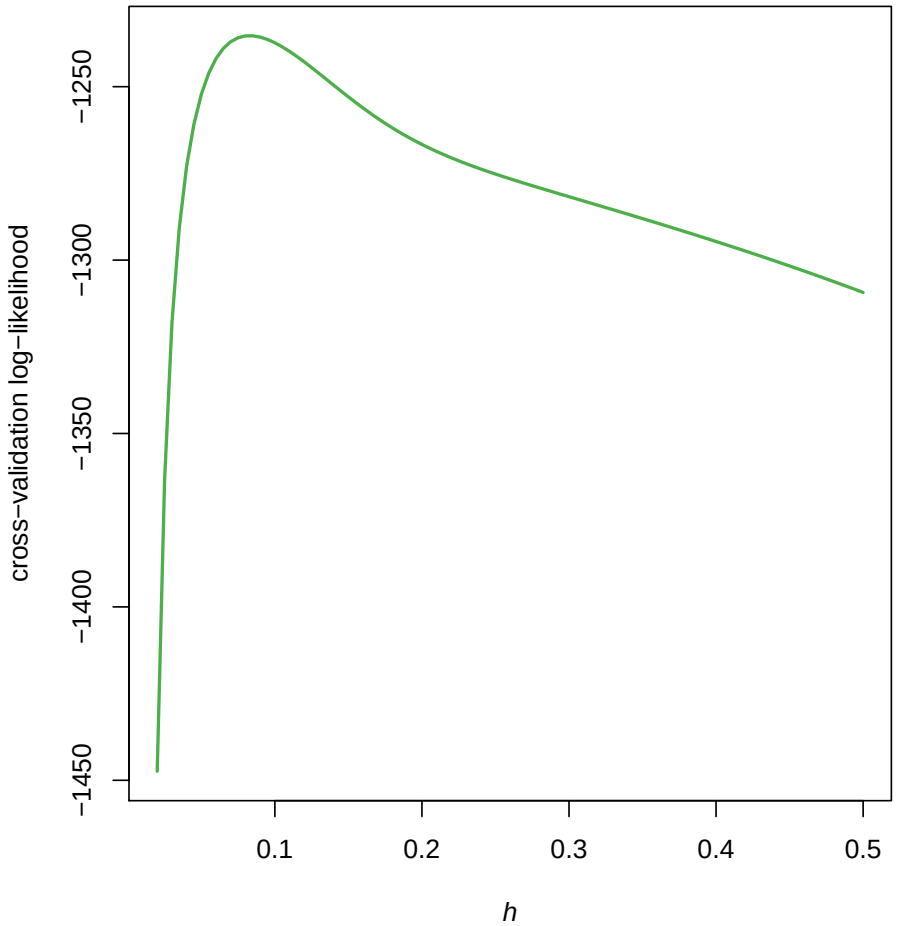
Η cross-validated πιθανοφάνεια ορίζεται ως

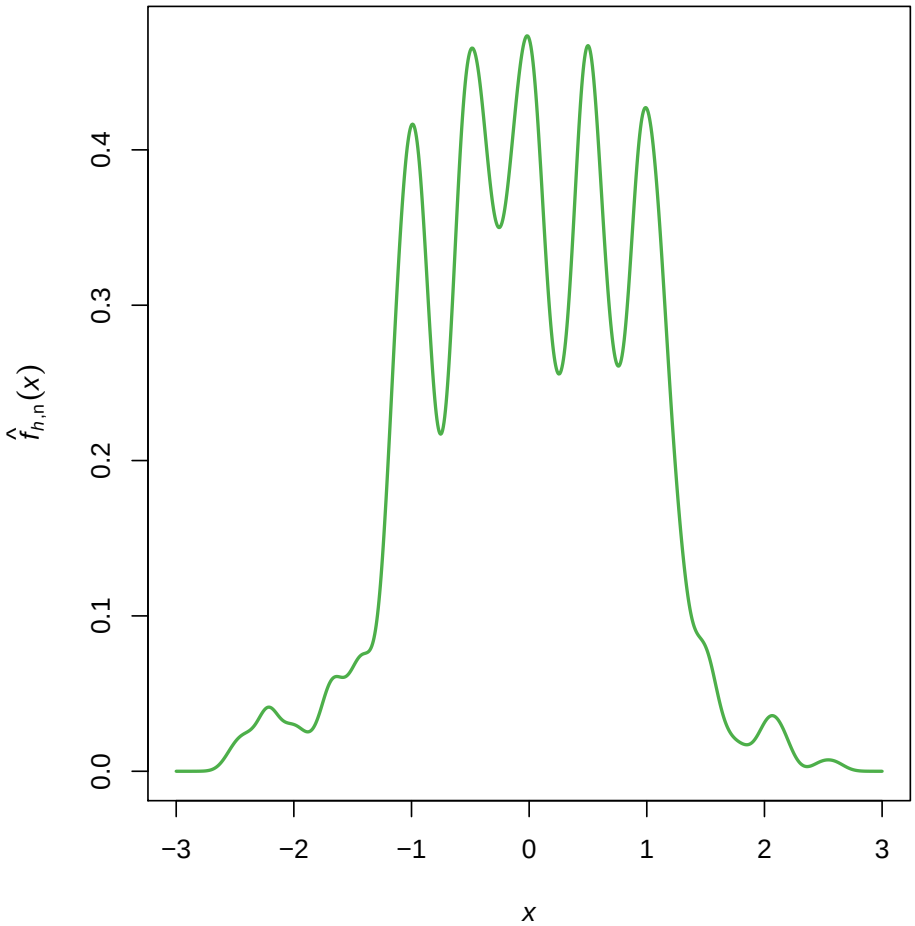
$$L^{\text{cv}}(h) := \prod_{i=1}^{n} \hat{f}^{(-i)}_{h,n-1}(x). \tag{16}$$

- Η τιμή του h που μεγιστοποιεί την (16) είναι η επιλογή μας για το bandwidth βάσει της cross-validated πιθανοφάνειας (cross-validated likelihood – CVL)[4].
- Στην πράξη η μεγιστοποίηση γίνεται δίνοντας ένα σύνολο εύλογων τιμών για το h και υπολογίζοντας τον λογάριθμο της (16) για κάθε μία τιμή.
- Το βέλτιστο h αντιστοιχεί στο μέγιστο μεταξύ όλων των τιμών.

[4] Habbema et al (1974) in *Compstat 1974: Proceedings in Computational Statistics*

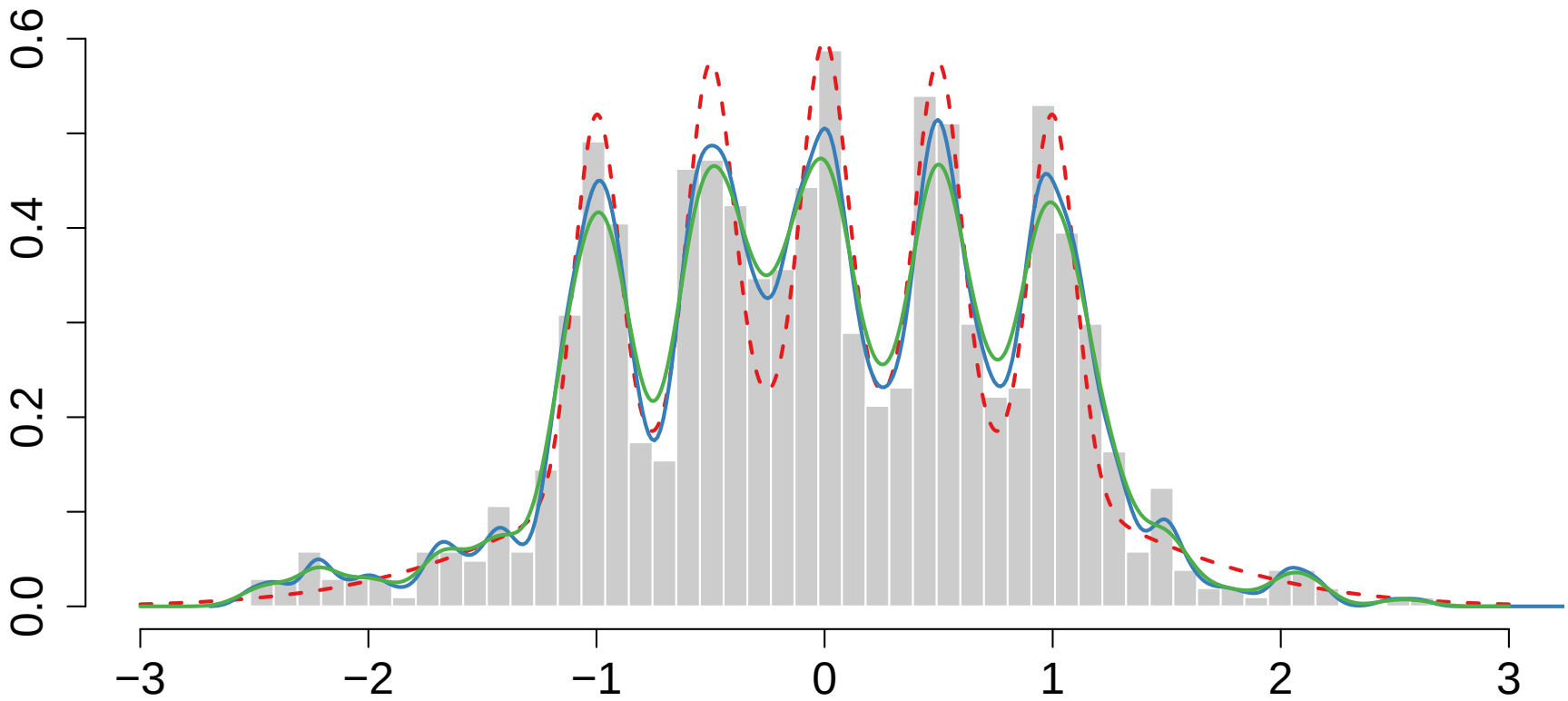
CVL: Bart Simpson data ($n = 1000$)





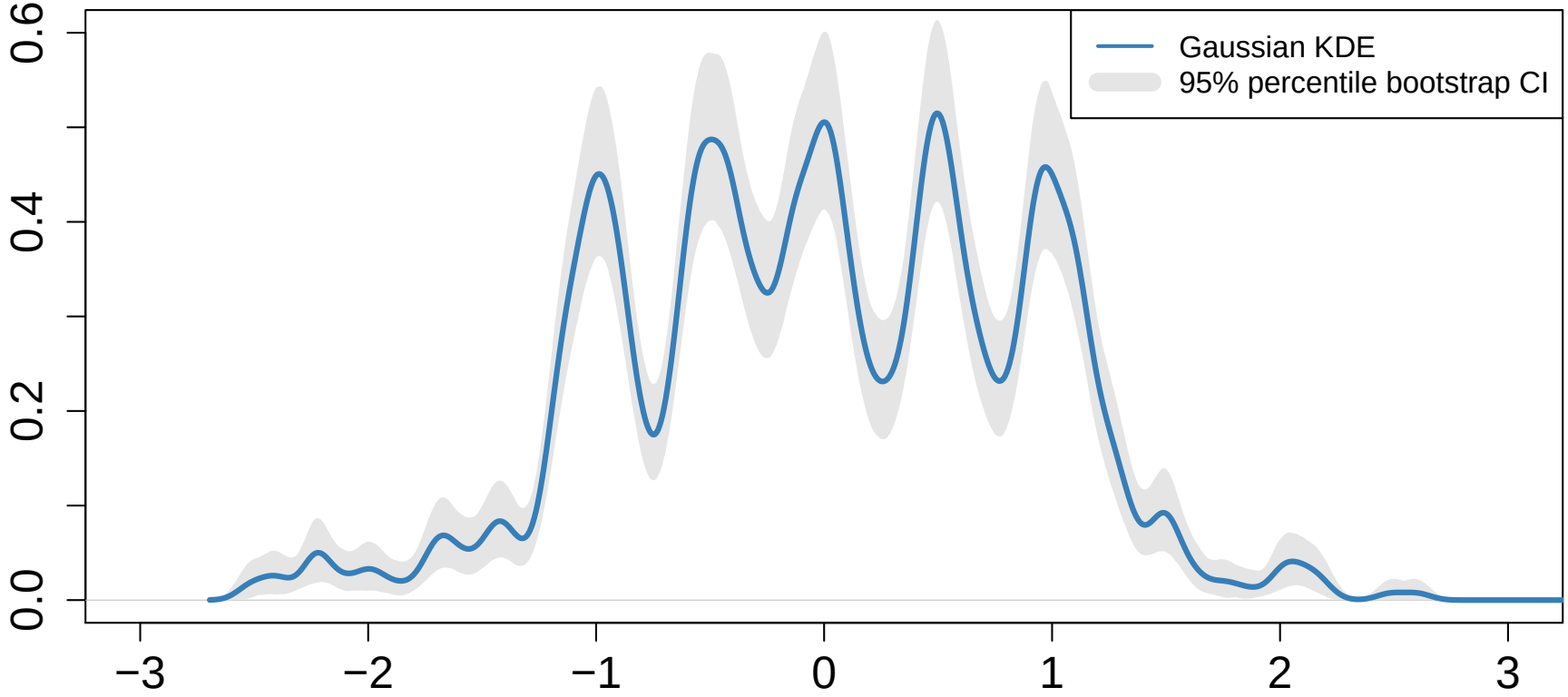
- Αριστερά: Η CVL λογαριθμική πιθανοφάνεια
- Δεξιά: Η εκτίμηση $\hat{f}_{h,n}(x)$ bandwidth που μεγιστοποεί την CVL ($h \approx 0.85$) και gaussian kernel

Παράδειγμα: Bart Simpson ($n = 1000$)



- Πραγματική συνάρτηση πυκνότητας $f(x)$
- Εκτίμηση μέ
  - ιστόγραμμα με εύρος κελιών $h = 0.104$ (βέλτιστο βάσει IMSE)
  - κανονικό πυρήνα και bandwidth $h = 0.059$ (βέλτιστο βάσει IMSE)
  - κανονικό πυρήνα και bandwidth $h = 0.085$ (βέλτιστο βάσει CVL)

Bootstrap διάστημα εμπιστοσύνης



- Εκτίμηση με κανονικό πυρήνα και bandwidth $h = 0.059$
- 95% bootstrap διάστημα εμπιστοσύνης ποσοστιαίων σημείων
- Μειονεκτήματα
  - Τεχνικά, είναι ΔΕ για την $\bar{f}_h(x) := \mathrm{E}\hat{f}_{h,n}(x)$ και όχι για την $f(x)$ ☹
  - Το Δ.Ε είναι σημειακό (όχι ταυτόχρονο για όλα τα x) ☹
- Αλλά: είναι χρήσιμο διότι μας σκιαγραφεί την αβεβαιότητα που έχουμε για τις πραγματικές κορυφές ☺

Input : $\alpha \in (0,1)$: $1-$συντελεστής εμπιστοσύνης

$B \in \mathbb{Z}_+$: αριθμός bootstrap δειγμάτων

$\boldsymbol{x} = (x_1, \ldots, x_n)$: παρατηρηθέντα δεδομένα

$M \in \mathbb{Z}_+$: ακέραιος που θα καθορίσει ένα σύνολο τιμών $(t_1, \ldots, t_M)$ στις οποίες θα υπολογιστεί η $\hat{f}_{h,n}(\cdot)$

$h > 0$: bandwidth.

Output: $C : M \times 2$ πίνακας με το άνω και κάτω όριο του ΔΕ στο t_m, $m = 1, \ldots, M$

Step 1: Bootstrap sampling

for $b = 1$ **to** B

Step 1.1 sampling:

Λάβε τυχαίο δείγμα $\boldsymbol{y}$ με επανάθεση από το $\boldsymbol{x}$, μεγέθους n

Step 1.2 Για κάθε t_m, υπολόγισε την εκτίμηση της πυκνότητας δοθέντος $\boldsymbol{y}$:

for all $m \in \{1, \ldots, M\}$: $\theta^{\star}_{b,m} = \hat{f}^{\star}_{h,n}(t_m)$;

endfor

Step 2: $100(1-\alpha)\%$ **percentile boostrap CI**

for $m = 1$ **to** M

$C_{m,1} = \alpha/2 -$ κάτω ποσοστιαίο σημείο του $\theta^{\star}_{1:B,m}$

$C_{m,2} = 1 - \alpha/2 -$ κάτω ποσοστιαίο σημείο του $\theta^{\star}_{1:B,m}$

endfor

END of algorithm

Παρατηρήσεις για επιλογή μέσω IMSE και CVL

- Γενικά, δεν υπάρχει βέλτιστη μέθοδος
- Και οι δύο τεχνικές έχουν μειονεκτήματα
- Η επιλογή μέσω της ελαχιστοποίησης του IMSE μέσω cross-validation έχει υψηλή μεταβλητότητα και συχνά καταλήγει σε undersmoothing
- Η επιλογή μέσω μεγιστοποίησης της CVL είναι ευαίσθητη σε ακραίες παρατηρήσεις και βαριές ουρές της πραγματικής f

Ιστόγραμμα ή χρήση πυρήνα ;

- Ασυμπτωτικά αποτελέσματα εξασφαλίζουν ότι η χρήση πυρήνα είναι προτιμότερη από την εκτίμηση μέσω ιστογράμματος
- Για το ιστόγραμμα ισχύει ότι το IMSE του εκτιμητή με βέλτιστο εύρος κελιών συγκλίνει στο 0 με ρυθμό της τάξης $n^{-2/3}$
- Ενώ με χρήση πυρήνα ισχύει ότι το IMSE του εκτιμητή με βέλτιστο bandwidth συγκλίνει στο 0 με ρυθμό της τάξης $n^{-4/5}$, δηλαδή πιο γρήγορα από ό,τι το ιστόγραμμα
- Δες σελ 74 και 84 σημειώσεων Ιωαννίδη για περισσότερες λεπτομέρειες.