# Πρόλογος – Ευχαριστίες

Η διπλωματική αυτή εργασία πραγματοποιήθηκε στα πλαίσια των σπουδών μου στη Σχολή των Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου, στον τομέα Μαθηματικών και ειδικότερα, στη ροή Στατιστικής. Ο σκοπός αυτής της διπλωματικής εργασίας είναι, πρώτον, να εξηγήσουμε πως μπορούμε να εκτιμήσουμε μια συνάρτηση πυκνότητας πιθανότητας έχοντας στη διάθεσή μας ένα δείγμα από ανεξάρτητες παρατηρήσεις, δηλαδή πως κατασκευάζεται μια εκτιμήτρια και, δεύτερον, να δείξουμε πως αυτές οι εκτιμήτριες μπορούν να χρησιμοποιηθούν, είτε με αυτοσκοπό, είτε ως ενδιάμεση διαδικασία άλλων στατιστικών διαδικασιών. Πιο συγκεκριμένα θα ασχοληθούμε με την εκτιμήτρια με την μέθοδο του πυρήνα για μονομεταβλήτες κυρίως παρατηρήσεις, θα δώσουμε τον ορισμό της, θα παρουσιάσουμε τρόπους που μειώνουν την ασυμφωνία της εκτιμήτριας σε σχέση με την πραγματική συνάρτηση, θα αποδείξουμε μεθόδους που σκοπό έχουν την επιλογή της βέλτιστης τιμής του πλάτους του κελιού και της συνάρτηση πυρήνα ώστε να μπορέσουμε να αποκτήσουμε μια βελτιωμένη εκτιμήτρια.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς Εξεταστικής Επιτροπής κ. Φουσκάκη Δ., κ. Βόντα Φ. και κ. Λουλάκη Μ. και ιδιαίτερα τον επιβλέποντα κ. Φουσκάκη Δημήτριο, Επίκουρο Καθηγητή του Ε.Μ.Π., που με εμπιστεύθηκε για την εκπόνηση της διπλωματικής αυτής εργασίας, καθώς και για την καθοδήγηση και τις συμβουλές που μου προσέφερε, και να αναφέρω πως χωρίς την συμβολή του, η άρτια ολοκλήρωση του παρόντος κειμένου δεν θα ήταν δυνατή.

# Περίληψη

Η συνάρτηση πυκνότητας πιθανότητας αποτελεί θεμέλιο για την στατιστική. Σε αυτήν την διπλωματική εργασία θα ασχοληθούμε με την κατασκευή εκτιμήσεων μιας άγνωστης συνάρτησης πυκνότητας πιθανότητας από δοθέντες παρατηρήσεις. Έτσι, στο δεύτερο κεφάλαιο θα δώσουμε τους ορισμούς των πιο διαδεδομένων εκτιμητριών και θα δείξουμε πώς κατασκευάζονται. Στο τρίτο κεφάλαιο θα ασχοληθούμε εξ' ολοκλήρου με την εκτιμήτρια με την μέθοδο του πυρήνα για μονομεταβλήτες παρατηρήσεις, θα δώσουμε τον ορισμό της, θα παρουσιάσουμε τρόπους που μειώνουν την ασυμφωνία της εκτιμήτριας σε σχέση με την πραγματική συνάρτηση και θα αποδείξουμε μεθόδους που σκοπό έχουν την επιλογή της βέλτιστης τιμής του πλάτους του κελιού και της συνάρτηση πυρήνα, ώστε να μπορέσουμε να αποκτήσουμε μια βελτιωμένη εκτιμήτρια. Στο τέταρτο κεφάλαιο θα γίνει η περιγραφή της εκτιμήτριας με την μέθοδο του πυρήνα για την περίπτωση που το δείγμα μας αποτελείται από πολυμεταβλητές παρατηρήσεις και στο τελευταίο κεφάλαιο θα ορίσουμε μια διαφορετική εκτιμήτρια, την Nadaraya-Watson εκτιμήτρια, η οποία κατασκευάζεται από παρατηρήσεις που προέρχονται από μια από κοινού συνάρτησης πυκνότητας πιθανότητας.

# Abstract

The probability density function is a fundamental concept in statistics. In this thesis we will deal with the construction of estimations of an unknown probability density function from a given data set. In second chapter we will give the definitions of the most widespread estimators and will explain how they are constructed. In third chapter we will concentrate to kernel estimator for univariate data, we will give their definitions, present some measures which decrease the discrepancy between the density estimator and the true density and will discuss various methods for choosing the optimal smoothing parameter so as to acquire an optimal estimator. In fourth chapter we will give the description of the kernel estimation for multivariate data and in the last chapter we will give the definition of the Nadaraya-Watson estimator, constructed from data from a jointly probability density function.

# ΠΕΡΙΕΧΟΜΕΝΑ

# Λίστα Πινάκων



# Λίστα Γραφημάτων

# ΚΕΦΑΛΑΙΟ 1

## 1. ΕΙΣΑΓΩΓΗ

### 1.1. Ορισμός

Η συνάρτηση πυκνότητας πιθανότητας αποτελεί θεμέλιο για την στατιστική. Θεωρούμε μία τυχαία μεταβλητή $X$ η οποία έχει συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $f$. Η σ.π.π $f$ μας δίνει την πλήρη περιγραφή της κατανομής της τ.μ. $X$ και βοηθάει στην εύρεση των πιθανοτήτων με την χρήση της σχέσης:

$$P(a < X < b) = \int_a^b f(x)\,dx,\ \forall a < b.$$

Υποθέτουμε ότι διαθέτουμε ένα δείγμα από παρατηρήσεις που προέρχονται από μια άγνωστη σ.π.π. $f$. Με τον όρο εκτιμήτρια $\hat{f}$ της σ.π.π. $f$ εννοούμε την κατασκευή μίας εκτίμησης της άγνωστης σ.π.π. από τις δοθέντες παρατηρήσεις, ή με άλλα λόγια μια συνάρτηση του τυχαίου δείγματος που χρησιμοποιείται για την εκτίμηση μιας άγνωστης παραμέτρου μιας συνάρτησης κατανομής. Από διαγνωστικής απόψεως η εκτίμηση της $f$, έναντι της συνάρτησης κατανομής $F$, είναι προτιμότερη διότι εμφανίζει ευκρινέστερα τις συγκεντρώσεις μάζας πιθανότητας, ενώ η $F$, λόγω της ολοκλήρωσης, τις εξομαλύνει.

### 1.2. Σκοπός

Ο σκοπός αυτής της διπλωματικής εργασίας είναι, πρώτον να εξηγήσουμε πως μπορούμε να εκτιμήσουμε μια σ.π.π. από δοθέντες παρατηρήσεις και, δεύτερον, να δείξουμε πως οι εκτιμήτριες αυτές μπορούν να χρησιμοποιηθούν, είτε με αυτοσκοπό είτε ως ενδιάμεση διαδικασία άλλων στατιστικών διαδικασιών.

Η προσέγγιση της εκτίμησης της σ.π.π μπορεί να είναι παραμετρική. Δηλαδή, εάν υποθέσουμε ότι επεξεργαζόμαστε παρατηρήσεις που ανήκουν στην οικογένεια των κανονικών κατανομών με μέση τιμή μ και συνδιακυμανση $\sigma^2$, η σ.π.π. $f$ θα μπορούσε να εκτιμηθεί βρίσκοντας εκτιμήσεις για τις παραμέτρους μ και $\sigma^2$ και αντικαθιστώντας αυτές στην σχέση της εκτίμησης για τις κανονικές κατανομές. Εμείς, όμως, θα

ασχοληθούμε κυρίως με μη παραμετρικές προσεγγίσεις έτσι ώστε να κάνουμε όσο το δυνατό λιγότερες υποθέσεις σε σχέση με την κατανομή των παρατηρήσεων.

Μια εκτιμήτρια μπορεί να χρησιμοποιηθεί είτε για εξερεύνηση είτε για παρουσίαση των δεδομένων. Όταν χρησιμοποιείται για εξερεύνηση, μπορεί να δώσει ενδείξεις για διάφορες δομές των δεδομένων, όπως λοξότητα και πολυπλοκότητα.

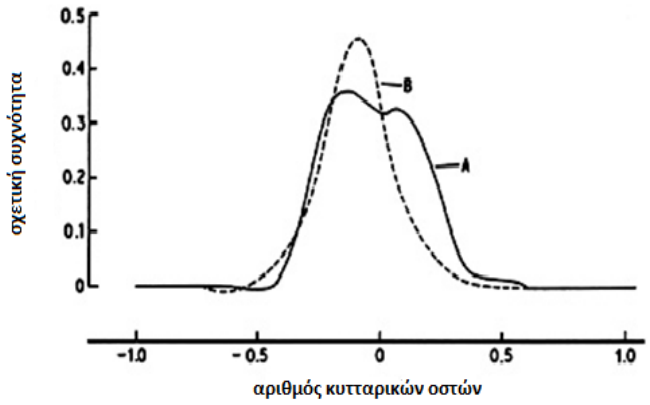
Ολοκληρώνοντας την εισαγωγή μας τονίζουμε ότι οι εκτιμήτριες χρησιμοποιούνται σε διάφορους τομείς, όπως π.χ. στην διακριτική ανάλυση, αλλά εμείς θα ασχοληθούμε στα επόμενα κεφάλαια κυρίως με το πώς μια εκτιμήτρια κατασκευάζεται. Για αυτό το λόγο στην επόμενη παράγραφο θα προσπαθήσουμε να προσεγγίσουμε την έννοια της εκτιμήτριας και μια από τις σημαντικότερές της εφαρμογές, την εξερεύνηση και την παρουσίαση των δεδομένων.

## 1.3. Η χρήση της εκτιμήτριας για την εξερεύνηση και παρουσίαση των δεδομένων

Κάνουμε χρήση των εκτιμητριών για να ανακαλύψουμε ιδιότητες των παρατηρήσεων ενός δείγματος, καθώς αυτές μας δίνουν χρήσιμες ενδείξεις για κάποια χαρακτηριστικά τους, όπως λοξότητα και πολυπλοκότητα. Σε μερικές περιπτώσεις αυτές οι ενδείξεις μπορούν να θεωρηθούν ισχύοντες ενώ σε άλλες μας δείχνουν τον δρόμο για περαιτέρω ανάλυση.

Ένα παράδειγμα δίνεται στο Γράφημα 1.1. Οι καμπύλες έχουν κατασκευαστεί από τους Emery και Carpenter (1974) στην προσπάθεια τους να εξετάσουν ένα σύνδρομο από ξαφνικούς θανάτους βρεφών. Η καμπύλη Α κατασκευάστηκε από παρατηρήσεις οι οποίες δείχνουν τον αριθμό των κατεστραμμένων κυτταρικών ιστών από κάθε ένα από τα 95 βρέφη που πέθαναν ξαφνικά και ανεξήγητα, ενώ η καμπύλη Β κατασκευάστηκε από παρατηρήσεις οι οποίες δείχνουν τον αριθμό των κατεστραμμένων κυτταρικών ιστών από κάθε ένα από τα 76 βρέφη τα οποία πέθαναν από γνωστή αιτία η οποία, όμως, δεν επηρέαζε των αριθμό των κατεστραμμένων κυτταρικών ιστών. Οι ερευνητές κατέληξαν προσωρινά στο συμπέρασμα ότι η σ.π.π. που προκύπτει στην περίπτωση των ξαφνικών θανάτων ίσως είναι μια μείξη της λεγόμενης σ.π.π. σε συνδυασμό, σε μικρή αναλογία, με μια επεξεργασμένη σ.π.π. με υψηλότερη μέση τιμή. Οπότε, φαίνεται ότι ο

αριθμός των κατεστραμμένων κυτταρικών ιστών ήταν απρόσμενα υψηλός σε ένα μικρό ποσοστό των βρεφών που βρήκαν ξαφνικό θάνατο. Σε αυτό το παράδειγμα το συμπέρασμα θα μπορούσε να θεωρηθεί ως μέσο για περαιτέρω ανάλυση.




**Γράφημα 1.1: Εκτιμήτριες που έχουν κατασκευαστεί από τα κατεστραμμένα κυτταρικά οστά σε μια έρευνα ξαφνικών θανάτων, Α) ανεξήγητοι θάνατοι, Β) γνωστής αιτίας θάνατοι.**

# ΚΕΦΑΛΑΙΟ 2

## 2. ΟΙ ΒΑΣΙΚΟΙ ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ

### 2.1. Το ιστόγραμμα

Το ιστόγραμμα είναι η παλαιότερη και η πιο διαδεδομένη μέθοδος εκτίμησης. Δοσμένης μιας αρχικής τιμής $x_0$ και ενός πλάτος κελιού $h$ μπορούμε να ορίσουμε το πλάτος των κελιών του ιστογράμματος να είναι το διάστημα:

$$[x_0 + mh\,, x_0 + (m+1)h\,],$$

για θετικούς και αρνητικούς ακεραίους $m$. Το διάστημα επιλέγετε κλειστό από τα αριστερά και ανοιχτό από τα δεξιά για λόγους οριστικότητας.

#### 2.1.1. Ορισμός

Το ιστόγραμμα είναι η γραφική απεικόνιση στατιστικών συχνοτήτων περιοχών τιμών ενός μεγέθους. Έτσι το ιστόγραμμα ορίζεται από την σχέση:

$$\hat{f}(x) = \frac{1}{nh}\ (o\ αριθ.των\ παρατηρ.X_i\ που\ βρίσκονται\ στο\ ίδιο\ κελί\ με\ το\ x). \quad (2.1)$$

Έτσι για να ορίσουμε ένα ιστόγραμμα αρκεί εκ των προτέρων να ορίσουμε την αρχική τιμή $x_0$ και την τιμή του πλάτους του κελιού $h$, που ελέγχει κατά κύριο λόγο την τιμή της εξομάλυνσης κατά την διαδικασία. Όμως, το ιστόγραμμα μπορεί να πάρει και μία πιο γενικευμένη μορφή, επιτρέποντας τα πλάτη των κελιών να ποικίλουν.

Η γενικευμένη μορφή του ιστογράμματος είναι:

$$\hat{f}(x) = \frac{1}{n}\ \frac{(o\ αριθ.των\ παρατηρ.X_ιπου\ βρίσκονται\ στο\ ίδιο\ κελί\ με\ το\ x)}{(το\ πλάτος\ \ του\ κελιού\ \ που\ περιέχει\ το\ x)}. \quad (2.2)$$

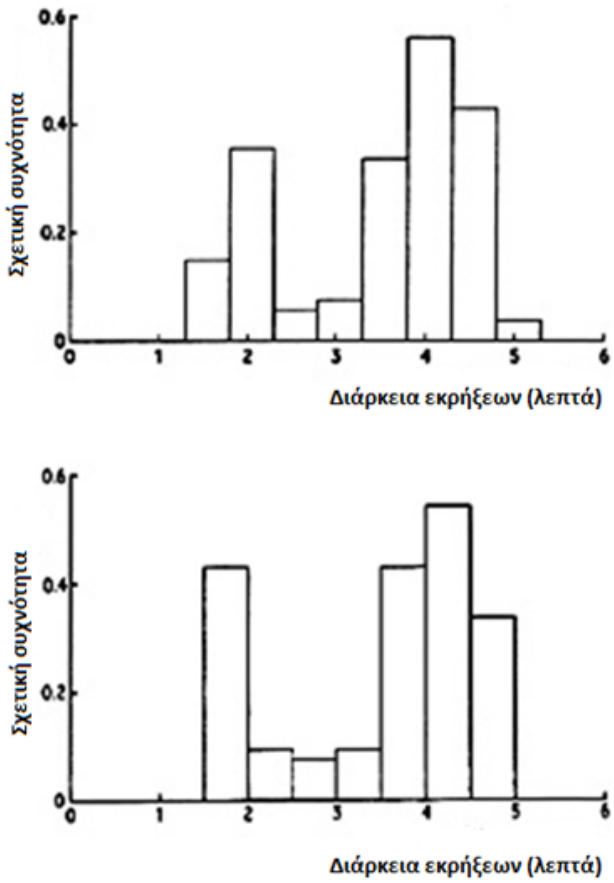
Οι τιμές στον παρακάτω πινάκα δείχνουν την διάρκεια σε λεπτά 107 εκρήξεων του old faithful ηφαιστείου.

| | | | | | | |
|---|---|---|---|---|---|---|
| 4.37 | 3.87 | 4.00 | 4.03 | 3.50 | 4.08 | 2.25 |
| 4.70 | 1.73 | 4.93 | 1.73 | 4.62 | 3.43 | 4.25 |
| 1.68 | 3.92 | 3.68 | 3.10 | 4.03 | 1.77 | 4.08 |
| 1.75 | 3.20 | 1.85 | 4.62 | 1.97 | 4.50 | 3.92 |
| 4.35 | 2.33 | 3.83 | 1.88 | 4.60 | 1.80 | 4.73 |
| 1.77 | 4.57 | 1.85 | 3.52 | 4.00 | 3.70 | 3.72 |
| 4.25 | 3.58 | 3.80 | 3.77 | 3.75 | 2.50 | 4.50 |
| 4.10 | 3.70 | 3.80 | 3.43 | 4.00 | 2.27 | 4.40 |
| 4.05 | 4.25 | 3.33 | 2.00 | 4.33 | 2.93 | 4.58 |
| 1.90 | 3.58 | 3.73 | 3.73 | 1.82 | 4.63 | 3.50 |
| 4.00 | 3.67 | 1.67 | 4.60 | 1.67 | 4.00 | 1.80 |
| 4.42 | 1.90 | 4.63 | 2.93 | 3.50 | 1.97 | 4.28 |
| 1.83 | 4.13 | 1.83 | 4.65 | 4.20 | 3.93 | 4.33 |
| 1.83 | 4.53 | 2.03 | 4.18 | 4.43 | 4.07 | 4.13 |
| 3.95 | 4.10 | 2.72 | 4.58 | 1.90 | 4.50 | 1.95 |
| 4.83 | 4.12 | | | | | |

**Πίνακας 2.1: Διάρκεια σε λεπτά 107 εκρήξεων του old faithful ηφαιστείου.**

Για την παρουσίαση και εξερεύνηση των δεδομένων του Πίνακα 2.1 που ακολουθεί το ιστόγραμμα είναι ικανοποιητικό. Ωστόσο, η επιλογή της αρχικής τιμής έχει μεγάλη επίδραση. Στο Γράφημα 2.1 παρουσιάζονται 2 ιστογράμματα που έχουν κατασκευαστεί με το ίδιο πλάτος κελιού $h$ και διαφορετικές αρχικές τιμές. Το γενικό μήνυμα και στις 2 περιπτώσεις είναι το ίδιο, όμως κάποιος μη-στατιστικός θα μπορούσε να λάβει διαφορετικές εντυπώσεις βλέποντάς τα.

Πολλοί θα αναρωτιούνται γιατί να μην μπορούμε να χρησιμοποιούμε το ιστόγραμμα σε όλες τις στατιστικές διαδικασίες και πολλές φορές ψάχνουμε μεθόδους πιο προχωρημένες. Αυτό οφείλεται στο γεγονός ότι το ιστόγραμμα έχει ένα σημαντικό μειονέκτημα, το οποίο μεταφράζεται ως αναποτελεσματική χρήση των δεδομένων, κυρίως όταν χρησιμοποιείται ως εκτιμήτρια σε διαδικασίες όπως η αθροιστική ανάλυση




**Γράφημα 2.1: Ιστογράμματα που έχουν κατασκευαστεί από τα δεδομένα του Πίνακα 2.1 έχοντας χρησιμοποιήσει διαφορετικές αρχικές τιμές.**

ή η μη-παραμετρική διακριτή ανάλυση. Το μειονέκτημα του είναι η ασυνέχεια που παρουσιάζει, η οποία προκαλεί δυσκολίες όταν απαιτούνται παράγωγα της εκτίμησης, και για αυτό το λόγο εάν θέλουμε η εκτιμήτρια να χρησιμοποιηθεί ως ενδιάμεση διαδικασία σε άλλες μεθόδους, προκύπτει η ανάγκη να χρησιμοποιήσουμε μια άλλη εκτιμήτρια αντί του ιστογράμματος.

Γενικά, το ιστόγραμμα είναι ικανοποιητικό μόνο για εξερεύνηση ή παρουσίαση των δεδομένων και κυρίως για μονομεταβλητές περιπτώσεις, γιατί σε διμεταβλητά ή τριμεταβλητά δεδομένα, είναι δύσκολο να το σχεδιάσουμε τρισδιάστατα αφού η εκτιμήτρια δεν εξαρτάται μόνο από την αρχική τιμή $x_0$ αλλά και από τις διευθύνσεις των κελιών.

## 2.2. Ο απλοϊκός εκτιμητής

Γνωρίζουμε ότι εάν έχουμε μια τυχαία μεταβλητή $X$, η σ.π.π. ορίζεται από την σχέση:

$$f(x) = lim_{h\to 0}\frac{1}{2h}P(x-h<X<x+h). \quad (2.3)$$

Για δοσμένο $h$, μπορούμε να εκτιμήσουμε την πιθανότητα $P(x-h<X<x+h)$ από την αναλογία του δείγματος που ανήκει στο διάστημα $(x-h, x+h)$. Έτσι προκύπτει και ο ορισμός του απλοϊκού εκτιμητή:

$$\hat{f}(x) = \frac{1}{2nh}[o\ αριθ.των\ παρατ.X_i\ που\ ανήκουν\ στο\ διάστ\ (x-h, x+h)]. \quad (2.4)$$

Για να εκφράσουμε καλύτερα τον απλοϊκό εκτιμητή πρέπει να ορίσουμε μια συνάρτηση βάρους $w$ ως έξης:

$$w(x) = \begin{cases} \frac{1}{2}, & εαν\ |x|<1 \\ 0, & διαφορετικα \end{cases}, \quad (2.5)$$

όπου $w(x)$ είναι η σ.π.π. μιας συνεχούς ομοιόμορφης κατανομής, οπότε, ο απλοϊκός εκτιμητής θα πάρει την μορφή :

$$\hat{f}(x) = \frac{1}{n}\sum_1^n \frac{1}{h}w(\frac{x-X_i}{h}). \quad (2.6)$$

Από την σχέση (2.5) φαίνεται ότι αντικαθιστώντας κάθε παρατήρηση με ένα ″κουτί″ πλάτους $2h$ και ύψους $(2nh)^{-1}$ και αθροίζοντας αυτά τα ΄΄κουτιά΄΄ προκύπτει η εκτίμηση μας.

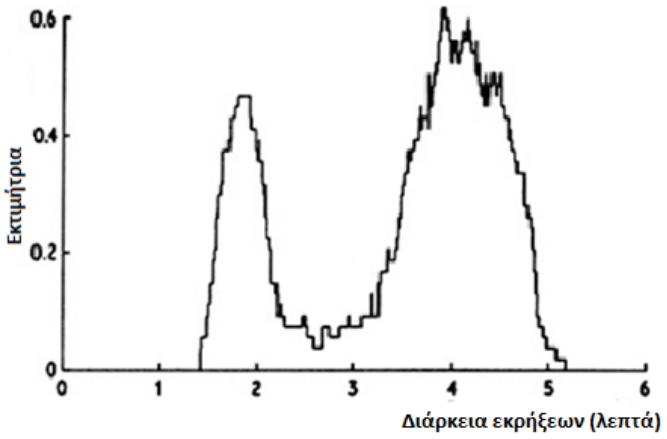
### 2.2.1. Σχέση του ιστογράμματος με τον απλοϊκό εκτιμητή

Θεωρούμε ότι το ιστόγραμμα κατασκευάζεται χρησιμοποιώντας κελιά πλάτους $2h$ και ότι δεν υπάρχουν παρατηρήσεις στην άκρη των κελιών. Εάν το $x$ βρίσκεται στο κέντρο

κάθε κελιού, από την σχέση (2.5) προκύπτει ότι ο απλοϊκός εκτιμητής $\hat{f}(x)$ θα είναι η τεταγμένη του ιστογράμματος στο $x$. Έτσι μπορούμε να θεωρήσουμε ότι ο απλοϊκός εκτιμητής είναι μια προσπάθεια κατασκευής ενός ιστογράμματος, όπου κάθε σημείο βρίσκεται στο κέντρο του δειγματικού διαστήματος, απελευθερώνοντας έτσι το ιστόγραμμα από την ανάγκη επιλογής της θέσης των κελιών. Ωστόσο, η επιλογή του πλάτους του κουτιού εξαρτάται από την παράμετρο $h$, η οποία ελέγχει την συνολική ποσότητα εξομάλυνσης.

Επιπλέον, ο απλοϊκός εκτιμητής, σε αντίθεση με το ιστόγραμμα δεν είναι τόσο χρήσιμος για παρουσίαση δεδομένων και αυτό συμβαίνει γιατί από τον ορισμό του έχουμε ότι η $\hat{f}$ δεν είναι συνεχής συνάρτηση, αλλά πηδά από τα σημεία $X_i \pm h$ και έχει παράγωγα 0 οπουδήποτε αλλού. Αυτό έχει ως συνέπεια την δημιουργία παραπλανητικών συμπερασμάτων. Για την αποφυγή, λοιπόν, αυτών των προβλημάτων χρησιμοποιούμε τον απλοϊκό εκτιμητή στην γενική του μορφή, την οποία θα παρουσιάσουμε στη επόμενη παράγραφο.

Μια εκτίμηση της σ.π.π., χρησιμοποιώντας τον απλοϊκό εκτιμητή για τις παρατηρήσεις του Πίνακα 2.1, που δείχνουν την διάρκεια σε λεπτά 107 εκρήξεων του old faithful ηφαιστείου, παρουσιάζεται στο Γράφημα 2.2. Τα ″κουτιά″ που χρησιμοποιήθηκαν στην εκτίμηση έχουν το ίδιο πλάτος με το πλάτος των κελιών του ιστογράμματος του Γραφήματος 2.1.




**Γράφημα 2.2: Απλοϊκός εκτιμητής που κατασκευάστηκε από τις παρατηρήσεις του Πίνακα 2.1 με πλάτος κελιού $h = 0.25$.**

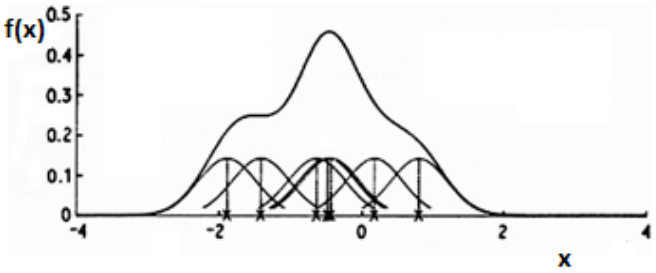
## 2.3. Η εκτιμήτρια με την μέθοδο του πυρήνα

Η εκτιμήτρια με την μέθοδο του πυρήνα είναι μια γενίκευση του απλοϊκού εκτιμητή που σκοπό έχει να αντιμετωπίσει τις διάφορες δυσκολίες που συζητήθηκαν προηγουμένως. Έτσι αντικαθιστώντας την συνάρτηση βάρους $w(x)$ με μια συνάρτηση πυρήνα $K$, που θα ικανοποιεί την συνθήκη $\int_{-\infty}^{+\infty} K(x)\,dx = 1$, η εκτιμήτρια με την μέθοδο του πυρήνα έχει την μορφή:

$$\hat{f}(x) = \frac{1}{nh}\sum_{1}^{n} K\left(\frac{x - X_i}{h}\right). \qquad (2.7)$$

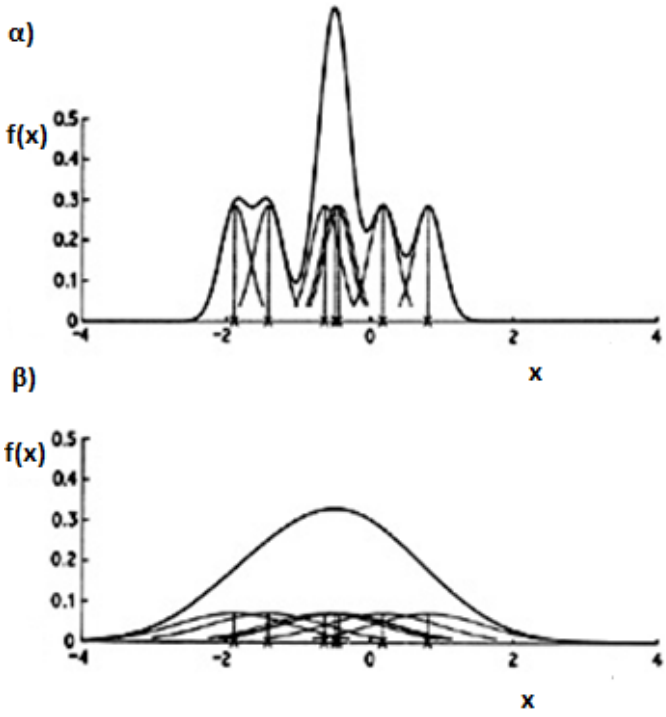
Όπως ο απλοϊκός εκτιμητής μπορεί να θεωρηθεί ως πρόσθεση των ΄΄κουτιών΄΄ έτσι και η εκτιμήτρια με την μέθοδο του πυρήνα μπορεί να θεωρηθεί ως πρόσθεση των καμπύλων (bumps). Η συνάρτηση πυρήνα $K$ καθορίζει το σχήμα της καμπύλης ενώ το $h$ καθορίζει το πλάτος της.

Στην συνέχεια θα παραθέσουμε κάποια γραφήματα από τα οποία θα δούμε πως η εκτιμήτρια με την μέθοδο του πυρήνα κατασκευάζεται και πως επηρεάζεται από την αλλαγή του $h$.




**Γράφημα 2.3: Εκτιμήτρια πυρήνα που προκύπτει από τις μεμονωμένες καμπύλες. Πλάτος κελιού $h = 0.4$.**

Στο Γράφημα 2.3 φαίνεται ότι η εκτιμήτρια πυρήνα κατασκευάζεται από το άθροισμα των μεμονωμένων καμπύλων. Στο Γράφημα 2.4, που ακολουθεί, φαίνεται η επίδραση της αλλαγής του $h$. Το όριο καθώς το $h$ τείνει στο 0 είναι το άθροισμα των αιχμών της δέλτα Dirac συνάρτησης (Γράφημα α), ενώ καθώς το $h$ μεγαλώνει πολλές λεπτομέρειες της κατανομής δεν γίνονται φανερές (Γράφημα b).




**Γράφημα 2.4: Εκτιμήτριες πυρήνα που προκύπτουν από τις μεμονωμένες καμπύλες. Πλάτος κελιού α) $h = 0.2$ β) $h = 0.8$.**

### 2.3.1. Βασικές ιδιότητες της εκτιμήτριας με την μέθοδο του πυρήνα

Δεδομένου ότι η συνάρτηση πυρήνα $K$ είναι παντού μη-αρνητική και ικανοποιεί την συνθήκη $\int_{-\infty}^{+\infty} K(x)\,dx = 1$, δηλαδή είναι μια σ.π.π., συνεπάγεται ότι και η εκτιμήτρια $\hat{f}$ θα είναι και αυτή μια σ.π.π.. Επιπλέον η $\hat{f}$ κληρονομεί όλες τις ιδιότητες της συνάρτησης πυρήνα $K$ όπως π.χ. την συνέχεια και την διαφορισιμότητα. Έτσι εάν η $K$ είναι μια κανονική σ.π.π. τότε συνεπάγεται ότι και η $\hat{f}$ θα είναι μια ομαλή καμπύλη.

.

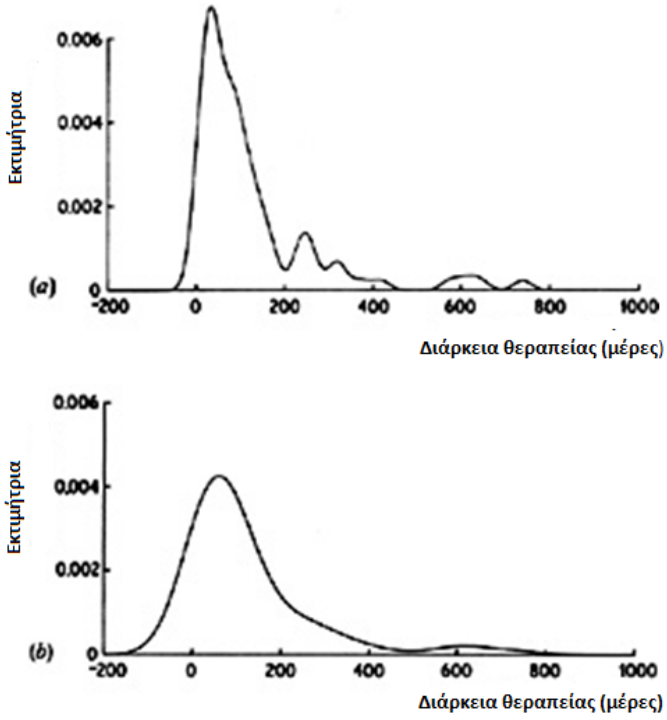
| | | | | | | |
|---|---|---|---|---|---|---|
| 4.37 | 3.87 | 4.00 | 4.03 | 3.50 | 4.08 | 2.25 |
| 4.70 | 1.73 | 4.93 | 1.73 | 4.62 | 3.43 | 4.25 |
| 1.68 | 3.92 | 3.68 | 3.10 | 4.03 | 1.77 | 4.08 |
| 1.75 | 3.20 | 1.85 | 4.62 | 1.97 | 4.50 | 3.92 |
| 4.35 | 2.33 | 3.83 | 1.88 | 4.60 | 1.80 | 4.73 |
| 1.77 | 4.57 | 1.85 | 3.52 | 4.00 | 3.70 | 3.72 |
| 4.25 | 3.58 | 3.80 | 3.77 | 3.75 | 2.50 | 4.50 |
| 4.10 | 3.70 | 3.80 | 3.43 | 4.00 | 2.27 | 4.40 |
| 4.05 | 4.25 | 3.33 | 2.00 | 4.33 | 2.93 | 4.58 |
| 1.90 | 3.58 | 3.73 | 3.73 | 1.82 | 4.63 | 3.50 |
| 4.00 | 3.67 | 1.67 | 4.60 | 1.67 | 4.00 | 1.80 |
| 4.42 | 1.90 | 4.63 | 2.93 | 3.50 | 1.97 | 4.28 |
| 1.83 | 4.13 | 1.83 | 4.65 | 4.20 | 3.93 | 4.33 |
| 1.83 | 4.53 | 2.03 | 4.18 | 4.43 | 4.07 | 4.13 |
| 3.95 | 4.10 | 2.72 | 4.58 | 1.90 | 4.50 | 1.95 |
| 4.83 | 4.12 | | | | | |

**Πίνακας 2.2: Η διάρκεια (σε μέρες) 86 ασθενών οι οποίοι υφίσταται ψυχιατρική θεραπεία έτσι ώστε να ελεγχθούν οι κίνδυνοι αυτοκτονίας.**

Τέλος, η εκτιμήτρια με την μέθοδο του πυρήνα είναι η εκτιμήτρια που χρησιμοποιείται πιο πολύ από όλες και για αυτό το λόγο είναι και η πιο μελετημένη μαθηματικώς.

Ωστόσο, αυτή η μέθοδος πάσχει από ένα σοβαρό μειονέκτημα όταν εφαρμόζεται σε δεδομένα που ανήκουν σε κατανομές με μακριές ουρές. Επειδή το $h$ είναι σταθερό σε όλο το μήκος των παρατηρήσεων, υπάρχει μια τάση ανωμαλίας στην ουρά της εκτίμησης. Ωστόσο, εάν η εκτίμηση είναι εξομαλυμένη ώστε να μπορεί να αντιμετωπίσει αυτό το πρόβλημα, τότε σημαντικές ιδιότητες από το κεντρικό μέρος της κατανομής δεν γίνονται φανερές

Η εκτίμηση που φαίνεται στο Γράφημα 2.5(α) με $h = 20$ παρουσιάζει θόρυβο στην δεξιά ουρά ενώ η εκτίμηση στο Γράφημα 2.5(b) με $h = 60$ παρουσιάζει μια πιο ομαλή καμπύλη στην ουρά. Ωστόσο, στην δεύτερη περίπτωση το μήκος της καμπύλης στο κεντρικό μέρος της κατανομής μεγαλώνει.




**Γράφημα 2.5: Εκτιμήτριες πυρήνα για τα δεδομένα του Πίνακα 2.2. Πλάτη κελιών α) $h = 20$, β) $h = 60$.**

## 2.4. Η εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα

Αυτή η μέθοδος περιγράφει μια προσπάθεια να προσαρμόσει το ποσοστό της εξομάλυνσης στην αρχική σ.π.π. των δεδομένων. Το ποσοστό της εξομάλυνσης ελέγχεται από έναν ακέραιο $k$ ο οποίος επιλέγεται να είναι ίσος με $k = n^{1/2}$. Ορίζουμε την απόσταση $d(x, y)$ μεταξύ 2 σημείων στην ευθεία να είναι $|x - y|$ και για κάθε t έχουμε:

$$d_1(t) \leq d_2(t) \leq \cdots \leq d_n(t), \tag{2.8}$$

να είναι οι αποστάσεις του t από τα σημεία του δείγματος σε αύξουσα σειρά.

Έτσι η εκτιμήτρια με την μέθοδο των K-κοντινότερων γειτόνων ορίζεται ως εξής:

$$\hat{f}(t) = \frac{k}{2nd_k(t)}. \tag{2.9}$$

Για να καταλάβουμε πως προκύπτει ο ορισμός, υποθέτουμε ότι η σ.π.π. στο $t$ είναι $f(t)$. Σε δείγμα μεγέθους $n$ αναμένεται ότι $2rnf(t)$ παρατηρήσεις θα ανήκουν στο διάστημα $(t - r, t + r)$ και ότι ακριβώς $k$ παρατηρήσεις θα ανήκουν στο διάστημα $[t - d_k(t), t + d_k(t)]$. Έτσι έχουμε ότι:

$$k = 2d_k(t)n\hat{f}(t). \tag{2.10}$$

### 2.4.1. Σχέση του απλοϊκού εκτιμητή με την εκτιμήτρια του κοντινότερου γείτονα

Ενώ ο απλοϊκός εκτιμητής βασίζεται στον αριθμό των παρατηρήσεων που ανήκουν σε ένα ΄΄κουτί΄΄ σταθερού πλάτους, η εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα είναι αντιστρόφως ανάλογη του μεγέθους του ΄΄κουτιού΄΄ που χρειάζεται ώστε να περιέχει ένα συγκεκριμένο αριθμό παρατηρήσεων.

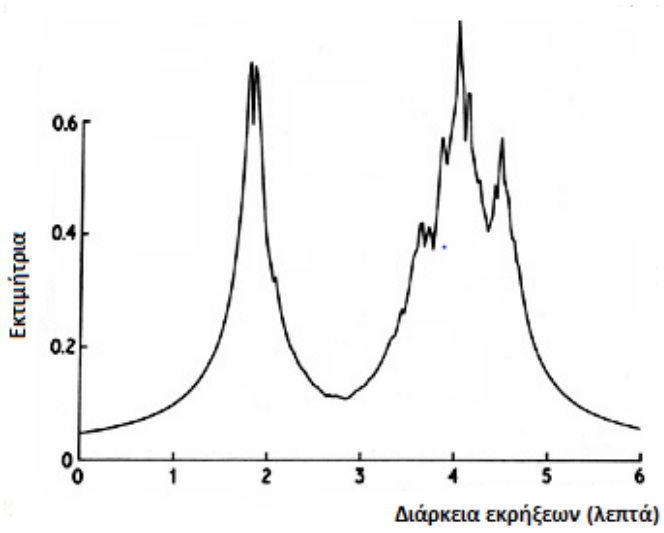
Παρατηρούμε ακόμη ότι η απόσταση $d_k(t)$ στην ουρά της κατανομής είναι μεγαλύτερη από ότι στο κεντρικό μέρος της και έτσι το πρόβλημα της εξομάλυνσης στην ουρά μειώνεται.

Όμως τόσο ο απλοϊκός εκτιμητής όσο και η εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα δεν παριστάνονται με μια ομαλή καμπύλη, για τον λόγο ότι η συνάρτηση $d_k(t)$ είναι συνεχής αλλά τα παράγωγά της παρουσιάζουν ασυνέχεια στα σημεία $2^{-1}(X_{(j)} +$

$X_{(j+k)}$) όπου $X_{(j)}$ είναι η j διατεταγμένη παρατήρηση. Έτσι η $\hat{f}$ θα είναι θετική και συνεχής παντού, αλλά θα έχει ασυνεχή παράγωγα στα σημεία που και η $d_k(t)$ έχει.

### 2.4.2. Βασικές ιδιότητες της εκτιμήτριας με την μέθοδο του κοντινότερου γείτονα

Σε αντίθεση με την εκτιμήτρια πυρήνα, η εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα δεν είναι μια σ.π.π. γιατί δεν έχει ολοκλήρωμα την μονάδα. Για $t$ μικρότερο από την μικρότερη παρατήρηση έχουμε $d_k(t) = X_{(n-k+1)}$ και για t>$X_n$ έχουμε $d_k(t) = t - X_{(n-k+1)}$. Εάν αντικαταστήσουμε στην σχέση (2.9), έχουμε ότι το ολοκλήρωμα $\int_{-\infty}^{+\infty} t\,dt$ είναι άπειρο και ότι η ουρά της $\hat{f}$ σβήνει με ρυθμό $t^{-1}$, δηλαδή πολύ αργά. Έτσι η εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα δεν θα είναι κατάλληλη εάν απαιτείται μια εκτίμηση ολόκληρης της σ.π.π.. Το Γράφημα 2.6 παρουσιάζει την εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα για τις παρατηρήσεις του Πίνακα 2.1. Η ασυνέχεια και η γεμάτη ουρά στα παράγωγα γίνονται φανερά.




**Γράφημα 2.6: Εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα για τις παρατηρήσεις του Πίνακα 2.1. με $k = 20$.**

Μπορούμε βέβαια να γενικεύσουμε αυτήν την εκτιμήτρια ώστε να προκύψει μια εκτιμήτρια σχετική με την εκτιμήτρια με την μέθοδο του πυρήνα, θέτοντας $K$ να είναι η συνάρτηση πυρήνα με ολοκλήρωμα την μονάδα. Έτσι, η εκτιμήτρια παίρνει την μορφή:

$$\hat{f}(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} K\left(\frac{t-X_i}{d_k(t)}\right), \qquad (2.11)$$

η οποία είναι ακριβώς η ίδια με την εκτιμήτρια με την μέθοδο του πυρήνα, που εκτιμήθηκε στα σημεία $t$ με πλάτος κελιού $d_k(t)$. Έτσι το ποσοστό της εξομάλυνσης εξαρτάται από την επιλογή του ακέραιου $k$ ενώ το $d_k(t)$ εξαρτάται από την σ.π.π. των παρατηρήσεων κοντά στο σημείο.

Η εκτιμήτρια με την μέθοδο των Κ-κοντινότερων γειτόνων είναι μια ειδική περίπτωση της σχέσης (2.11), όπου η συνάρτηση πυρήνα $K$ είναι μια ομοιόμορφη συνάρτηση όπως ορίστηκε στην σχέση (2.5). Έτσι η παράσταση (2.11) έχει την ίδια σχέση με την (2.9), όπως έχει η εκτιμήτρια πυρήνα με τον απλοϊκό εκτιμητή. Ωστόσο, τα παράγωγα της γενικευμένης εκτιμήτριας με την μέθοδο του κοντινότερου γείτονα θα είναι ασυνεχή στα σημεία όπου και η συνάρτηση $d_k(t)$ έχει ασυνεχή παράγωγα.

## 2.5. Η εκτιμήτρια μεταβλητού πυρήνα

Η εκτιμήτρια μεταβλητού πυρήνα σχετίζεται με την προσέγγιση του κοντινότερου γείτονα αφού είναι και αυτή μια μέθοδος που προσαρμόζει το ποσοστό της εξομάλυνσης στην αρχική σ.π.π. των δεδομένων. Η εκτιμήτρια αυτή κατασκευάζεται με παρόμοιο τρόπο με την εκτιμήτρια με την μέθοδο του πυρήνα αλλά η διαφορά είναι ότι η παράμετρος που καθορίζει το σχήμα της καμπύλης μπορεί να διαφέρει από το ένα σημείο στο άλλο. Για την κατασκευή της, λοιπόν, θέτουμε $K$ να είναι η συνάρτηση πυρήνα, $k$ θετικός ακέραιος και ορίζουμε τα $d_{j,k}$ να είναι η απόσταση των $X_j$ από το $k^o$- κοντινότερο σημείο συμπεριλαμβάνοντας και τα άλλα $n-1$ σημεία. Έτσι η εκτιμήτρια παίρνει την μορφή :

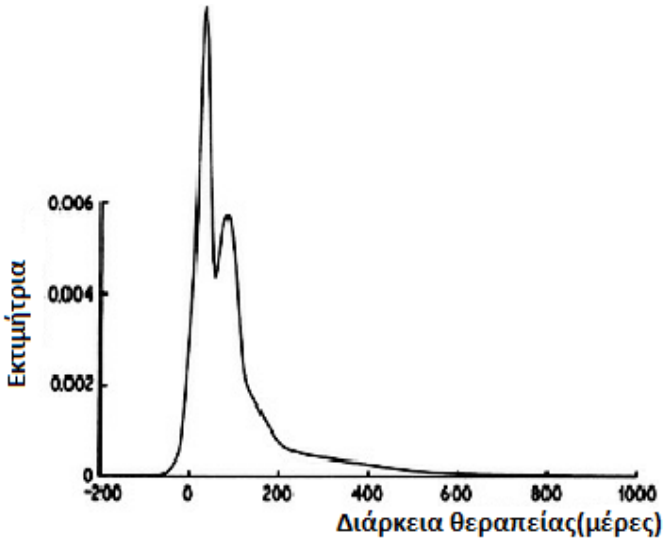
$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h d_{j,k}} K\left(\frac{t-X_j}{h d_{j,k}}\right). \qquad (2.12)$$

Το $h$ στα σημεία $X_j$ σχετίζεται με το $d_{j,k}$, έτσι οι παρατηρήσεις στις περιοχές όπου αυτές είναι αραιές, θα έχουν πιο επίπεδη συνάρτηση πυρήνα, η οποία θα σχετίζεται με αυτές. Ακόμη, για δεδομένο $k$, το ποσοστό της εξομάλυνσης θα εξαρτάται από την παράμετρο $h$. Η επιλογή του $k$ καθορίζει το πόσο καλή είναι η επιλογή του $h$ ώστε να μας δίνει περισσότερες πληροφορίες.

### 2.5.1. Σχέση εκτιμήτριας μεταβλητού πυρήνα με την εκτιμήτρια του κοντινότερου γείτονα

Το πλάτος κελιού που χρησιμοποιείται για την κατασκευή της εκτιμήτριας με την μέθοδο του κοντινότερου γείτονα στο $t$ εξαρτάται από την απόσταση του $t$ από τα σημεία των παρατηρήσεων, ενώ το πλάτος κελιού που χρησιμοποιείται για την κατασκευή της εκτιμήτριας μεταβλητού πυρήνα στο $t$ δεν εξαρτάται από το σημειο $t$, αλλά εξαρτάται μόνο από τις αποστάσεις μεταξύ των παρατηρήσεων.

Επιπλέον, σε αντίθεση με την εκτιμήτρια με την μέθοδο του κοντινότερου γείτονα, η εκτιμήτρια μεταβλητού πυρήνα είναι μια σ.π.π., με την προϋπόθεση ότι η συνάρτηση πυρήνα είναι, και έτσι όλες οι ιδιότητές της κληρονομούνται στην εκτιμήτρια.




**Γράφημα 2.7: Η εκτιμήτρια μεταβλητού πυρήνα που προκύπτει από τα δεδομένα του Πίνακα 2.2 με συνάρτηση πυρήνα $K = 8$ και πλάτος κελιού $h = 5$.**

Στο Γράφημα 2.7 παρουσιάζεται η εκτιμήτρια μεταβλητού πυρήνα η οποία προκύπτει από τα δεδομένα του Πίνακα 2.2. Ο θόρυβος στην ουρά της κατανομής έχει ελαττωθεί, αλλά πρέπει να σημειώσουμε ότι επιδεικνύεται μια δομή στο κεντρικό μέρος της κατανομής που δεν γίνεται φανερή στο Γράφημα 2.5.

## 2.6. Η εκτιμήτρια μεταβλητού πλάτους κελιού

Η εκτιμήτρια μεταβλητού πλάτους κελιού κατασκευάζεται με παρόμοιο τρόπο με την εκτιμήτρια με την μέθοδο του πυρήνα αλλά η διαφορά είναι ότι η παράμετρος που καθορίζει το πλάτος της καμπύλης μπορεί να διαφέρει από το ένα σημείο στο άλλο με σκοπό η εξομάλυνση να είναι καλύτερη και να έχουμε μια πιο επίπεδη καμπύλη, ειδικά στις ουρές των κατανομών όπου οι παρατηρήσεις είναι πιο αραιά κατανεμημένες.

Για την κατασκευή της, λοιπόν, θέτουμε $K$ να είναι η συνάρτηση πυρήνα, και $h(X_J)$ θα είναι το πλάτος του κελιού για κάθε $X_j$ παρατήρηση. Έτσι η εκτιμήτρια παίρνει την μορφή :

$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h(X_j)} K\left(\frac{t-X_j}{h(X_j)}\right). \quad (2.12)$$

### 2.6.1. Η επιλογή του πλάτους των κελιών

Η επιλογή του πλάτους των κελιών μπορεί να γίνει με τρεις τρόπους:

- Εάν θέσουμε $h(X_j) = d_{k,j}$, όπου $d_{k,j}$ είναι η απόσταση της παρατήρησης $X_j$ από το $k^o$- κοντινότερο σημείο και $k$ θετικός ακέραιος. Για δεδομένο $k$, το ποσοστό της εξομάλυνσης θα εξαρτάται από την παράμετρο $h$. Η επιλογή του $k$ καθορίζει το πόσο καλή είναι η επιλογή του $h$ ώστε να μας δίνει περισσότερες πληροφορίες.

- Εάν θέσουμε $h(X_j) = \frac{h}{\sqrt{f(X_j)}}$. Για να υπολογίσουμε, όμως, αυτήν την σχέση εργαζόμαστε ως εξής:

  i. Βρίσουμε το βέλτιστο σταθερό για όλες τις παρατηρήσεις πλάτος κελιού $h_{opt}$, με τους τρόπους που θα δείξουμε σε επόμενη παράγραφο.
  ii. Υπολογίζουμε την εκτιμήτρια $\hat{f}(t)$ με βάση το σταθερό πλάτος κελιού που έχουμε βρει.
  iii. Υπολογίζουμε τα διαφορά πλάτη κελιών $h(X_j)$ από την σχέση $h(X_j) = \frac{h}{\sqrt{f(X_j)}}$.

iv. Τέλος, βρίσκουμε την βελτιωμένη εκτιμήτρια $\hat{f}(t)$ με βάση τα πλάτη κελιών $h(X_j)$ που υπολογίσαμε για κάθε παρατήρηση.

- Υπάρχει και ένας εναλλακτικός τρόπος για τον υπολογισμό των $h(X_j)$ από την σχέση $h(X_j) = \frac{h}{\sqrt{f(X_j)}}$ ,ο οποίος είναι ο ακόλουθος

  i. Υπολογίζουμε την εκτιμήτρια $\hat{f}(t)$ με βάση κάποιο σταθερό πλάτος κελιού $h$.

  ii. Βρίσκουμε τον γεωμετρικό μέσο των εκτιμήσεων $\hat{f}(X_j)$ που δίνεται από την σχέση $G = (\prod \hat{f}(X_j))^{\frac{1}{n}}$.

  iii. Υπολογίζουμε την ποσότητα $\lambda_i = \sqrt{\frac{G}{\hat{f}(X_j)}}$.

  iv. Θέτουμε $h(X_j) = h\lambda_i$.

Στην τελευταία περίπτωση η επιλογή του $h$ στο πρώτο βήμα δεν έχει πολύ μεγάλη σημασία, αφού τα $h(X_j)$ καθορίζονται κυρίως από τα δεδομένα, έτσι η διαδικασία απαλλάσσεται από την επιλογή του βέλτιστου πλάτους κελιού $h_{opt}$.

## 2.7. Η εκτιμήτρια με την μέθοδο των ορθογώνιων σειρών

Σε αυτό το σημείο θα προσπαθήσουμε να προσεγγίσουμε μια εκτίμηση με λίγο διαφορετικό τρόπο και θα το κάνουμε με ένα παράδειγμα. Υποθέτουμε ότι έχουμε να εκτιμήσουμε μια σ.π.π. $f$ στο διάστημα [0,1]. Η κεντρική ιδέα είναι να εκτιμήσουμε την $f$ εκτιμώντας τους συντελεστές της επέκτασης Fourier.

Ορίζουμε την σειρά $\varphi_\nu(x)$ ως εξης :

$$\varphi_0(x) = 1$$
$$\varphi_{2r-1}(x) = \sqrt{2}\cos 2\pi r x, \quad \text{για r=1,2,....}$$
$$\varphi_{2r}(x) = \sqrt{2}\sin 2\pi r x.$$

Η $f$ μπορεί να γραφεί ως σειρά Fourier $\sum_{\nu=1}^{\infty} f_\nu \varphi_\nu$, όπου για κάθε ν>0 έχουμε:

$$f_\nu = \int_0^1 f(x)\varphi_\nu(x)dx. \qquad (2.13)$$

Υποθέτουμε ότι $X$ είναι μια τυχαία μεταβλητή με σ.π.π. $f$ . Έτσι η σχέση (2.13) μπορεί να πάρει την μορφή:

$$\mathrm{f_\nu} = \mathrm{E}\varphi_\nu(\mathrm{X}). \tag{2.14}$$

Οπότε μια αμερόληπτη εκτιμήτρια της $f_\nu$ είναι η ακόλουθη:

$$\widehat{f_\nu} = \frac{1}{n}\sum_{i=1}^{n} \varphi_\nu(X_i). \tag{2.15}$$

Δυστυχώς, το άθροισμα $\sum_{\nu=1}^{\infty} \widehat{f_\nu}\varphi_\nu$ δεν είναι μια καλή εκτίμηση της $f$, αλλά συγκλίνει στο άθροισμα των δέλτα συναρτήσεων των παρατηρήσεων. Αυτό προκύπτει θέτοντας:

$$w(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i), \tag{2.16}$$

όπου δ είναι η Dirac δέλτα συνάρτηση.

Έτσι για κάθε ν έχουμε :

$$\widehat{\mathrm{f_\nu}} = \int_0^1 \mathrm{w(x)}\varphi_\nu(\mathrm{x})\mathrm{dx} \tag{2.17}$$

και η $\widehat{f_\nu}$ είναι ίση με τους συντελεστές της σειράς Fourier της συνάρτησης $w(x)$.

Για να πετύχουμε μια χρήσιμη εκτιμήτρια είναι απαραίτητο να εξομαλύνουμε την $w(x)$ εφαρμόζοντας μια ακολουθία στους συντελεστές της $\widehat{f_\nu}$. Ο ευκολότερος τρόπος για να γίνει αυτό είναι να περικόψουμε την επέκταση $\sum_{\nu=1}^{\infty} \widehat{f_\nu}\varphi_\nu$ σε κάποιο σημείο. Επιλεγούμε, έτσι, έναν ακέραιο $N$ και ορίζουμε την εκτιμήτρια ως έξης:

$$\hat{\mathrm{f}}(\mathrm{x}) = \sum_{\nu=1}^{\mathrm{N}} \widehat{\mathrm{f_\nu}}\,\varphi_\nu(\mathrm{x}). \tag{2.18}$$

Η επιλογή του σημείου όπου θα περικοπεί η επέκταση καθορίζει το ποσοστό της εξομάλυνσης.

Μια άλλη προσέγγιση, πιο γενική, είναι να μικρύνουμε την σειρά με την βοήθεια μιας ακολουθίας $\lambda_\nu$, η οποία θα ικανοποιεί την σχέση:

$$\lambda_\nu \to 0 \quad \text{καθώς } \nu \to \infty,$$

ώστε να αποκτήσουμε την ακόλουθη εκτιμήτρια :

$$\hat{\mathrm{f}}(\mathrm{x}) = \sum_{\nu=\mathrm{n}}^{\infty} \lambda_\nu \widehat{\mathrm{f_\nu}}\,\varphi_\nu(\mathrm{x}). \tag{2.19}$$

Ο ρυθμός με τον οποίο η ακολουθία $\lambda_\nu$ συγκλίνει στο μηδέν καθορίζει το ποσοστό της εξομάλυνσης.

## 2.8. Η εκτιμήτρια μέγιστης ποινικοποιημένης πιθανοφάνειας

Η πιθανοφάνεια μιας καμπύλης $g$, η οποία είναι μια σ.π.π. κατασκευασμένη από ανεξάρτητες κατανεμημένες παρατηρήσεις, δίνεται από την σχέση:

$$L(g|X_1, X_2, \dots\dots, X_n) = \prod_{i=1}^{n} g(X_i). \tag{2.20}$$

Η πιθανοφάνεια δεν έχει πάντα πεπερασμένο μέγιστο για κάθε σ.π.π. και αυτό φαίνεται αν θέσουμε $\hat{f}$ να είναι ο απλοϊκός εκτιμητής με πλάτος κελιού $1/2h$. Έτσι για κάθε i έχουμε :

$$\widehat{f_h}(X_i) > \frac{1}{nh} \quad \text{και έτσι} \quad \prod \widehat{f_n}(X_i) > \frac{1}{nh} \to \infty \ \kappa\alpha\theta\acute{\omega}\varsigma\ h \to 0. \tag{2.21}$$

Η πιθανοφάνεια μπορεί να γίνει πολύ μεγάλη εάν έχουμε μια σ.π.π. η οποία προσεγγίζεται με το άθροισμα των δέλτα συναρτήσεων $w(x)$ όπως ορίστηκε στην σχέση (2.16), οπότε είναι αδύνατον να χρησιμοποιήσουμε άμεσα την εκτιμήτρια μέγιστης πιθανοφάνειας χωρίς να θέσουμε κάποιους περιορισμούς στις σ.π.π. στις οποίες θέλουμε η πιθανοφάνεια να γίνει μέγιστη.

Ωστόσο υπάρχει η δυνατότητα προσέγγισης της εκτίμησης με μια μέθοδο παρόμοια με αυτή της μέγιστης πιθανοφάνειας. Ειδικότερα, μπορούμε να ενσωματώσουμε στην πιθανοφάνεια έναν όρο ο οποίος θα περιγράφει την τραχύτητα της καμπύλης με κάποιες προϋποθέσεις. Υποθέτουμε ότι $R(g)$ είναι η συνάρτηση που μετρά την τραχύτητα της καμπύλης $g$. Μια πιθανή τέτοια συνάρτηση είναι η:

$$R(g) = \int_{-\infty}^{+\infty} (g'')^2. \tag{2.22}$$

Ορίζουμε την ποινικοποιημένη λογαριθμική πιθανοφάνεια να είναι :

$$l_a(g) = \sum_{i=1}^{n} log g(X_i) - aR(g), \tag{2.23}$$

η οποία με κάποιο τρόπο ρυθμίζει την αναλογία μεταξύ της εξομάλυνσης και του πόσο καλά ταιριάζει η καμπύλη στα δεδομένα, αφού το $\sum_{i=1}^{n} log g(X_i)$ υπολογίζει το πόσο καλά ταιριάζει η καμπύλη στα δεδομένα και το $\alpha$ καθορίζει το ποσοστό της

εξομάλυνσης. Έτσι, η συνάρτηση $\hat{f}$ θα λέμε ότι είναι μια εκτιμήτρια μέγιστης ποινικοποιημένης πιθανοφάνειας εάν μεγιστοποιεί την $l_a(g)$ για όλες τις καμπύλες $g$ που ικανοποιούν τις συνθήκες:

$$\int_{-\infty}^{+\infty} g = 1 \ , g(x) \geq 0 \ για \ όλα \ τα \ x \ , \ \ και \ \ R(g) < \infty. \qquad (2.24)$$

Τέλος, παρατηρούμε, πρώτον, ότι όσο μικρότερη είναι η τιμή της παραμέτρου $\alpha$, τόσο πιο τραχιά θα είναι η καμπύλη της εκτιμήτριας μέγιστης ποινικοποιημένης πιθανοφάνειας και, δεύτερον, ότι αυτές οι εκτιμήτριες που προκύπτουν με αυτήν την μέθοδο θα είναι σ.π.π.

## 2.9. Η εκτιμήτρια με την χρήση της γενικής συνάρτησης βάρους

Σε αυτήν την τελευταία παράγραφο θα ορίσουμε μια κατηγορία εκτιμητριών η οποία περιλαμβάνει πολλές από τις εκτιμήτριες που συζητήσαμε στις προηγούμενες παραγράφους. Υποθέτουμε ότι η $w(x, y)$ είναι μια συνάρτηση με 2 ορίσματα η οποία στις περισσότερες περιπτώσεις θα ικανοποιεί τις συνθήκες:

$$\int_{-\infty}^{+\infty} w(x, y)dy = 1 \quad και \quad w(x, y) \geq 0 \ \ για \ όλα \ τα \ x, y \ , \qquad (2.25)$$

και, επίσης, ότι είναι ορισμένη με τέτοιο τρόπο ώστε η πλειοψηφία των συναρτήσεων βάρους της σ.π.π. $w(x,)$ να βρίσκονται κοντά στο $x$. Έτσι η εκτιμήτρια που προκύπτει είναι η ακόλουθη:

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} w(X_i, t), \qquad (2.26)$$

και θα την ονομάσουμε εκτιμήτριας γενικής συνάρτησης βάρους .

Είναι φανερό ότι οι συνθήκες της σχέσης (2.25) διαβεβαιώνουν ότι η $\hat{f}$ είναι μια σ.π.π.. Αυτή η κατηγορία εκτιμητριών είναι χρήσιμη για δυο λογούς. Πρώτον, γιατί είναι μια ενοποιημένη έννοια η οποία δίνει την δυνατότητα να αποκτήσουμε θεωρητικά συμπεράσματα που μπορούν να εφαρμοστούν σε μεγάλο πλήθος εκτιμητριών, και

επιπλέον, γιατί δίνει την δυνατότητα να ορίσουμε εκτιμήτριες οι οποίες δεν ανήκουν σε καμιά από τις κατηγορίες που έχουμε ήδη αναφέρει.

Στην συνέχεια θα παραθέσουμε μερικές ειδικές περιπτώσεις της εκτιμήτριας της γενικής συνάρτησης βάρους.

- Για να αποκτήσουμε το ιστόγραμμα ως ειδική περίπτωση της σχέσης (2.26), θέτουμε:

$$w(x,y) = \begin{cases} \frac{1}{h(x)} & \varepsilon\alpha\nu\ \tau\alpha\ x, y\ \alpha\nu\acute{\eta}\kappa\sigma\upsilon\nu\ \sigma\tau\sigma\ \iota\delta\iota\sigma\ \kappa\varepsilon\lambda\acute{\iota} \\ 0 & \delta\iota\alpha\varphi\sigma\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}, \quad (2.27)$$

  όπου $h(x)$ είναι το πλάτος του κελιού που περιέχει το $x$.

- Για να αποκτήσουμε την εκτιμήτρια με την μέθοδο του πυρήνα , θέτουμε :

$$w(x,y) = \frac{1}{h} K\left(\frac{y-x}{h}\right). \quad (2.28)$$

- Για να αποκτήσουμε την εκτιμήτρια των ορθογωνίων σειρών, θέτουμε :

$$w(x,y) = \sum_{\nu=0}^{K} \varphi_\nu(x)\varphi_\nu(y). \quad (2.29)$$

# ΚΕΦΑΛΑΙΟ 3

## 3. Η ΕΚΤΙΜΗΤΡΙΑ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΤΟΥ ΠΥΡΗΝΑ ΓΙΑ ΜΟΝΟΜΕΤΑΒΛΗΤΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

### 3.1. 3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα δώσουμε μια αναλυτική περιγραφή της εκτιμήτριας με την μέθοδο του πυρήνα. Η ενασχόληση μας με αυτήν την μέθοδο πρώτα από όλες δεν σημαίνει ότι αυτή είναι η καλύτερη για όλες τις περιπτώσεις, απλώς την μελετάμε πρώτη για το λόγο ότι μπορεί να εφαρμοστεί σε πολλές διαδικασίες και ότι είναι αρκετά κατανοητή έτσι ώστε να μας βοηθήσει να ορίσουμε και άλλες μεθόδους.

#### 3.1.1. Συμβολισμοί και συνθήκες

Αρχικά υποθέτουμε ότι έχουμε ένα δείγμα $X_1, \dots, X_n$ από ανεξάρτητες κατανεμημένες παρατηρήσεις που ανήκουν σε μια συνεχή κατανομή με σ.π.π. f την όποια θα προσπαθήσουμε να εκτιμήσουμε. Ωστόσο, υπάρχουν κάποιες περιπτώσεις όπου αυτή μόνο η υπόθεση δεν αρκεί, και για αυτό το λόγο πρέπει να ορίσουμε ένα πλαίσιο πάνω στο οποίο θα συζητήσουμε τις ιδιότητες των εκτιμητριών. Επιπλέον, σε όλο το κεφάλαιο θα συμβολίζουμε με $\hat{f}$ την εκτιμήτρια πυρήνα, με K την συνάρτηση πυρήνα και με h το πλάτος του κελιού.

Η βασική μεθοδολογία της θεωρητικής προσέγγισης είναι να βρούμε πόσο κοντά είναι η τιμή της εκτιμήτριας $\hat{f}$ με την πραγματική συνάρτηση f. Η εκτίμηση $\hat{f}$ εξαρτάται τόσο από τα δεδομένα όσο και από την συνάρτηση πυρήνα K και από το πλάτος των κελιών h.

Τέλος, το σύμβολο $\int$ θα σημαίνει ολοκλήρωση στο διάστημα $(-\infty, +\infty)$ και τα x θα ανήκουν στο X.

### 3.1.2. Μέτρα ασυμφωνίας

Υπάρχουν διάφορα μέτρα τα όποια υπολογίζουν την ασυμφωνία μεταξύ της εκτιμήτριας $\hat{f}$ και της πραγματικής συνάρτησης f. Σε αυτήν την παράγραφο, θα αναφέρουμε τα δύο πιο διαδεδομένα.

1. Μέσο τετραγωνικό σφάλμα (Mean square error-MSE): Χρησιμοποιείται για εκτιμήσεις σε ένα σημείο και ορίζετε ως εξής :

$$MSE_x(\hat{f}) = E\left[\{\hat{f}\,(x) - f\,(x)\}^2\right], \quad x \in X. \tag{3.1}$$

Χρησιμοποιώντας τις ιδιότητες της μέσης τιμής και της διασποράς, η σχέση παίρνει την μορφή:

$$MSE_x(\hat{f}) = \{E[\hat{f}\,(x)] - f\,(x)\}^2 + var[\hat{f}\,(x)]. \tag{3.2}$$

Άρα το μέσο τετραγωνικό σφάλμα γράφεται ως άθροισμα του τετραγώνου της μεροληψίας συν την διακύμανση στο σημείο x. Παρατηρώντας όμως αυτήν την σχέση βλέπουμε ότι η μεροληψία μπορεί να μειωθεί μόνο σε βάρος της διακύμανσης, δηλαδή εάν αυτή αυξηθεί, και αντίστροφα, ρυθμίζοντας στο ποσοστό της εξομάλυνσης.

Ωστόσο, το MSE είναι ένα μέτρο το οποίο μετράει την ασυμφωνία στα σημεία x και έτσι δεν αποδίδει την συνολική ασυμφωνία. Έτσι προκύπτει η ανάγκη για ένα μέτρο που θα μπορεί να υπολογίσει την συνολική ασυμφωνία ανάμεσα στην εκτιμήτρια $\hat{f}$ και την πραγματική συνάρτηση f, το οποίο είναι το ακόλουθο.

2. Μέσο ολοκληρώσιμο τετραγωνικό σφάλμα (Mean integrated square error-MISE): Είναι το πιο διαδεδομένο μέτρο που μας δίνει την συνολική ασυμφωνία, λόγω της ολοκλήρωσης, ανάμεσα στην εκτιμήτρια $\hat{f}$ και την πραγματική συνάρτηση $f$ και μας βοηθά να προσεγγίσουμε με σχετική ακρίβεια την $\hat{f}$ ως εκτιμήτρια της $f$.

Ορίζεται ως εξής:

$$MISE(\hat{f}) = E[\int\{\hat{f}\,(x) - f\,(x)\}^2\,dx]. \tag{3.3}$$

Ωστόσο υπάρχουν και άλλα μέτρα που είναι καταλληλότερα σε κάποιες περιπτώσεις και στα οποία θα αναφερθούμε στην συνέχεια. Όμως το MISE είναι το πιο βολικό μέτρο και για αυτό το λόγο αξίζει να το μελετήσουμε πρώτο από όλα.

Καθώς οι όροι του ολοκληρώματος είναι θετικοί, το μέτρο μπορεί να πάρει την ακόλουθη μορφή:

$$MISE(\hat{f}) = \int E[\{\hat{f}\,(x) - f\,(x)\}^2]\,dx = \int MSE_x(\hat{f})\,dx$$

$$= \int \{E[\hat{f}\,(x)] - f\,(x)\}^2\,dx + \int var[\hat{f}\,(x)]dx\,, \qquad (3.4)$$

δηλαδή γράφεται ως άθροισμα του ολοκληρώματος του τετραγώνου της μεροληψίας συν το ολοκλήρωμα της διακύμανσης.

## 3.2. Στοιχειώδεις ιδιότητες

Εάν υποθέσουμε ότι $\hat{f}$ είναι η εκτιμήτρια με την χρήση της γενικής συνάρτησης βάρους της μορφής $\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} w(X_i, t)$, όπως έχει οριστεί στην Παράγραφο 2.9, οι προσεγγιστικές εκφράσεις για την μέση τιμή και την διακύμανση είναι οι ακόλουθες. Έτσι για κάθε $t$ έχουμε :

$$E[\hat{f}(t)] = \frac{1}{n}\sum_{i=1}^{n} E[w(X_i, t)] = \int w(x,t) f(x)\,dx \qquad (3.5)$$

και

$$var[\hat{f}(t)] = \frac{1}{n} var[w(X_i, t)] = \frac{1}{n}[\int w(x,t)^2 f(x)\,dx - \{\int w(x,t) f(x)\,dx\}^2], \qquad (3.6)$$

αντίστοιχα.

Αντικαθιστώντας τις εκφράσεις αυτές στις σχέσεις (3.3) και (3.4) μπορούμε να αποκτήσουμε εκφράσεις για το MSE και το MISE αντίστοιχα. Μια σημαντική ιδιότητα της σχέσης (3.5) είναι ότι η μεροληψία $E[\hat{f}(t)] - f(t)$ δεν εξαρτάται άμεσα από το μέγεθος του δείγματος, αλλά εξαρτάται μόνο από την συνάρτηση βάρους. Αυτό δείχνει ότι όσο μεγάλα δείγματα και να πάρουμε δεν θα μειωθεί η μεροληψία.

### 3.2.1. Εφαρμογή της εκτιμήτριας με την μέθοδο του πυρήνα

Εάν υποθέσουμε ότι η συνάρτηση βάρους είναι της μορφής :

$$w(x,y) = \frac{1}{h}K\left(\frac{x-y}{h}\right), \tag{3.7}$$

μπορούμε να αποκτήσουμε εκφράσεις για το MSE και το MISE αντικαθιστώντας πρώτα αυτήν την σχέση στις εκφράσεις (3.5) και (3.6) για την μέση τιμή και την διακύμανση:

$$E[\hat{f}(x)] = \int \frac{1}{h}K(\frac{x-y}{h})f(y)\,dy \tag{3.8}$$

και

$$var\left[\hat{f}(x)\right] = \int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)\,dy - \left\{\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)\,dy\right\}^2, \tag{3.9}$$

και έπειτα τις εκφράσεις που προκύπτουν, δηλαδή τις (3.8) και (3.9), στις σχέσεις (3.2) και (3.4). Όμως σε μερικές περιπτώσεις οι υπολογισμοί δεν είναι εύκολοι και για το λόγο αυτό οι προσεγγίσεις δέχονται κάποιους περιορισμούς, τους οποίους θα αναφέρουμε στην επόμενη παράγραφο.

Σε αυτό το σημείο είναι χρήσιμο να ορίσουμε μια γενική σχέση που ισχύει για όλες τις εκτιμήτριες:

| Εκτιμήτρια = εξομαλυμένη εκδοχή της πραγματικής συνάρτησης + το τυχαίο σφάλμα |
|---|

όπου η εξομαλυμένη εκδοχή εξαρτάται από την συγκεκριμένη επιλογή των παραμέτρων και όχι από το μέγεθος του δείγματος.

## 3.3. Περιορισμοί

Σε αυτήν την παράγραφο θα ορίσουμε κάποιες προσεγγιστικές εκφράσεις για την μεροληψία και την διακύμανση, τις οποίες θα χρησιμοποιήσουμε για να ανακαλύψουμε πως συμπεριφέρονται τα MSE και MISE. Για να το πετύχουμε αυτό θεωρούμε ότι η συνάρτηση πυρήνα K ικανοποιεί τους παρακάτω περιορισμούς:

$$\int K(t)dt = 1 \quad , \quad \int tK(t)dt = 0 \quad \text{και} \quad \int t^2K(t)dt = k_2 \neq 0,$$

και ότι η άγνωστη συνάρτηση πυκνότητας πιθανότητας f έχει συνεχή παράγωγα. Έτσι η συνάρτηση πυρήνα θα είναι μια συμμετρική σ.π.π., όπως η κανονική σ.π.π., και η σταθερά $k_2$ θα είναι η διακύμανση της εν λόγω κατανομής.

### 3.3.1. Προσεγγιστικές εκφράσεις για την μεροληψία και την διακύμανση

Όπως είναι ήδη γνωστό, η μεροληψία δεν εξαρτάται άμεσα από το μέγεθος του δείγματος, αλλά από το h (ή από την συνάρτηση βάρους). Φυσικά, εάν το h εξαρτάται από το n, τότε έμμεσα και η μεροληψία εξαρτάται από το μέγεθος του δείγματος. Έτσι ορίζουμε την σχέση:

$$\begin{aligned} bias_h(x) &= E[\hat{f}(x)] - f(x) \\ &= \int h^{-1} K\left(\tfrac{x-y}{h}\right) f(y)dy - f(x), \end{aligned} \tag{3.10}$$

από την οποία μπορούμε να αποκτήσουμε μια προσεγγιστική έκφραση για την μεροληψία.

Κάνοντας αλλαγή μεταβλητών $y = x - ht$ προκύπτει ότι:

$$\begin{aligned} bias_h(x) &= \int K(t) f(x - ht)dt - f(x) \\ &= \int K(t)\{f(x - ht) - f(x)\}dt, \end{aligned} \tag{3.11}$$

και χρησιμοποιώντας το ανάπτυγμα της σειράς Taylor,

$$f(x - ht) = f(x) - htf'(x) + \tfrac{1}{2}h^2t^2f''(x) + \ \ldots \tag{3.12}$$

προκύπτει η έκφραση για την μεροληψία:

$$\begin{aligned} bias_h(x) &= -hf'(x)\int tK(t)dt + \tfrac{1}{2}h^2f''(x)\int t^2\,K(t)dt + \cdots \\ &= \tfrac{1}{2}h^2f''(x)k_2 + \upsilon\psi\eta\lambda\acute{o}\tau\varepsilon\rho\eta\varsigma\ \sigma\varepsilon\iota\rho\acute{a}\varsigma\ \acute{o}\rho o\upsilon\varsigma\ \tau o\upsilon\ h, \end{aligned} \tag{3.13}$$

όπου οι όροι υψηλότερης σειράς του h παραλείπονται γιατί είναι αμελητέοι.

Ολοκληρώνοντας, λοιπόν, την μεροληψία προκύπτει :

$$\int bias_h(x)^2dx \approx \tfrac{1}{4}h^4k_2^2\int f''(x)^2\,dx. \tag{3.14}$$

Τώρα από τις σχέσεις (3.9) και (3.8), χρησιμοποιώντας την αντικατάσταση $y = x - ht$ και τις σχέσεις (3.12) και (3.13) προκύπτει μια προσεγγιστική έκφραση για την διακύμανση:

$$\begin{aligned} var[\hat{f}(x)] &= \frac{1}{n}\int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)\, dy - \frac{1}{n}\{f(x) + \ bias_h(x)\}^2 \\ &= \frac{1}{n}\frac{1}{h}\int f(x-ht)K(t)^2 dt - \frac{1}{n}\{f(x) + O(h^2)\}^2 \\ &= \frac{1}{n}\frac{1}{h}\int \{f(x) - htf'(x) + \cdots\}K(t)^2 dt + O(n^{-1}) \\ &= \frac{1}{n}\frac{1}{h} f(x)\int K(t)^2 dt + O(n^{-1}) \\ &\approx \frac{1}{n}\frac{1}{h} f(x)\int K(t)^2 dt. \end{aligned} \tag{3.15}$$

Ολοκληρώνοντας, λοιπόν, την (3.15), ως προς x προκύπτει η παρακάτω προσεγγιστική σχέση:

$$\int var[\hat{f}(x)]dx \approx \frac{1}{n}\frac{1}{h}\int K(t)^2 dt. \tag{3.16}$$

Υποθέτουμε τώρα ότι θέλουμε να επιλέξουμε ένα h έτσι ώστε να αποκτήσουμε MISE όσο το δυνατόν μικρότερο. Συγκρίνοντας, όμως, τις εκφράσεις (3.14) και (3.16) ως προς τις δύο συνιστώσες του MISE παρατηρούμε ότι, εάν χρησιμοποιήσουμε όσο το δυνατόν μικρότερη τιμή h για την μεροληψία, αυτή ελαχιστοποιείται ενώ αυτόματα μεγαλώνει η διακύμανση, και αντίστροφα, όσο μεγαλύτερη τιμή h χρησιμοποιήσουμε, η διακύμανση θα ελαχιστοποιηθεί ενώ η μεροληψία θα μεγαλώσει. Οπότε, προκύπτει ένα από τα θεμελιώδη προβλήματα για την προσέγγιση της εκτίμησης. Έτσι στην επόμενη παράγραφο θα προσπαθήσουμε να δώσουμε μια τιμή για το h η οποία θα είναι η βέλτιστη για την ελαχιστοποίηση του MISE.

### 3.3.2. Η βέλτιστη τιμή του h και της συνάρτησης πυρήνα

Η βέλτιστη τιμή του h που ελαχιστοποιεί την τιμή του MISE προκύπτει παραγωγίζοντας το MISE ως προς h και θέτοντάς την παράγωγο ίση με μηδέν, εάν βέβαια η δεύτερη παράγωγος του είναι θετική ώστε να αποδεικνύεται η ελαχιστοποίηση.

Έχοντας την ακόλουθη σχέση για το MISE

$$MISE = \frac{1}{4}h^4 k_2^2 \int f''(x)^2\, dx + \frac{1}{n}\frac{1}{h}\int K(t)^2 dt, \tag{3.17}$$

παραγωγίζοντας και μηδενίζοντας, προκύπτει η βέλτιστη τιμή για το h:

$$h_{opt} = {k_2}^{-\frac{2}{5}}\{\int K(t)^2 dt\}^{\frac{1}{5}}\{\int f''(x)^2 dx\}^{-\frac{1}{5}} n^{-\frac{1}{5}}. \quad (3.18)$$

Η σχέση που προκύπτει, όμως, δεν είναι ικανοποιητική καθώς παρατηρούμε ότι το h εξαρτάται από την άγνωστη συνάρτηση την οποία θέλουμε να εκτιμήσουμε. Ωστόσο προκύπτουν κάποια χρήσιμα συμπεράσματα. Πρώτον, ότι το h συγκλίνει στο 0 καθώς το μέγεθος του δείγματος αυξάνεται, αλλά με αργό ρυθμό και δεύτερον, ότι μικρότερες τιμές για το h θα είναι καταλληλότερες για πιο γρήγορες εξομαλύνσεις, καθώς ο όρος $\int f''(x)^2 dx$ επηρεάζει την διακύμανση της σ.π.π. f.

Αντικαθιστώντας την τιμή $h_{opt}$ στην σχέση (3.17), για τον υπολογισμό του MISE, προκύπτει :

$$MISE = \frac{5}{4} C(K)\{\int f''(x)^2 dx\}^{\frac{1}{5}} n^{-\frac{4}{5}}, \quad (3.19)$$

όπου

$$C(K)={k_2}^{\frac{2}{5}}\{\int K(t)^2 dt\}^{\frac{4}{5}}. \quad (3.20)$$

Αφού όλοι οι όροι είναι σταθεροί, πρέπει να επιλέξουμε μια κατάλληλη συνάρτηση πυρήνα K ώστε να αποκτήσουμε μικρή τιμή για το C(K), καθώς αυτό θα μας βοηθήσει θεωρητικά να αποκτήσουμε μικρή τιμή για το MISE, εάν βέβαια κάνουμε καλή επιλογή για το h.

Μεγάλη σημασία αυτή τη στιγμή πρέπει να δώσουμε στις συναρτήσεις πυρήνα που είναι από μόνες τους σ.π.π., καθώς διαβεβαιώνουν ότι η εκτιμήτρια θα είναι παντού θετική, εάν και σε μερικές περιπτώσεις αυτή η απαίτηση δεν είναι τόσο ισχυρή. Εάν η τιμή του $k_2$ δεν είναι ίση με την μονάδα, μπορούμε να αντικαταστήσουμε την συνάρτηση πυρήνα K με την έκφραση ${k_2}^{1/2}K({k_2}^{1/2}t)$ χωρίς να επηρεαστεί η τιμή του C(K).

Έτσι, το πρόβλημα της ελαχιστοποίησης του C(K) μεταφέρεται στο πρόβλημα της ελαχιστοποίησης του $\int K(t)^2 dt$ υπό τους περιορισμούς ότι $\int K(t)dt = 1$ και $\int t^2 K(t)dt = 1$.

Για την λύση αυτού του προβλήματος μπορούμε να αντικαταστήσουμε την συνάρτηση πυρήνα K με την σχέση:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), & -\sqrt{5} \le t \le \sqrt{5} \\ 0 \quad , & διαφορετικά \end{cases}, \tag{3.21}$$

γνωστή ως συνάρτηση Epanechnikov.

Τέλος, μπορούμε να συγκρίνουμε κάθε συμμετρική συνάρτηση πυρήνα K με την συνάρτηση Epanechnikov, ορίζοντας την απόδοση Eff(K), με την παρακάτω σχέση:

$$\begin{aligned} Eff(K) &= \{C(K_e)/C(K)\}^{\frac{5}{4}} \\ &= \frac{3}{5\sqrt{5}}\{\int t^2 K(t)dt\}^{\frac{1}{2}}\{\int K(t)^2 dt\}^{-1}. \end{aligned} \tag{3.22}$$

Ο λόγος που χρησιμοποιούμε την δύναμη $\frac{5}{4}$ είναι ότι για μεγάλα n το MISE θα είναι το ίδιο, είτε χρησιμοποιήσουμε n παρατηρήσεις και την συνάρτηση πυρήνα K είτε χρησιμοποιήσουμε neff(K) παρατηρήσεις και την συνάρτηση Epanechnikov $K_e$.

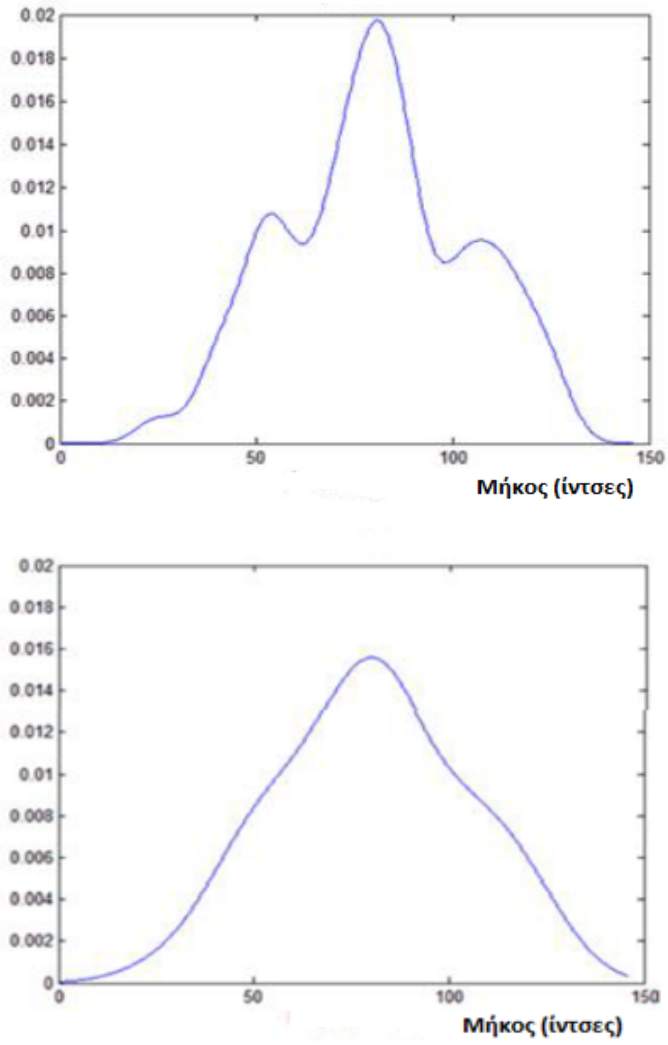
### 3.4. Επιλογή του $h$

Η σωστή επιλογή του h είναι ζωτικής σημασίας κατά την διαδικασία της εκτίμησης. Πριν αναφέρουμε αναλυτικά διάφορες μεθόδους για την καλή επιλογή του h, πρέπει να σκεφτούμε τον σκοπό για τον οποίο η εκτιμήτρια πρόκειται να χρησιμοποιηθεί. Έτσι εάν η εκτιμήτρια πρόκειται να χρησιμοποιηθεί για εξερεύνηση των δεδομένων, έτσι ώστε να προτείνει πιθανά μοντέλα ή υποθέσεις, τότε η υποκειμενική επιλογή του h, που θα περιγράψουμε στην επόμενη παράγραφο, είναι ικανοποιητική, ενώ, εάν η εκτιμήτρια πρόκειται να χρησιμοποιηθεί για παρουσίαση συμπερασμάτων, η επιλογή του h απαιτεί περαιτέρω ανάλυση.

Ωστόσο, σε πολλές περιπτώσεις απαιτείται αυτόματη επιλογή του h, κυρίως όταν η εκτιμήτρια πρόκειται να εφαρμοστεί σε μεγάλο αριθμό παρατηρήσεων ή ως ενδιάμεση διαδικασία άλλου στατιστικού προβλήματος.

### 3.4.1. Υποκειμενική επιλογή

Μια μέθοδος για να επιλέξουμε μια κατάλληλη τιμή για το h είναι να σχεδιάσουμε διάφορες καμπύλες και έπειτα να επιλέξουμε αυτήν που βρίσκεται σε καλύτερη συμφωνία με την αρχική μας ιδέα σχετικά με την εκτίμηση. Σε μερικές διαδικασίες αυτή η προσέγγιση είναι ικανοποιητική γιατί όντως μας δίνει την δυνατότητα να είμαστε πιο διορατικοί για τα δεδομένα από το να θεωρήσουμε μια αυτόματα κατασκευασμένη καμπύλη.




**Γράφημα 3.1: Εκτιμήτριες πυρήνα που έχουν προκύψει από παρατηρήσεις οι οποίες περιγράφουν το μέγεθος χιονοπτώσεων (σε ίντσες) στο Μπάφαλο της Νέας Υόρκης. Πλάτη κελιών α) $h = 5.489$, β) $h = 10.97$.**

Για παράδειγμα, στο Γράφημα 3.1, παρουσιάζονται εκτιμήσεις που έχουν προκύψει από παρατηρήσεις οι οποίες περιγράφουν το μέγεθος χιονοπτώσεων (σε ίντσες) στο Μπάφαλο της Νέας Υόρκης για 63 χειμώνες, από το 1910/11 έως το 1972/73. Από το Γράφημα αυτό φαίνεται ότι δύο διαφορετικές τιμές του h αποδίδουν δυο πιθανές επεξηγήσεις των δεδομένων. Είτε μια κανονική κατανομή, είτε μια τρικόρυφη καμπύλη που προτείνει μια μείξη από τρεις πληθυσμούς σε αναλογία 1:3:1. Η επιλογή ανάμεσα στα δύο εναλλακτικά μοντέλα είναι ένα πολύ σημαντικό βήμα.

### 3.4.2. Αναφορά στην κανονική κατανομή

Μια εύκολη προσέγγιση για να αναθέσουμε μια τιμή στον όρο $\int f''(x)^2 dx$ της έκφρασης (3.18) για το βέλτιστο h είναι να χρησιμοποιήσουμε συναρτήσεις που ανήκουν στην οικογένεια των κανονικών κατανομών. Για παράδειγμα, εάν θεωρήσουμε ότι η f είναι η σ.π.π. της κανονικής κατανομής με μέση τιμή 0 και διακύμανση $\sigma^2$ και με φ συμβολίσουμε την σ.π.π. της τυποποιημένης κανονικής κατανομής, η τιμή του ολοκληρώματος θα πάρει την μορφή:

$$\int f''(x)^2 dx = \int [f(x)\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)]^2 dx = \int [\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2}\right)]^2 dx$$

$$= \int [\varphi(y)\left(\frac{y^2}{\sigma^2} - \frac{1}{\sigma^2}\right)]^2 \frac{1}{\sigma} dy = \sigma^{-5} \int \varphi''(x)^2 dx = \sigma^{-5} \int [\varphi(y)(y^2-1)]^2 dy$$

$$= \sigma^{-5} \int \frac{1}{2\pi} e^{-y^2}(y^4 - 2y^2 + 1) dy = \sigma^{-5} \int \frac{1}{2\pi} e^{-\frac{x^2}{2}}(\frac{x^4}{4} - 2\frac{x^2}{2} + 1)\frac{1}{\sqrt{2}} dx$$

$$= \sigma^{-5} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} [\frac{1}{4} \int x^4 \varphi(x) dx - \int x^2 \varphi(x) dx + 1] = \sigma^{-5} \frac{1}{2} \pi^{-\frac{1}{2}} \left[\frac{3}{4}\right] = \frac{3}{8} \pi^{-\frac{1}{2}} \sigma^{-5}, \quad (3.23)$$

χρησιμοποιώντας τις αντικαταστάσεις $\frac{x^2}{\sigma^2} = y^2$ και $y = \frac{x}{\sqrt{2}}$ αντίστοιχα.

Εάν τώρα και η συνάρτηση πυρήνα ανήκει στην κανονική κατανομή, η σχέση (3.18) για το βέλτιστο h θα πάρει την μορφή:

$$h_{opt} = (4\pi)^{-\frac{1}{10}} (\frac{3}{8} \pi^{-\frac{1}{2}})^{-\frac{1}{5}} \sigma n^{-\frac{1}{5}} = (\frac{4}{3})^{\frac{1}{5}} \sigma n^{-\frac{1}{5}} = 1.06 \sigma n^{-\frac{1}{5}}, \quad (3.24)$$

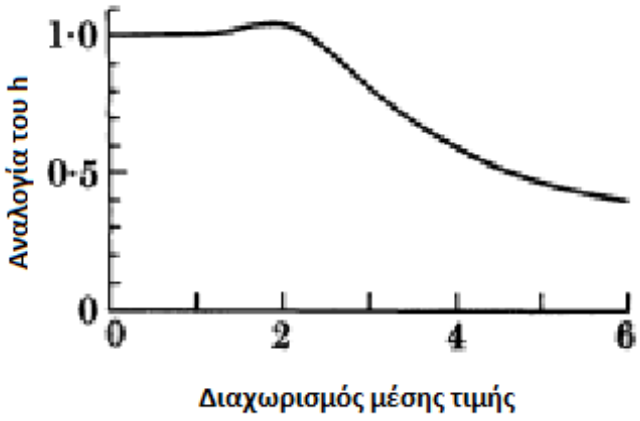
αφού έχουμε ότι:

$$\int K(t)^2 dt = \int (\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}})^2 dx = \int \frac{1}{2\pi} e^{-x^2} dx = \frac{1}{\sqrt{2\pi}} \int \frac{1}{\sqrt{2\pi}} e^{-x^2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int \varphi(x) \frac{1}{\sqrt{2}} dx = (4\pi)^{-\frac{1}{2}},$$

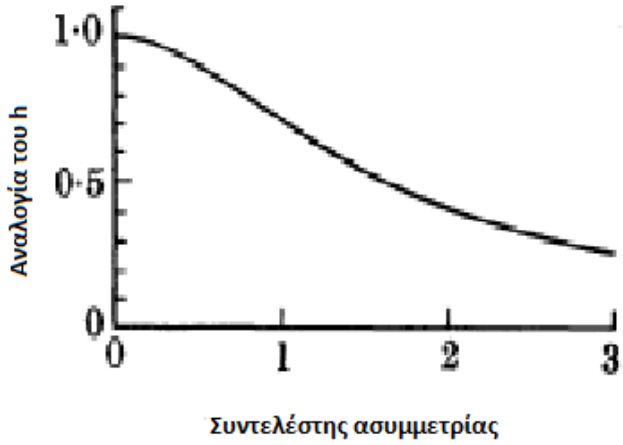
χρησιμοποιώντας την αντικατάσταση $t = \frac{x}{\sqrt{2}}$.

Μια τιμή για το $h_{opt}$ θα μπορούσε να προκύψει υπολογίζοντας το σ από τα δεδομένα και αντικαθιστώντας αυτό στη σχέση (3.23). Παρατηρούμε, όμως, ότι αυτή η σχέση δουλεύει ικανοποιητικά για πληθυσμούς που πράγματι ανήκουν στην κανονική κατανομή και όχι για πληθυσμούς πολυκόρυφων κατανομών, εξαιτίας του γεγονότος ότι η τιμή του ολοκληρώματος $\int f''(x)^2 dx$ γίνεται πολύ μεγάλη σε σχέση με την τυπική απόκλιση $\sigma^2$. Αυτή η επίδραση απεικονίζεται στο Γράφημα 3.2, το οποίο δείχνει την αναλογία του βέλτιστου πλάτους κελιού που δίνεται από την σχέση (3.18) σε σχέση με την τιμή που αποκτάται με την χρήση της σχέσης (3.24), εάν η πραγματική συνάρτηση είναι μια μείξη από δύο κανονικές κατανομές με μέσες τιμές που διαχωρίζονται.

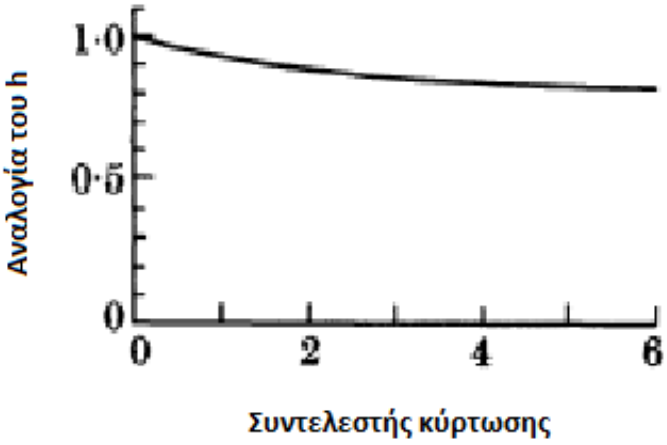



**Γράφημα 3.2: Η αναλογία του βέλτιστου πλάτους κελιού, εάν η πραγματική συνάρτηση είναι μια μείξη από δύο κανονικές κατανομές με μέσες τιμές που διαχωρίζονται**

Το γράφημα δείχνει ότι στο διάστημα του διαχωρισμού (0,2) η σχέση (3.24) δουλεύει καλά. Φυσικά η νέα σ.π.π. είναι μονοκόρυφη σε αυτό το διάστημα. Ωστόσο όσο η νέα συνάρτηση γίνεται δικόρυφη, το πλάτος του κελιού, που υπολογίζεται από την σχέση (3.24), θα εξομαλύνεται όλο και περισσότερο.




**Γράφημα 3.3: Η αναλογία του βέλτιστου πλάτους κελιού για την λογαριθμική κατανομή με δοσμένους τους συντελεστές ασυμμετρίας.**




**Γράφημα 3.4: Η αναλογία του βέλτιστου πλάτους κελιού για την λογαριθμική κατανομή με δοσμένους τους συντελεστές κύρτωσης.**

Για να εξετάσουμε, τώρα, το πόσο ευαίσθητο είναι το βέλτιστο h στην ασυμμετρία και στην κύρτωση για μονοκόρυφες κατανομές, κατασκευάζουμε τις καμπύλες που φαίνονται στα Γραφήματα 3.3 και 3.4 για την λογαριθμική και την t κατανομή σε σχέση με την αναλογία του h. Παρατηρούμε ότι για δεδομένα με μεγάλη ασυμμετρία το h της σχέσης (3.24) πάλι θα εξομαλυνθεί πολύ, ενώ δεν είναι ευαίσθητο στην κύρτωση.

Τέλος, καλύτερα αποτελέσματα μπορούν να αποκτηθούν εάν μετατρέψουμε την σχέση (3.24) χρησιμοποιώντας το ενδοτεταρτημοριακό εύρος R της κανονικής κατανομής. Έτσι το βέλτιστο h γίνεται:

$$h_{opt} = 0.79Rn^{-\frac{1}{5}}. \tag{3.25}$$

Χρησιμοποιώντας, τώρα, την έκφραση (3.25) σε κατανομές με μεγάλη ουρά ή σε ασύμμετρες κατανομές έχουμε καλύτερα αποτελέσματα. Δυστυχώς, όμως, εάν χρησιμοποιήσουμε την σχέση (3.25) σε δικόρυφες κατανομές δεν παίρνουμε καλά αποτελέσματα. Για αυτό το λόγο το καλύτερο που μπορούμε να κάνουμε είναι να χρησιμοποιήσουμε το:

$$A = \min(\tau\upsilon\pi\iota\kappa\acute{\eta}\ \alpha\pi\acute{o}\kappa\lambda\iota\sigma\eta, \epsilon\nu\delta o\tau\epsilon\tau\alpha\rho\tau o\mu o\rho\iota\alpha\kappa o\ \epsilon\upsilon\rho\omega\varsigma/1.34),$$

στην θέση του σ. Αυτό μπορεί να χρησιμοποιηθεί τόσο για μονοκόρυφες όσο και για δικόρυφες κατανομές.

Τέλος, μια διαφορετική προσέγγιση για το βέλτιστο h είναι να μειώσουμε τον συντελεστή 1.06 της σχέσης (3.24), ώστε να προκύψει η ακόλουθη τιμή για το h:

$$h = 0.9An^{-\frac{1}{5}}, \tag{3.26}$$

που θα αποδώσει ένα μέσο ολοκληρώσιμο τετραγωνικό σφάλμα μειωμένο κατά 10% .

### 3.4.3. Least-square cross-validation

Η least-square cross-validation είναι μια αυτόματη μέθοδος για την επιλογή του h. Επειδή παρουσιάζονται κάποια προβλήματα όταν αυτή η μέθοδος χρησιμοποιείται σε παρατηρήσεις που ανήκουν σε κατανομές με μεγάλη ουρά, μπορούμε να θεωρήσουμε την εναλλακτική μέθοδο που εξάγεται θεωρώντας την ελαχιστοποίηση της Kullback-Leibler απόστασης. Έτσι, η Kullback-Leibler απόσταση ορίζεται ως εξής:

$$I(p,q) = \int p(x)\log(\frac{\mathrm{p(x)}}{\mathrm{q(x)}})dx\,.$$

Για κάθε εκτιμήτρια $\hat{f}$ της συνάρτησης πυκνότητας πιθανότητας f το MISE δίνεται από την σχέση:

$$MISE = \int E[\{\hat{f}\,(x) - f\,(x)\}^2]\,dx.$$

Όμως μπορούμε να αλλάξουμε το I ώστε να αποκτήσει την ακόλουθη μορφή:

$$I(p,q) = \int \{p(x) - q(x)\}^2 dx. \quad (3.27)$$

Εάν θέλουμε, λοιπόν, να περιγράψουμε την εκτιμήτρια που βρίσκεται από την σχέση (3.27), θεωρούμε ότι έχουμε ένα δείγμα $X_1, \dots, X_n$ από ανεξάρτητες κατανεμημένες παρατηρήσεις από μια σ.π.π. f και θέτουμε $\hat{f}_{-i}$ να είναι η εκτιμήτρια που κατασκευάζεται από όλα τα δεδομένα εκτός του $X_i$. Δηλαδή:

$$\hat{f}_{-i}(x) = (n-1)^{-1} h^{-1} \sum_{j \neq i} K\{h^{-1}(x - X_j)\}. \quad (3.28)$$

Έχουμε, όμως, ότι:

$$\hat{f}(x) = \frac{1}{nh}\left[\sum_{j \neq i} K\left(\frac{x-X_j}{h}\right) + K\left(\frac{x-X_i}{h}\right)\right] = \frac{(n-1)h}{nh}\frac{1}{(n-1)h}\left[\sum_{j \neq i} K\left(\frac{x-X_j}{h}\right) + K\left(\frac{x-X_i}{h}\right)\right]$$
$$= \frac{(n-1)}{n}\hat{f}_{-i}(x) + \frac{1}{nh}K\left(\frac{x-X_i}{h}\right).$$

Άρα έχουμε ότι:

$$\hat{f}_{-i}(x) = \frac{n}{n-1}\hat{f}(x) - \frac{1}{(n-1)h}K\left(\frac{x-X_i}{h}\right).$$

και ότι:

$$\hat{f}_{-i}^{\,2}(x) = \left(\frac{n}{n-1}\right)^2 \hat{f}^2(x) + \frac{1}{(n-1)^2 h^2}K^2\left(\frac{x-X_i}{h}\right) - 2\frac{n}{n-1}\frac{1}{(n-1)h}\hat{f}(x)K\left(\frac{x-X_i}{h}\right). \quad (3.29)$$

Ο Bowman (1982) έδειξε ότι η εκτιμήτρια με την χρήση least-square cross-validation μεθόδου με βάση την Kullback-Leibler απόσταση επιτυγχάνεται χρησιμοποιώντας την τιμή του h που ελαχιστοποιεί την ακόλουθη σχέση:

$$\alpha_n(h) = n^{-1}\sum_{i=1}^{n}\int \hat{f}_{-i}^{\,2}(x)dx - 2n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i).$$

Όμως, το δεύτερο άθροισμα της συνάρτησης $\alpha_n(h)$ ισούται με:

$$n^{-1}\sum_{i=1}^{n}\int \hat{f}_{-i}^{\,2}(x)dx =$$
$$= \left(\frac{n}{n-1}\right)^2 \int \hat{f}^2(x)dx + \frac{1}{n(n-1)^2h^2}\int \sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right)dx - \frac{2}{(n-1)^2 h}\hat{f}(x)\int \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)dx$$
$$= \left[\left(\frac{n}{n-1}\right)^2 - \frac{2n}{(n-1)^2}\right]\int \hat{f}^2(x)dx + \frac{1}{n(n-1)^2h^2}\int \sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right)dx$$
$$= (1-n^{-1})^{-2}(1-2n^{-1})\int \hat{f}^2(x)dx + \frac{1}{n(n-1)^2h^2}\int \sum_{i=1}^{n} K^2\left(\frac{x-X_i}{h}\right)dx$$

$$= (1-n^{-1})^{-2}(1-2n^{-1})\int \hat{f}^2(x)dx + \frac{1}{(n-1)^2h^2}\int K^2\left(\frac{x-y}{h}\right)dx$$

$$= (1-n^{-1})^{-2}(1-2n^{-1})\int \hat{f}^2(x)dx + \frac{1}{(n-1)^2h^2}\int K^2(u)du$$

$$= (1-n^{-1})^{-2}(1-2n^{-1})\int \hat{f}^2(x)dx + \frac{1}{(n-1)^2h^2}\int K^2(x)dx$$

$$= \int \hat{f}^2(x)dx + O(\frac{1}{n^2h}), \qquad (3.30)$$

σε πιθανότητα, καθώς το $n \to \infty$. Αυτό προκύπτει από το γεγονός ότι ένας από τους όρους της επέκτασης της $\alpha_n(h)$ είναι $1/nh$ και έτσι οι υπόλοιποι όροι (3.30) είναι αμελητέοι.

Έτσι η συνάρτηση $\alpha_n(h)$ μπορεί να παρει την μορφή:

$$M_0(h) = \int \hat{f}^2(x)dx - 2n^{-1}\sum_i \hat{f}_{-i}(X_i), \qquad (3.31)$$

Για να εκφράσουμε, όμως, την συνάρτηση $M_0(h)$ σε μια μορφή η οποία θα είναι καταλληλότερη για υπολογισμούς, ορίζουμε πρώτα $K^{(2)}$ να είναι η συνέλιξη της συνάρτησης πυρήνα με τον εαυτό της. Εάν για παράδειγμα, K είναι η τυπική κανονική κατανομή, τότε $K^{(2)}$ θα είναι η κανονική κατανομή με διακύμανση 2.

Υποθέτοντας ότι η K είναι συμμετρική και χρησιμοποιώντας την αντικατάσταση $u = h^{-1}x$ έχουμε:

$$\begin{aligned}\int \hat{f}(x)^2dx &= \int \sum_i n^{-1}h^{-1}K\{h^{-1}(x-X_i)\} \\ &\quad \times \sum_j n^{-1}h^{-1}K\{h^{-1}(x-X_j)\}dx \\ &= n^{-2}h^{-1}\sum_i\sum_j \int K(h^{-1}X_i - u)K(u-h^{-1}X_j)du \\ &= n^{-2}h^{-1}\sum_i\sum_j K^{(2)}\{h^{-1}(X_i-X_j)\}. \qquad (3.32)\end{aligned}$$

Επίσης,

$$\begin{aligned}n^{-1}\sum_i \hat{f}_{-i}(X_i) &= n^{-1}\sum_i (n-1)^{-1}\sum_{j\neq i} h^{-1}K\{h^{-1}(X_i-X_j)\} \\ &= n^{-1}(n-1)^{-1}\sum_i\sum_j h^{-1}K\{h^{-1}(X_i-X_j)\} - (n-1)^{-1}h^{-1}K(0). \qquad (3.33)\end{aligned}$$

Οπότε αντικαθιστώντας αυτές τις σχέσεις στη σχέση (3.31) υπολογίζουμε το $M_0(h)$.

Ωστόσο μια παραπλήσια συνάρτηση score, η $M_1(h)$, η οποία μπορεί πολύ πιο εύκολα να υπολογιστεί, προκύπτει αλλάζοντας τον παράγοντα $(n-1)^{-1}$ της σχέσης (3.33) με τον απλούστερο, $n^{-1}$ ,και αντικαθιστώντας αυτή στην σχέση (3.31):

$$M_1(h) = n^{-2}h^{-1}\sum_i\sum_j K^*\{h^{-1}(X_i - X_j)\} + 2n^{-1}h^{-1}K(0), \qquad (3.34)$$

όπου η συνάρτηση $K^*$ είναι ίση με $K^*(t) = K^{(2)}(t) - 2K(t)$.

Περισσότερες πληροφορίες για την υπολογιστική πτυχή αυτών θα δοθεί στην παράγραφο 3.5 όπου θα δούμε ότι οι σειρές Fourier βοηθούν στο να υπολογίσουμε εύκολα το $M_1(h)$.

Έτσι έχοντας ένα δείγματα $X_1, \dots, X_n$ από μια σ.π.π. f, ορίζουμε $I_{ISXV}(X_1, \dots, X_n)$ να είναι το MISE εάν το h που έχει χρησιμοποιηθεί ελαχιστοποιεί την συνάρτηση $M_1(h)$, και ορίζουμε $I_{opt}(X_1, \dots, X_n)$ να είναι το MISE εάν το h έχει υπολογιστεί αυτόματα από το δείγμα ελαχιστοποιώντας τη σχέση $\int(\hat{f}(x) - f(x))^2 dx$ για όλα τα h, κρατώντας τα δεδομένα σταθερά.

Επίσης έχει δειχθεί ότι $\frac{I_{ISXV}}{I_{opt}} \to 1$ καθώς το $n \to \infty$ . Έτσι η least-square cross-validation μέθοδος πετυχαίνει την καλύτερη δυνατή επιλογή του h, ελαχιστοποιώντας το MISE .

### 3.4.4. Likelihood cross - validation

Η likelihood cross - validation είναι μια μέθοδος που βασίζεται στην χρήση της πιθανοφάνειας για να ορίσει την επάρκεια ενός στατιστικού μοντέλου. Η αρχή της μεθόδου, όπως αυτή εφαρμόζεται στην διαδικασία της εκτίμησης, είναι η ακόλουθη. Υποθέτουμε ότι στην διάθεσή μας έχουμε, εκτός από τα δεδομένα μας, και μια ανεξάρτητη παρατήρηση Y. Η λογαριθμική πιθανοφάνεια μιας συνάρτησης f ως μια σ.π.π. που κατασκευάστηκε από την παρατήρηση Y θα είναι η $\log f(Y)$. Θεωρώντας ότι η $\hat{f}$ ανήκει στην παραμετρική οικογένεια σ.π.π. που εξαρτάται από το h ενώ τα δεδομένα $X_1, \dots, X_n$ είναι σταθερά, η πιθανοφάνεια θα είναι $\log \hat{f}(Y)$, μια συνάρτηση ως προς h.

Εάν δεν έχουμε στην διάθεσή μας μια ανεξάρτητη παρατήρηση Y, μπορούμε να παραλείψουμε μια από τις παρατηρήσεις $X_i$ του δείγματός μας, οι οποίες χρησιμοποιήθηκαν για την κατασκευή της εκτιμήτριας, και να χρησιμοποιήσουμε αυτήν

στη θέση της παρατήρησης Y. Έτσι προκύπτει η λογαριθμική πιθανοφάνεια $\log \hat{f}_{-i}(X_i)$, όπου η $\hat{f}_{-i}(X_i)$ δίνετε από την σχέση (3.28). Αφού δεν έχει σημασία ποιά παρατήρηση θα παραληφτεί, η λογαριθμική πιθανοφάνεια ισούται με τον μέσο όρο των πιθανοφανειών, που προκύπτουν εάν παραλείπουμε κάθε φορά μια από τις παρατηρήσεις $X_i$, ώστε να αποκτήσουμε την συνάρτηση score:

$$CV(h) = n^{-1} \sum_{i=1}^{n} \log \hat{f}_{-i}(X_i). \tag{3.35}$$

Η σωστή επιλογή του h με την likelihood cross - validation μέθοδο είναι η τιμή του h που μεγιστοποιεί την συνάρτηση CV(h). Η συνάρτηση score έχει δυνατή διαισθητική φύση και δεν παρουσιάζει σοβαρές υπολογιστικές δυσκολίες. Οπότε έχουμε μια επιχειρηματολογία παραπλήσια με αυτή της Παραγράφου 3.4.3, η οποία δείχνει ότι η μεγιστοποίηση της CV(h) θα αποδώσει μια εκτιμήτρια η οποία θα είναι πολύ κοντά στην πραγματική συνάρτηση από την άποψη της Kullback-Leibler.

Θέτοντας $\hat{f}_i$ να είναι η εκτιμήτρια που βασίζεται μόνο σε $(n-1)$ παρατηρήσεις, έχουμε ότι :

$$\begin{aligned} E\{CV(h)\} &= E \log \hat{f}_{-n}(X_n) = E \int f(x) \log \hat{f}_i(x) dx \\ &\approx E \int f(x) \log \hat{f}(x) dx = -E\{I(f, \hat{f})\} + \int f(x) \log f \, dx, \end{aligned} \tag{3.36}$$

έτσι ώστε CV(h) να είναι μια αμερόληπτη εκτιμήτρια του προσδοκώμενου Kullback-Leibler σφάλματος μιας εκτίμησης που έχει χρησιμοποιήσει το ίδιο πλάτος κελιού με το δείγμα μεγέθους n-1. Δυστυχώς, όμως, αυτό το επιχείρημα ισχύει μόνο υπό αυστηρές προϋποθέσεις για την συνάρτηση f. Εάν, για παράδειγμα, η συνάρτηση f έχει απεριόριστο στήριγμα και η συνάρτηση πυρήνα K έχει περιορισμένο στήριγμα, τότε η $I(f, \hat{f})$ θα είναι $-\infty$ για κάθε h.

Σε αυτό το σημείο πρέπει να σημειώσουμε ότι η συνάρτηση CV(h) δέχεται μεγάλη επίδραση από τις παρατηρήσεις που βρίσκονται μακριά από τα υπόλοιπα δεδομένα. Για να καταλάβουμε γιατί συμβαίνει αυτό, εξετάζουμε τι θα συμβεί εάν το στήριγμα της K είναι περιορισμένο στο $(-1,1)$ και εάν μια παρατήρηση, πχ η $X_1$, είναι σε απόσταση R από τις υπόλοιπες. Έτσι, εάν $h < R$ τότε το $\hat{f}_{-1}(X_1)$ θα είναι 0 και η CV(h) θα είναι $-\infty$ για όλα τα $h < R$, ανεξαρτήτως της συμπεριφορά των υπολοίπων παρατηρήσεων. Έτσι

η τιμή που μεγιστοποιεί τη $CV(h)$ αναγκάζεται να είναι μεγαλύτερη από την R, και αυτό οδηγεί σε μεγάλη εξομάλυνση.

Ακόμη και αν αυτή η προφανής δυσκολία παρακαμφθεί, χρησιμοποιώντας μια συνάρτηση πυρήνα K με μικρότερη ουρά, είναι πάλι πιθανό η ουρά αυτή του δείγματος να ασκεί μεγάλη επιρροή, εξαιτίας της διασποράς του $\log f$ στο $-\infty$, καθώς το $|x|$ γίνεται πολύ μεγάλο.

Δεν είναι όμως μόνο οι παρατηρήσεις που βρίσκονται μακριά από τις υπόλοιπες που εμποδίζουν την διαδικασία. Για να είναι μια στατιστική διαδικασία ικανοποιητική απαιτούνται κάποιες προϋποθέσεις, όπως να υπάρχει ειρμός στα στάδια της διαδικασίας και να χρησιμοποιούνται ικανοποιητικά μεγάλου μεγέθους δείγματα.

Τώρα υποθέτουμε ότι η μία ή η άλλη ουρά της f είναι μονοτονική και σβήνει με αργό ρυθμό. Σε αυτή την περίπτωση ανήκουν όλες οι τυποποιημένες συναρτήσεις πυκνότητας πιθανότητας, εκτός από την κανονική, και όλες εκείνες που έχουν περιορισμένο στήριγμα. Η χρήση της likelihood cross - validation μεθόδου οδηγεί σε άστατες εκτιμήσεις και αυτό βασίζεται στο γεγονός ότι τα κενά ανάμεσα στις ακραίες παρατηρήσεις στις ουρές δεν μικραίνουν καθώς το μέγεθος του δείγματος μεγαλώνει. Συνεπώς το h που μεγιστοποιεί τη $CV(h)$ μπορεί να συγκλίνει στο 0 καθώς το $n \to \infty$. Οπότε καταλήγουμε στο συμπέρασμα ότι όταν η likelihood cross - validation μέθοδος εφαρμόζεται σε διακριτά δεδομένα αντιμετωπίζει τις ίδιες δυσκολίες με την least square cross-validation μέθοδο.

Η συνάρτηση score $CV(h)$ θα τείνει στο άπειρο καθώς το $h \to 0$ εάν δεν υπάρχουν απομονωμένα δεδομένα, με εξαίρεση την ειδική περίπτωση που έχουμε στη διάθεση μας πολύ μεγάλο πλήθος παρατηρήσεων. Και ακόμη εάν υπάρχουν πολλά απομονωμένα δεδομένα, αλλά κάποια από αυτά συμπίπτουν, τότε το όριο της συνάρτησης $CV(h)$ καθώς $h \to 0$ θα εξαρτάται από τις ιδιότητες της ουράς της συνάρτησης πυρήνα.

### 3.4.5. Η test graph μέθοδος

Η test graph μέθοδος είναι μια εντελώς διαφορετική μέθοδος από τις δυο cross – validation μεθόδους που αναφέραμε στις προηγούμενες παραγράφους. Σκοπός της είναι να αποδώσει μια εκτιμήτρια η οποία θα είναι πολύ κοντά στην πραγματική συνάρτηση. Μία πολύ σημαντική απαίτηση, εκτός από το να έχουμε μικρό MISE, είναι να εκτελείται η μέθοδος σε πεπερασμένο διάστημα, καθώς η σύγκλιση του $\sup\left|\hat{f}(x) - f(x)\right|$ στο μηδέν είναι ικανοποιητική, αλλά όχι απαραίτητη για την σύγκλιση του $\int(\hat{f}(x) - f(x))^2$ στο μηδέν.

Η αρχή της μεθοδολογίας βασίζεται σε ένα Θεώρημα που απέδειξε ο Silverman (1978), το οποίο δίνει το ακόλουθο αποτέλεσμα. Για την ισχύ του Θεωρήματος, όμως, υποθέτουμε ότι η συνάρτηση πυρήνα K είναι συμμετρική, δυο φορές διφορούμενη και, επίσης, ότι $\int x^2 K(x)dx$ δεν είναι μηδέν. Υποθέτουμε, ακόμη, ότι η άγνωστη συνάρτηση f είναι συνεχής και ότι έχει περιορισμένη δεύτερη παράγωγο. Επιπλέον, το h έχει επιλεγεί να είναι συναρτήσει του n για να βεβαιώσουμε την γρήγορη σύγκλιση του $\sup\left|\hat{f}(x) - f(x)\right|$ στο μηδέν. Με άλλα λόγια, το h έχει επιλεγεί ώστε να ελαχιστοποιεί το μέγιστο σφάλμα της εκτιμήτριας.

Επομένως έχουμε ότι, καθώς το $n \to \infty$:

$$\frac{\sup|\hat{f}(x)'' - E[\hat{f}(x)'']|}{sup|E[\hat{f}(x)'']|} \to k, \tag{3.37}$$

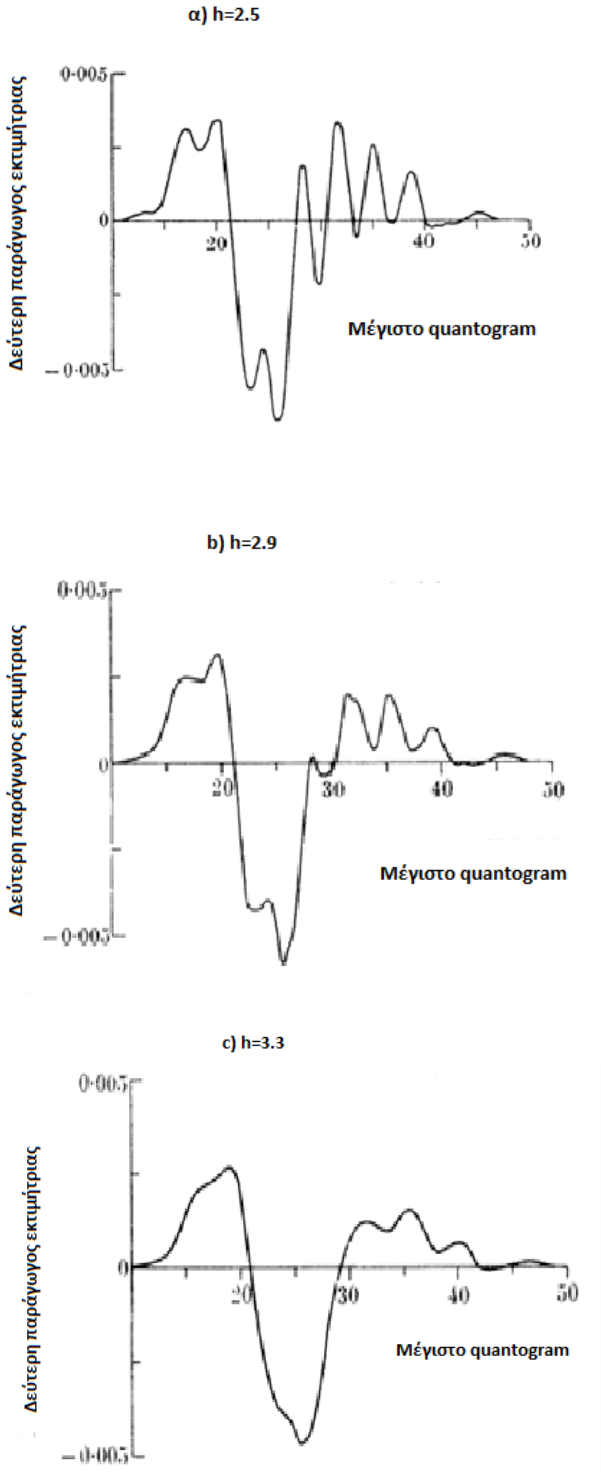
όπου η σταθερά k εξαρτάται μόνο από την συνάρτηση πυρήνα K και δίνεται από την σχέση:

$$k = \frac{1}{2}\int|x^2K(x)|\{\int(K'')^2dx/\int K^2dx\}^{\frac{1}{2}}. \tag{3.38}$$

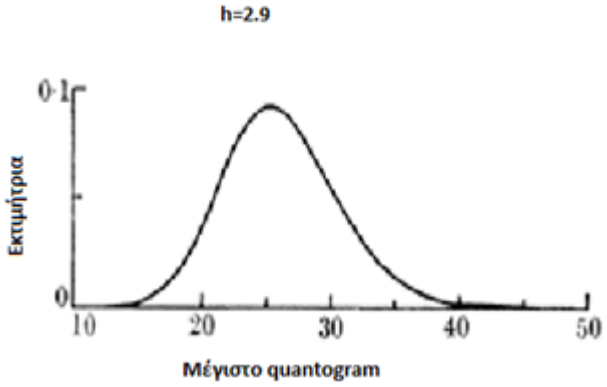
Ο όρος $\hat{f}(x)'' - E[\hat{f}(x)'']$ της σχέσης (3.37) παρουσιάζει τον τυχαίο θόρυβο της καμπύλης $\hat{f}(x)''$, ενώ ο όρος του παρανομαστή $E[\hat{f}(x)'']$ παρουσιάζει την τάση της καμπύλης. Από την σχέση (3.37) προκύπτει ότι για να έχουμε μια καλή εκτιμήτρια, οι διαστάσεις του θορύβου θα πρέπει να είναι περίπου στο μισό της τιμής της τάσης στην καμπύλη. Για μεγάλα δείγματα, ο θόρυβος θα εμφανίζεται ως γρήγορες αυξομειώσεις της καμπύλης $\hat{f}(x)''$.

Η προτεινόμενη μέθοδος για την επιλογή του h είναι η ακόλουθη. Σχεδιάζουμε την test graph της δεύτερης παραγώγου της $\hat{f}$ για διάφορες τιμές το h. Το ιδανικό test graph θα πρέπει να έχει γρήγορες αυξομειώσεις οι οποίες θα φαίνονται αλλά δεν θα επισκιάζουν την διακύμανση. Επιλέγουμε το h το οποίο αποδίδει ένα test graph που θα ταιριάζει στην συγκεκριμένη διαδικασία και θα χρησιμοποιήσουμε αυτό για την εκτίμηση της συνάρτησης f.

Ένα παράδειγμα της εφαρμογής της test graph δίνεται στο Γράφημα 3.5 για 300 ανεξάρτητες προσομοιωμένες παρατηρήσεις που χρησιμοποιήθηκαν για την εύρεση της κατανομής του μεγίστου του συνημιτονικού quantogram. Τα test graph που φαίνονται είναι για τρεις διαφορετικές τιμές του πλάτους του κελιού, και παρατηρούμε ότι, καθώς το πλάτος του κελιού αυξάνεται, το test graph γίνεται ολοένα και ομαλότερη και η γρήγορη εξομάλυνση γίνεται φανερή. Εάν το πλάτος του κελιού είναι 2.5, το test graph παρουσιάζει πολύ θόρυβο, ενώ εάν το πλάτος του κελιού είναι 3.3, το test graph είναι μια εξομαλυμένη καμπύλη με λίγο ή σχεδόν καθόλου θόρυβο. Οπότε το test graph προτείνει την τιμή 2.9 για το πλάτος του κελιού που είναι κατάλληλο για την εκτίμηση, η οποία παρουσιάζεται στο Γράφημα 3.6, όπου γίνεται φανερή η ασυμμετρία της κατανομής.




**Γράφημα 3.5: Test graphs των δεδομένων για πλάτη κελιού a)$h = 2.5$, b)$h = 2.9$, c)$h = 3.3$.**




**Γράφημα 3.6: Εκτιμήτρια των δεδομένων με πλάτος κελιού $\mathbf{h = 2.9}$.**

### 3.4.6. Εσωτερική εκτίμηση της τραχύτητας

Θεωρούμε την εξίσωση (3.18) που εκφράζει την βέλτιστη τιμή για το h. Εάν ορίσουμε $\alpha(\mathrm{K}) = \mathrm{K}_2{}^{-\frac{2}{5}}\{\int \mathrm{K(t)}^2\mathrm{dt}\}^{\frac{1}{5}}$ και $\beta(\mathrm{f}) = (\int \mathrm{f''})^{-\frac{1}{5}}$ η εξίσωση (3.18) γίνεται:

$$h_{opt} = a(K)\beta(f)n^{-\frac{1}{5}}. \tag{3.39}$$

Μπορούμε να χρησιμοποιήσουμε ένα αρχικό $\mathrm{h}_0$ για να προβλέψουμε μια εκτίμηση $\hat{\beta}(\mathrm{h}_0)$ της $\beta(\mathrm{f})$. Αυτή η εκτίμηση θα μπορούσε να αντικατασταθεί στην σχέση (3.39) για να δώσει το h που χρησιμοποιήθηκε για την εκτίμηση της σ.π.π..

Η εκτίμηση του $\beta(\mathrm{f})$ θα δοθεί από την σχέση:

$$\hat{\beta}(h_0) = (\int f_o{}''^2)^{-\frac{1}{5}} = \beta(\hat{f}_0),$$

όπου $\hat{\mathrm{f}}_0$ είναι η εκτιμήτρια που κατασκευάστηκε από τα δεδομένα με $\mathrm{h}_0$. Εάν η συνάρτηση πυρήνα είναι η τυποποιημένη κανονική συνάρτηση πυκνότητας πιθανότητας τότε προκύπτει ότι:

$$\hat{\beta}(h)^{-5} = \frac{3}{8}\pi^{\frac{-1}{2}}n^{-2}h^{-5}\sum_{j=1}^{n}\sum_{k=1}^{n}\Psi\{(X_j - X_k)/h\},$$

όπου $\Psi(\mathrm{u}) = (1 - \mathrm{u}^2 + \mathrm{u}^4/12)\exp\left(-\frac{1}{4}\mathrm{u}^2\right)$.

Οπότε το πλάτος κελιού που χρησιμοποιήθηκε θα μπορούσε να είναι το $\mathrm{h}_1$, δηλαδή το:

$$h_1 = a(K)\hat{\beta}(h_0)n^{-\frac{1}{5}}. \tag{3.40}$$

Η διατύπωση αυτή βασίζεται στο μέσο τετραγωνικό σφάλμα (MSE) για εκτιμήσεις σε ένα σημείο και όχι στο μέσο ολοκληρώσιμο τετραγωνικό σφάλμα (MISE), εάν και η βασική ιδέα είναι η ίδια. Ένα μειονέκτημα αυτής της προσέγγισης είναι ότι είναι πάλι απαραίτητο να ορίσουμε ένα αρχικό $h_0$. Ωστόσο η επιλογή του $h_0$ επηρεάζει λιγότερο την εκτίμηση από την άμεση επιλογή του h.

Είναι φανερό ότι η μέθοδος θα είναι αυτοεξυπηρετική μέχρι κάποιο σημείο. Μια μεγάλη τιμή για το $h_0$ θα οδηγήσει σε πιο εξομαλυμένο $\hat{f}_0$ και συνεπώς σε μεγαλύτερη τιμή για το $h_1$.

Παρατηρώντας τη γραφική παράσταση που δίνεται στο ακόλουθο Γράφημα 3.7 γίνεται ολοφάνερη η δυνατή σχέση μεταξύ του $h_0$ και του $h_1$. Το Γράφημα έχει κατασκευαστεί από δεδομένα τα οποία έχουν μετρηθεί από έναν απομακρυσμένο δορυφόρο και τα οποία μοιράζονται σε 70 κελιά όπως φαίνεται στον Πίνακα 3.1. Ειδικότερα, παρατηρούμε ότι εάν το $h_0$ είναι αρκετά μικρό, το $h_1$ θα είναι σχεδόν ίσο με το $h_0$.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 4 | 6 | 13 | 20 | 20 | 32 | 24 | 31 | 33 | 30 | 35 | 28 |
| 34 | 29 | 20 | 27 | 15 | 19 | 11 | 10 | 9 | 15 | 9 | 8 | 15 | 9 | 8 |
| 11 | 12 | 10 | 6 | 10 | 10 | 12 | 6 | 12 | 4 | 7 | 6 | 7 | 5 | 1 |
| 3 | 5 | 0 | 5 | 3 | 3 | 2 | 2 | 5 | 3 | 3 | 0 | 2 | 4 | 0 |
| 0 | 4 | 0 | 1 | 1 | 1 | 2 | 1 | 0 | 1 | | | | | |

**Πίνακας 3.1: Ο διαχωρισμός των δεδομένων στα 70 κελιά.**

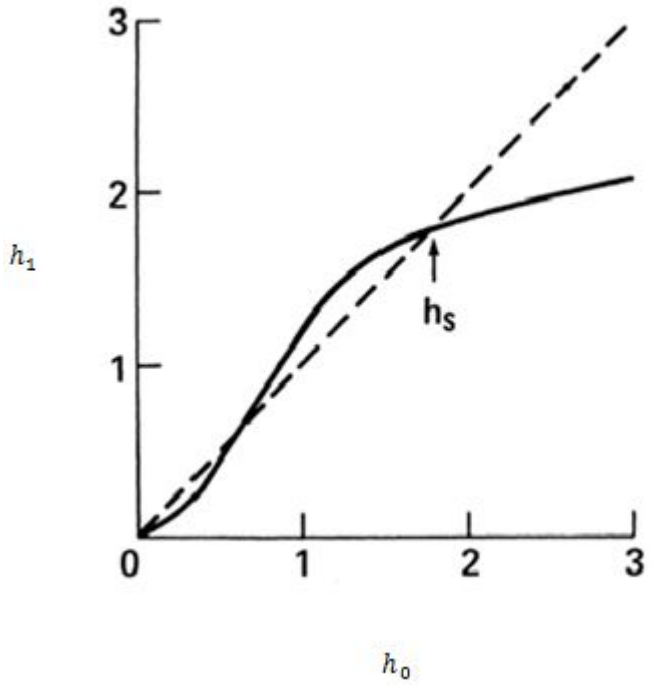
Στην προσπάθεια να αποφευχθεί το πρόβλημα της επιλογής ενός αρχικού $h_0$, ο Scott, Tapia και Thompson (1977) πρότειναν μια επαναληπτική προσέγγιση, ξεκινώντας με μια μεγάλη τιμή $h_0$, ώστε να βρίσκονται οι τιμές $h_1 h_2$ ... από τη σχέση:

$$h_i = \alpha(K)\hat{\beta}(h_{i-1})n^{-\frac{1}{5}}. \tag{3.41}$$

Η επανάληψη συνεχίζεται μέχρι να συγκλίνουν οι δυο τιμές. Αυτό το σημείο σύγκλισης ισοδυναμεί με την επιλογή του h που θα δίνει την λύση της εξίσωσης:

$$h = \alpha(K)\hat{\beta}(h)n^{-\frac{1}{5}}. \tag{3.42}$$

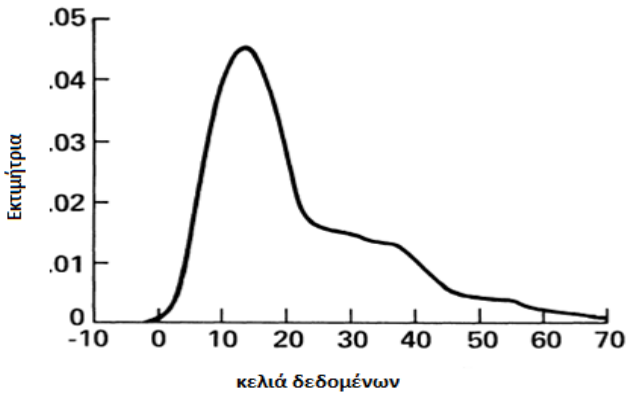
Αυτή η λύση θα ονομαστεί $h_s$.




**Γράφημα 3.7: Η σχέση ανάμεσα στο $h_0$ και $h_1$ για τα δεδομένα του Πίνακα 3.1.**

Στο Γραφήματος 3.7, φαίνεται ότι αυτή η διαδικασία οδηγεί στην επιλογή $h_s$=20. Στην πραγματικότητα, όμως, είναι πιο πρακτικό να λύσουμε την εξίσωση (3.42) με τη μέθοδο του Νεύτωνα από ότι με την μέθοδο της επανάληψης εάν και η μόνη προφανής ρίζα της θα είναι η εκφυλισμένη $h = 0$.

Η εκτίμηση των παραπάνω δεδομένων με $h = 2$ παρουσιάζεται στο Γράφημα 3.8. Είναι ενδιαφέρον να σημειώσουμε ότι η σχέση (3.26) αποδίδει μια μεγαλύτερη τιμή για το h, την $h = 3{,}3$ για αυτά τα δεδομένα. Μια πιο ξεκάθαρη εικόνα για την διαδικασία των Scott - Tapia - Thompson θα μπορούσε να δοθεί χρησιμοποιώντας την θεωρία της test - graph μεθόδου της προηγούμενης παραγράφου.




**Γράφημα 3.8: Εκτιμήτρια των δεδομένων με πλάτος κελιού $h_s$.**

Ένα καλό h για την εκτίμηση της σ.π.π. δεν θα αποδώσει απαραίτητα μια καλή εκτίμηση της δεύτερης παραγώγου κι έτσι δεν μπορούμε να προσδοκάμε ότι $\hat{\beta}(f)$ θα είναι μια καλή εκτίμηση του β(f).

Ωστόσο, έρευνες των Scott - Tapia - Thompson και των Scott - Factor έχουν δείξει ότι το $h_s$ είναι μια αρκετά καλή επιλογή για πολλά μοντέλα. Μια τελευταία έρευνα που περιλαμβάνει μια σύγκριση με την likelihood cross - validation μέθοδο, αποδεικνύει ότι αυτή η μέθοδος προσεγγίζει καλύτερα την εκτίμηση, κυρίως όμως για συναρτήσεις που έχουν μικρή ουρά. Ωστόσο, όμως, η μέθοδος Scott - Tapia - Thompson δεν επηρεάζεται από τις παρατηρήσεις που βρίσκονται μακριά, σε αντίθεση με την likelihood cross - validation μέθοδο.

Από την άλλη μεριά, ο Bowman (1985) παρουσίασε μια έρευνα στην οποία αναφέρει ότι κάποιες από τις μεθόδους που περιγράψαμε προηγουμένως εφαρμόζονται σε δείγματα μεγάλου μεγέθους που ανήκουν σε διάφορες κατανομές, συμπεριλαμβανομένων διμεταβλητών και μικρής ουράς σ.π.π.. Σε αυτή την έρευνα η Scott - Tapia - Thompson μέθοδος είναι απογοητευτική, ενώ η likelihood cross - validation μέθοδος είναι αρκετά ικανοποιητική, κυρίως για κατανομές με μικρή ουρά, ενώ δίνει ελλιπή αποτελέσματα για κατανομές με μεγάλη ουρά. Ειδικά, για μεγάλα δείγματα η least square cross-validation είναι πολύ καλύτερη.

Η εμπειρική σχέση (3.26) δίνει καλύτερα αποτελέσματα από την least square cross-validation μέθοδο για μονομεταβλητές σ.π.π., κυρίως για μικρά δείγματα.

### 3.5. Υπολογιστική εξέταση

Η πρώτη δημοσίευση άρθρου σχετικά με την εκτιμήτρια με την μέθοδο του πυρήνα έγινε το 1956, θεωρώντας την βασική ιδέα πολύ απλή. Πριν την ευρεία διάδοση των ηλεκτρονικών υπολογιστών, οι υπολογισμοί και οι γραφικές παραστάσεις των εκτιμητριών ήταν ένα αξεπέραστο πρόβλημα. Τώρα με τη βοήθεια των ηλεκτρονικών προγραμμάτων είναι πολύ πιο εύκολο να ξοδεύουμε υπερβολικό υπολογιστικό χρόνο χρησιμοποιώντας αλγορίθμους οι οποίοι μπορεί και να μην είναι και τόσο ικανοποιητικοί. Η χρήση ενός ακατάλληλου αλγορίθμου δεν είναι καταστρεπτική όταν αυτός χρησιμοποιείται για εκτιμήσεις μετρίου μεγέθους δείγματος, αλλά οι δυσκολίες μεγαλώνουν όταν πρόκειται να εφαρμοστεί σε μεγάλα δείγματα ή όταν μεγάλος αριθμός εκτιμητριών πρέπει να υπολογιστεί. Ωστόσο η χρήση τεχνικών όπως της least square cross-validation ξοδεύουν πολύ υπολογιστικό χρόνο, αφού χρειάζονται $^1/_2\,\mathrm{n}(\mathrm{n}-1)$ εκτιμήσεις της συνάρτησης $K^*$, για να υπολογίσουμε κάθε τιμή της συνάρτησης score $M_1(h)$, όπως φαίνεται από την σχέση (3.34).

Η χρήση της εκτιμήτριας με την μέθοδο του πυρήνα, όπως αυτή ορίστηκε στο δεύτερο Κεφάλαιο, δεν είναι ικανοποιητική. Είναι πολύ πιο εύκολο να θεωρήσουμε ότι η εκτιμήτρια με την μέθοδο του πυρήνα είναι η συνέλιξη των δεδομένων με την συνάρτηση πυρήνα και να χρησιμοποιήσουμε τον μετασχηματισμό Fourier για να παρουσιάσουμε αυτήν την συνέλιξη. Η χρήση του γρήγορου μετασχηματισμού Fourier δίνει τη δυνατότητα για άμεσους και αντίστροφους μετασχηματισμούς Fourier πολύ γρήγορα. Έτσι, θα δούμε ότι η least square cross-validation συνάρτηση score μπορεί να βρεθεί γρήγορα με την χρήση μετασχηματισμών Fourier.

Δοσμένης κάθε συνάρτησης g, ορίζουμε $\tilde{g}$ να είναι ο μετασχηματισμός Fourier:

$$\tilde{g}(s) = (2\pi)^{-\frac{1}{2}} \int e^{ist} g(t) dt.$$

Ορίζουμε $u(s)$ να είναι ο μετασχηματισμός Fourier από τα δεδομένα:

$$u(s) = (2\pi)^{-\frac{1}{2}} n^{-1} \sum_{j=1}^{n} \exp(isX_j),$$

και $\widetilde{f_n}(s)$ να είναι ο μετασχηματισμός Fourrier της εκτιμήτριας πυρήνα:

$$\tilde{f}_n(s) = (2\pi)^{\frac{1}{2}}\tilde{K}(hs)u(s), \tag{3.43}$$

χρησιμοποιώντας την τυπική σχέση της συνέλιξης για τους μετασχηματισμούς Fourier και την ιδιότητα ότι ο μετασχηματισμός Fourier της συνάρτησης $h^{-1}K(h^{-1}t)$ είναι $\widetilde{K}(hs)$. Η σχέση (3.43) γίνεται πολύ χρήσιμη όταν K είναι η κανονική συνάρτηση πυρήνα. Σε αυτήν την περίπτωση ο μετασχηματισμός Fourier της $\tilde{f}_n$ παίρνει την μορφή:

$$\tilde{f}_n(s) = \exp\left(-\frac{1}{2}h^2s^2\right)u(s). \tag{3.44}$$

Η βασική ιδέα του αλγορίθμου που θα αναπτύξουμε σε αυτήν την παράγραφο είναι να χρησιμοποιήσουμε τον γρήγορο μετασχηματισμό Fourrier για να βρούμε την συνάρτηση $u(s)$ και να αντιστρέψουμε την $\widetilde{f_n}(s)$ ώστε να βρούμε την εκτίμηση $\hat{f}$.

Είναι, επίσης, πολύ εύκολο να βρούμε τη least square cross-validation συνάρτηση score $M_1(h)$ από τον μετασχηματισμό Fourrier. Ορίζουμε :

$$\begin{aligned} \upsilon(s) &= (2\pi)^{-\frac{1}{2}}n^{-2}\sum\sum exp\{is(X_j - X_k)\} \\ &= (2\pi)^{-\frac{1}{2}}|u(s)|^2. \end{aligned} \tag{3.45}$$

Ο μετασχηματισμός Fourier της συνάρτησης $K^*$, η οποία έχει οριστεί στην Παράγραφο 3.4.3. είναι:

$$\begin{aligned} \widetilde{K}^*(s) &= \widetilde{K}^{(2)}(s) - 2\widetilde{K}(s) \\ &= (2\pi)^{\frac{1}{2}}\widetilde{K}(s)^2 - 2\widetilde{K}(s) && (3.46) \\ &= (2\pi)^{-\frac{1}{2}}\{\exp(-s^2) - 2\exp\left(-\frac{1}{2}s^2\right)\}, && (3.47) \end{aligned}$$

για την ειδική περίπτωση της κανονικής συνάρτησης πυρήνα.

Ορίζουμε, επίσης, την συνάρτηση:

$$\psi(t) = n^{-2}\sum_i\sum_j h^{-1}K^*\{(X_i - X_j)h^{-1} - t\}. \tag{3.48}$$

Έτσι, η least square cross-validation συνάρτηση score γίνεται:

$$M_1(h) = \psi(0) + 2n^{-1}h^{-1}K(0). \tag{3.49}$$

Έτσι έχουμε:

$$\tilde{\psi}(s) = (2\pi)^{\frac{1}{2}}\widetilde{K}^*(hs)\upsilon(s) = 2\pi\widetilde{K}^*(hs)|u(s)|^2$$

και

$$\psi(0) = (2\pi)^{-\frac{1}{2}}\int \tilde{\psi}(s)\,ds = (2\pi)^{\frac{1}{2}}\int \widetilde{K}^*(hs)|u(s)|^2\,ds$$
$$= \int\{\exp(-h^2s^2) - 2\exp\,(-\tfrac{1}{2}h^2s^2)\}\,|u(s)|^2 ds, \qquad (3.50)$$

εάν χρησιμοποιήσουμε την κανονική συνάρτηση πυρήνα. Αντικαθιστώντας την σχέση (3.50) στην σχέση (3.49) προκύπτει η least square cross-validation συνάρτηση score.

Αφού ο γρήγορος μετασχηματισμός Fourier δίνει έναν διακριτό μετασχηματισμό Fourier μιας ακολουθίας αντί για ένα μετασχηματισμό Fourier μιας συνάρτησης, είναι απαραίτητο να κάνουμε κάποιες ρυθμίσεις στην διαδικασία. Θεωρούμε ένα διάστημα $[\alpha, b]$ στο οποίο ανήκουν όλα τα δεδομένα. Ο σκοπός αυτής της μεθόδου είναι να θέσουμε περιοδικές συνοριακές συνθήκες, ορίζοντας τα τελικά σημεία α και b, έτσι ώστε το διάστημα που θα έχουμε επιλέξει να είναι αρκετά μεγάλο ώστε να μην παρουσιαστούν άλλες δυσκολίες. Δηλαδή, θέτουμε $\alpha < \min(X_i) - 3h$ και $b > \max(X_i) + 3h$ ώστε να έχουμε μια επαρκή συνθήκη για αυτό το σκοπό, ειδικά αν χρησιμοποιηθεί η κανονική συνάρτηση πυρήνα.

Διαλέγουμε $M = 2^r$ για κάποιο ακέραιο r και οριζουμε:

$$\delta = (b - \alpha)/M$$
$$t_k = \alpha + k\delta \qquad \gamma\iota\alpha\; k = 0{,}1 \ldots, M - 1,$$

και προσπαθούμε να κάνουμε τα δεδομένα διακριτά ως εξής: Εάν μια παρατήρηση X ανήκει στο διάστημα $[t_k, t_{k+1}]$, αυτό χωρίζεται σε ένα βάρος $n^{-1}\delta^{-2}(t_{k-1} - X)$ για $t_k$ και σε ένα βάρος $n^{-1}\delta^{-2}(X - t_k)$ για $t_{k+1}$. Αυτά τα βάρη συσσωρεύονται για όλα τα $X_i$ ώστε να δώσουν μια ακολουθία βαρών $(\xi_k)$ των οποίων η πρόσθεση δίνει $\delta^{-1}$. Τώρα, για $-\frac{1}{2}M \le l \le \frac{1}{2}M$, ορίζουμε $Y_l$ να είναι ο διακριτός μετασχηματισμός Fourier :

$$Y_l = M^{-1}\sum_{k=0}^{M-1} \xi_k\, exp\{i2\pi kl/M\},$$

ο οποίος βρίσκεται από τον γρήγορο μετασχηματισμό Fourier.

Ορίζουμε, επίσης,:

$$s_l = 2\pi l(b-\alpha)^{-1},$$

και υποθέτουμε ότι το $\alpha = 0$. Έπειτα, χρησιμοποιώντας τον ορισμό της ακολουθίας $\xi_k$, έχουμε :

$$\begin{aligned} Y_l &= M^{-1}\sum_{k=0}^{M-1} \xi_k \, exp\{it_k s_l\} \\ &\approx n^{-1}M^{-1}\delta^{-1}\textstyle\sum_j \exp\,(isX_j) \qquad (3.51) \\ &= (2\pi)^{\frac{1}{2}}(b-\alpha)^{-1}u(s_l). \qquad (3.52) \end{aligned}$$

Η προσέγγιση (3.52) δεν θα είναι ικανοποιητική καθώς το $|s_l|$ θα αυξάνεται, αλλά καθώς το επόμενο βήμα του αλγορίθμου, πολλαπλασιάζει όλα τα $Y_l$ για μεγάλο $|l|$ με έναν πολύ μικρό παράγοντα, στην πράξη δεν επηρεάζει.

Ορίζω μια ακολουθία $\zeta_l^*$ να είναι :

$$\zeta_l^* = \exp\left(-\frac{1}{2}h^2 {s_l}^2\right) Y_l, \qquad (3.53)$$

και θέτω $\zeta_k$ να είναι ο αντίστροφος μετασχηματισμός Fourier του $\zeta_l^*$, οπότε έχουμε:

$$\begin{aligned} \zeta_k &= \textstyle\sum_{l=-M/2}^{M/2} exp\{-i2\pi kl/M\}\zeta_l^* \\ &\approx \textstyle\sum_l \exp(is_l t_k)\,(2\pi)^{\frac{1}{2}}(b-\alpha)^{-1}\exp\left(-\frac{1}{2}h^2 {s_l}^2\right)u(s_l) \\ &\approx (2\pi)^{-\frac{1}{2}}\textstyle\int \exp(-ist_k)\exp\left(-\frac{1}{2}h^2 s^2\right)u(s) \qquad (3.54) \\ &= \hat{f}(t_k), \end{aligned}$$

καθώς (3.54) είναι ο αντίστροφος μετασχηματισμός Fourier της $\hat{f}$ όπως ορίστηκε στην σχέση (3.43).

Η περίπτωση του γενικού α απαιτεί άλγεβρα η οποία είναι πιο περίπλοκη αλλά το τελικό αποτέλεσμα (3.53) είναι ακριβώς το ίδιο. Η λεπτομερής ανάλυση των διάφορων σφαλμάτων που παράγονται από αυτόν τον αλγόριθμο δίνεται από τους Jones και Lotwick(1983).Για πολλούς σκοπούς αυτά τα σφάλματα θεωρούνται αμελητέα.

Ωστόσο, η εκτιμήτρια μπορεί να βρεθεί με τον ακόλουθο αλγόριθμο:

Βήμα 1. Διακριτοποιούμε τα δεδομένα για να βρούμε την ακολουθία βαρών $\xi_k$.

Βήμα 2. Χρησιμοποιούμε τον γρήγορο μετασχηματισμό Fourier για να βρούμε την ακολουθία $Y_l$.

Βήμα 3. Χρησιμοποιούμε την σχέση (3.53) για να βρούμε την ακολουθία $\zeta_l^*$.

Βήμα 4. Αντιστρέφουμε τον γρήγορο μετασχηματισμό Fourier για να υπολογίσουμε την ακολουθία $\hat{f}(t_k)$.

Βήμα 5. Εάν απαιτούνται εκτιμήτριες με διαφορετικά h, επαναλαμβάνουμε το βήμα 3 και το βήμα 4 μόνο.

Επιστρέφουμε πάλι στον υπολογισμό της least square cross-validation score. Προσεγγίζοντας το ολοκλήρωμα της σχέση (3.50) με ένα άθροισμα και αντικαθιστώντας στην σχέση (3.52) έχουμε:

$$\psi(0) = (b-a)\sum_{l=-M/2}^{M/2}\{\exp(-h^2 s_l{}^2) - 2\exp\left(-\tfrac{1}{2}h^2 s_l{}^2\right)\}|Y_l|^2$$

$$= -1 + 2(b-\alpha)\sum_{l=1}^{\frac{M}{2}}\left\{\exp(-h^2 s_l{}^2) - 2\exp\left(-\tfrac{1}{2}h^2 s_l{}^2\right)\right\}|Y_l|^2, \qquad (3.55)$$

καθώς $Y_0 = M^{-1}\sum \xi_k = M^{-1}\delta^{-1} = (b-\alpha)^{-1}$ και $|Y_l| = |Y_{-l}|$ για όλα τα l.

Αντικαθιστώντας, τώρα, την (3.55) στην (3.49) έχουμε:

$$\tfrac{1}{2}\{1 + M_1(h)\} = (b-\alpha)\sum_{l=1}^{\frac{M}{2}}\exp(-h^2 s_l{}^2) - \ 2\exp\left(-\tfrac{1}{2}h^2 s_l{}^2\right)\}|Y_l|^2$$

$$+n^{-1}h^{-1}(2\pi)^{-\frac{1}{2}}, \qquad (3.56)$$

όπου αυτή η σχέση βρίσκεται εύκολα για μια σειρά τιμών του h. Για τιμές που μας ενδιαφέρουν, οι εκθετικοί όροι γίνονται γρήγορα αμελητέοι και έτσι το άθροισμα θα έχει λιγότερους από $\frac{1}{2}M$ όρους.

Λίγη έρευνα έχει γίνει σε σχέση με την συνάρτηση score $M_1(h)$ και έτσι δεν μπορούν να γίνουν υποδείξεις σχετικά με την ελαχιστοποίηση του αλγορίθμου για το $M_1(h)$. Ωστόσο, στην Παράγραφο 3.4.2, αναφορές δίνουν μια ένδειξη του πιθανού διαστήματος στο οποίο θα ψάξουμε το ελάχιστο της $M_1$, το οποίο είναι :

$$\frac{1}{4}n^{-\frac{1}{5}}\sigma < h < \frac{3}{2}n^{-\frac{1}{5}}\sigma \qquad (3.57)$$

και το οποίο μπορούμε να επεκτείνουμε, εάν το ελάχιστο πέφτει στην άκρη του διαστήματος. Όμως καθώς η $M_1$ είναι μια καλή εκτίμηση του MISE και καθώς η προσεγγιστική σχέση (3.17) που δίνει το MISE είναι κυρτή στο h, θα ήταν λίγο παράξενο να μην βρούμε κυρτότητα και στο $M_1$.

### 3.6. Μια πιθανή τεχνική μείωσης της μεροληψίας

Σε αυτήν την παράγραφο θα ασχοληθούμε με την περίπτωση όπου η συνάρτηση πυρήνα είναι μια συμμετρική σ.π.π. και ικανοποιεί τις συνθήκες της Παραγράφου 3.3. Υπάρχουν γενικά επιχειρήματα που ευνοούν τη χρήση συναρτήσεων πυρήνα οι οποίες παίρνουν τόσο αρνητικές όσο και θετικές τιμές και αυτά θα συζητήσουμε στη συνέχεια.

#### 3.6.1. Ασυμπτωτικά επιχειρήματα

Υποθέτουμε ότι δεν ισχύει ο αυστηρός περιορισμός ότι η συνάρτηση πυρήνα K πρέπει να είναι μη-αρνητική και επιλέγουμε ως συνάρτηση πυρήνα μια συμμετρική συνάρτηση K η οποία ικανοποιεί τις ακόλουθες συνθήκες:

$$\int K(t)dt = 1 \quad , \quad \int t^2K(t)dt = 0 \quad , \quad \kappa\alpha\iota \quad \int t^4K(t)dt = k_4 \neq 0. \qquad (3.58)$$

Σημειώνουμε ότι οι συνθήκες αυτές δεν θα μπορούσαν να ικανοποιηθούν εάν η συνάρτηση πυρήνα K είναι μη-αρνητική, επειδή τότε θα ήταν αδύνατο το $\int \mathrm{t}^2\mathrm{K(t)dt}$ να είναι μηδέν.

Παρόμοια επιχειρήματα με αυτά της Παραγράφου 3.3.1 μπορούν να χρησιμοποιηθούν ώστε να αποκτήσουμε μια προσέγγιση για τη μεροληψία της εκτιμήτριας.

Το ανάπτυγμα Taylor της $\mathrm{f(x-ht)}$ με όρους της τάξης $\mathrm{h}^4$ είναι :

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) - \frac{1}{6}h^3t^3f'''(x) + \frac{1}{24}h^4t^4f^{iv}(x) + \cdots \qquad (3.59)$$

Αντικαθιστώντας την (3.59) στην έκφραση για την μεροληψία της σχέσης (3.14), έχουμε ένα ανάπτυγμα στο οποίο οι όροι $\mathrm{h}, \mathrm{h}^2$ και $\mathrm{h}^3$ είναι όλοι μηδέν. Ο h και ο $\mathrm{h}^3$

είναι μηδέν εξαιτίας της συμμετρίας της K, ενώ ο $h^2$ εξαιτίας της συνθήκης (3.58), η οποία λέει ότι ο συντελεστής $k_2$ είναι 0. Έτσι η μεροληψία γίνεται:

$$bias_k(x) = \frac{1}{24}h^4 f^{iv}(x)k_4 + \upsilon\psi\eta\lambda\acute{o}\tau\varepsilon\rho\eta\varsigma\ \tau\acute{a}\xi\eta\varsigma\ \acute{o}\rho o\iota\ \tau o\upsilon\ h. \qquad (3.60)$$

Η χρήση μιας συνάρτησης πυρήνα K που ικανοποιεί τις συνθήκες (3.58) μειώνει την προσέγγιση της μεροληψίας της τάξης $h^2$ στην τάξη $h^4$. Ο υπολογισμός της προσεγγιστικής έκφρασης για την διακύμανση γίνεται όπως στην Παράγραφο 3.3.1. και έτσι οι προσεγγιστικές εκφράσεις για την μεροληψία και την διακύμανση μπορούν να χρησιμοποιηθούν για να αποκτήσουμε μια προσεγγιστική έκφραση για το μέσο ολοκληρώσιμο τετραγωνικό σφάλμα:

$$\mathrm{MISE} = \frac{1}{576}\mathrm{h}^8{\mathrm{k}_4}^2 \int \mathrm{f}^{\mathrm{iv}}(\mathrm{x})^2\mathrm{dx} + \mathrm{n}^{-1}\mathrm{h}^{-1} \int \mathrm{K}(\mathrm{t})^2\mathrm{dt}. \qquad (3.61)$$

Ελαχιστοποιώντας την σχέση (3.61) ως προς h παίρνουμε μια προσέγγιση της βέλτιστης τιμής του h, η οποία θα είναι:

$$\mathrm{h_{opt}} = 72^{\frac{1}{9}}\mathrm{k_4}^{-\frac{2}{9}}\{\int \mathrm{K}(\mathrm{t})^2\mathrm{dt}\}^{\frac{1}{9}}\{\int \mathrm{f}^{\mathrm{iv}}(\mathrm{x})^2\mathrm{dx}\}^{-\frac{1}{9}}\mathrm{n}^{-\frac{1}{9}}. \qquad (3.62)$$

Τελικά, αυτή η τιμή του h εάν αντικατασταθεί στη σχέση (3.61) μας δίνει μια προσεγγιστική έκφραση του $\mathrm{MISE_{opt}}$:

$$\mathrm{MISE_{opt}} = \mathrm{C_4(K)}\{\int \mathrm{f}^{\mathrm{iv}}(\mathrm{x})^2\mathrm{dx}\}^{\frac{1}{9}}\mathrm{n}^{-\frac{8}{9}}, \qquad (3.63)$$

όπου η σταθερά $\mathrm{C_4(K)}$ δίνεται από την σχέση :

$$\mathrm{C_4(K)} = 9^{\frac{8}{9}}2^{-\frac{10}{3}}\mathrm{k_4}^{\frac{2}{9}}\{\int \mathrm{K}(\mathrm{t})^2\mathrm{dt}\}^{\frac{8}{9}}. \qquad (3.64)$$

Το βασικό μήνυμα που προκύπτει από την χρήση αυτών των ασυμπτωτικών υπολογισμών, δηλαδή από την χρήση μιας συνάρτησης πυρήνα K που ικανοποιεί τις παραπάνω συνθήκες, είναι μια μικρή βελτίωση της τάξης μεγέθους του MISE από $\mathrm{n}^{-\frac{4}{5}}$ σε $\mathrm{n}^{-\frac{8}{9}}$.

### 3.6.2. Επιλογή της συνάρτησης πυρήνα

Μια προσέγγιση για το h προκύπτει κάνοντας μια προσπάθεια να ελαχιστοποιήσουμε την ποσότητα $\mathrm{C_4(K)}$ υπό τους περιορισμούς ότι η συνάρτηση πυρήνα K θα ικανοποιεί

τις συνθήκες της σχέσης (3.58) και θα έχει ολοκλήρωμα την μονάδα. Ωστόσο αυτή η προσπάθεια δεν είχε αποτέλεσμα καθώς, η $C_4(K)$ μπορεί να γίνει αυθαίρετα μικρή ακόμη και υπό αυτούς τους περιορισμούς.

Μια διαφορετική προσέγγιση είναι να ελαχιστοποιηθεί η ασυμπτωτική διακύμανση υπό τους παραπάνω περιορισμούς και επιπλέον υπό τον περιορισμό ότι η K έχει περιορισμένο στήριγμα. Αυτό οδηγεί σε μια συνάρτηση πυρήνα K που ορίζετε από την σχέση :

$$K(y) = \begin{cases} \frac{3}{8}(3 - 5y^2) & |y| < 1 \\ 0 & \text{διαφορετικα} \end{cases}, \tag{3.65}$$

Όμως, αυτή η συνάρτηση πυρήνα K είναι ασυνεχής στο $\pm 1$ και έτσι συνεπάγεται ότι οι εκτιμήτριες θα είναι και αυτές ασυνεχείς, κάτι το οποίο απαιτεί η άγνωστη σ.π.π. f να είναι τέσσερις φορές διφορούμενη.

Ο Schucany και ο Sommers (1977) πρότειναν μια εναλλακτική μέθοδο. Δοσμένου ενός δείγματος $X_1, \dots, X_n$, θέτουμε $f_h$ να είναι η εκτιμήτρια με πλάτος κελιού h που κατασκευάστηκε από ένα δείγμα, χρησιμοποιώντας μια θετική συνάρτηση πυρήνα $K_0$ όπως την κανονική σ.π.π.. Ο όρος $O(h^2)$ στην μεροληψία της $f_h$ ελαττώνεται κατασκευάζοντας μια εκτιμήτρια:

$$\hat{f}(t) = \frac{f_h(t) - c^{-2} f_{ch}(t)}{1 - c^{-2}}, \tag{3.66}$$

δηλαδή ένα γραμμικό συνδυασμό των δύο εκτιμητριών με διαφορετικά πλάτη κελιού. Ο χρήστης μπορεί να δώσει τιμή στην σταθερά c, εάν και οι Schuman και Summers πρότειναν μια τιμή για το c κοντά στη μονάδα. Εάν η τεχνική του μετασχηματισμού Fourier χρησιμοποιηθεί για να υπολογίσουμε την $f_h$, τότε το να βρούμε μια δεύτερη εκτιμήτρια $f_{ch}$ που βασίζετε στα ίδια δεδομένα, και να αξιολογήσουμε την $\hat{f}$ από την σχέση (3.66), είναι εύκολο.

Η εκτιμήτρια $\hat{f}$ της σχέσης (3.66) μπορεί να δειχθεί ότι είναι μια εκτιμήτρια με συνάρτηση πυρήνα K που δίνεται από την σχέση:

$$K(t) = \frac{K_0(t) - c^{-3} K_0(c^{-1} t)}{1 - c^{-2}}. \tag{3.67}$$

Το όριο της (3.67) καθώς το $c \to \infty$ είναι:

$$K_1(t) = \frac{3}{2}K_0(t) + \frac{1}{2}tK_0'(t). \tag{3.68}$$

Αυτή η σχέση αποκομίζεται από ένα θεώρημα του Taylor. Εάν $K_0$ είναι η κανονική σ.π.π. φ, τότε η σχέση (3.68) γίνεται:

$$K_1(t) = \left(\frac{3}{2} - \frac{1}{2}t^2\right)\varphi(t) = \varphi(t) - \frac{1}{2}\varphi''(t), \tag{3.69}$$

και η μέθοδος με την χρήση του μετασχηματισμού Fourier μπορεί να χρησιμοποιηθεί για να βρούμε την εκτιμήτρια πυρήνα με συνάρτηση πυρήνα $K_1$.

Ο μετασχηματισμός Fourier της $K_1$ θα είναι :

$$K_1{}^*(s) = \varphi^*(s) + \frac{1}{2}s^2\varphi^*(s), \tag{3.70}$$

οπότε μπορούμε να αντικαταστήσουμε αυτήν την σχέση στην σχέση (3.44) ώστε από αποκτήσουμε τον μετασχηματισμό Fourier της εκτιμήτριας:

$$\hat{f}_n(s) = \left(1 + \frac{1}{2}h^2s^2\right)\exp\left(-\frac{1}{2}h^2s^2\right)u(s). \tag{3.71}$$

## 3.7. Ασυμπτωτικές Ιδιότητες

Σε αυτή την παράγραφο θα αναφέρουμε μερικές από τις πιο σημαντικές ασυμπτωτικές ιδιότητες των εκτιμητριών για να δώσουμε μια γενική ιδέα του ότι έχει αποδειχτεί και μερικές διαισθητικές λεπτομέρειες σχετικά με τη συμπεριφορά μεγάλου δείγματος εκτιμητριών.

Το πιο σύνηθες ασυμπτωτικό πλαίσιο, στο οποίο πολλά Θεωρήματα σχετικά με την εκτιμήτρια πυρήνα έχουν αποδειχτεί, είναι ότι συνάρτηση πυρήνα K και η άγνωστη σ.π.π. f είναι καθορισμένες και ικανοποιούν κάποιες συνθήκες. Οι εκτιμήτριες θεωρούμε ότι είναι κατασκευασμένες από τις πρώτες n παρατηρήσεις σε μια ανεξάρτητη κατανεμημένη ακολουθία $X_1, \dots, X_n$ που δίνεται από την f. Υποτίθεται ακόμη ότι το πλάτος του κελιού h εξαρτάται από το μέγεθος του δείγματος n και για να κάνουμε αυτή την εξάρτηση σαφή θα γράφουμε $h_n$ στην θέση του πλάτους του κελιού h σε αυτή την παράγραφο.

### 3.7.1. Αποτελέσματα συνοχής

Μεγάλη προσοχή πρέπει να δώσουμε στους περιορισμούς που ισχύουν για τον υπολογισμό της εκτιμήτριας με την μέθοδο του πυρήνα. Οι περιορισμοί αυτοί δεν είναι πολύ ισχυροί και για το λόγο αυτό ο ρυθμός με τον οποίο η εκτιμήτρια συγκλίνει στην πραγματική συνάρτηση είναι πολύ μικρός.

Η συνοχή της εκτιμήτριας f σε ένα σημείο x μελετήθηκε από τον Parzen, ο οποίος έθεσε τους παρακάτω περιορισμούς, δηλαδή, ότι η συνάρτηση πυρήνα K είναι μια περιορισμένη Borel συνάρτηση που ικανοποιεί τις ακόλουθες συνθήκες:

$$\int |K(t)|dt < \infty \qquad \text{και} \qquad \int K(t)dt = 1 \tag{3.72}$$

και

$$|tK(t)| \to 0 \quad \text{καθώς} \quad |t| \to \infty. \tag{3.73}$$

Αυτές οι συνθήκες ικανοποιούνται σχεδόν από κάθε συνάρτηση πυρήνα. Το πλάτος του κελιού $h_n$ υποθέτουμε ότι ικανοποιεί τις ακόλουθες σχέσεις:

$$h_n \to 0 \quad \text{και} \quad nh_n \to \infty \quad \text{καθώς } n \to \infty. \tag{3.74}$$

Υπό αυτές τις συνθήκες και με την προϋπόθεση ότι η f είναι συνεχής συνάρτηση έχουμε ότι:

$$\hat{f}(x) \to f(x) \quad \text{καθώς} \quad n \to \infty.$$

Οι συνθήκες (3.74) είναι αυτές που κυρίως απαιτούνται για την συνοχή. Αυτές συνεπάγονται ότι, ενώ το πλάτος του κελιού πρέπει να γίνεται ολοένα και μικρότερο καθώς το μέγεθος του δείγματος αυξάνεται, αυτό δεν μπορεί να συγκλίνει στο μηδέν με γρηγορότερο ρυθμό από $n^{-1}$. Δηλαδή ο προσδοκώμενος αριθμός των σημείων του δείγματος που ανήκουν στο διάστημα $x \pm h_n$ πρέπει να τείνει στο άπειρο τόσο αργά όσο το n τείνει στο άπειρο.

Εάν δεν ασχοληθούμε με την συνοχή σε ένα σημείο, είναι απαραίτητο να διευκρινίσουμε με ποιό τρόπο η καμπύλη της εκτιμήτριας $\hat{f}$ προσεγγίζει την πραγματική συνάρτηση f.

Υποθέτουμε ότι η συνάρτηση πυρήνα K είναι περιορισμένη, έχει περιορισμένη διακύμανση και ικανοποιεί τις συνθήκες (3.72) και ότι τα σημεία ασυνέχειας της K έχουν μέτρο Lebesgue μηδέν. Πάλι αυτές οι συνθήκες ικανοποιούνται από σχεδόν κάθε συνάρτηση.

Υποθέτουμε ότι:

$$\text{η f είναι μια ομοιόμορφα συνεχής συνάρτηση στο } (-\infty, +\infty), \qquad (3.75)$$

και ότι το $h_n$ ικανοποιεί τις ακόλουθες σχέσεις:

$$h_n \to 0 \quad \text{και} \quad nh_n(\log n)^{-1} \to \infty \quad \text{καθώς} \quad n \to \infty, \qquad (3.76)$$

και, επίσης, ότι έχουμε:

$$\sup\left|\hat{f}(x) - f(x)\right| \to 0 \quad \text{καθώς} \quad n \to \infty.$$

Αυτοί οι περιορισμοί είναι απαραίτητοι και εξίσου ικανοποιητικοί για ομοιόμορφη συνοχή και, επίσης, η συνθήκη (3.76) είναι ισχυρότερη από την (3.74), η οποία απαιτείται για την συνοχή σε ένα σημείο.

# ΚΕΦΑΛΑΙΟ 4

## 4. Η ΕΚΤΙΜΗΤΡΙΑ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΤΟΥ ΠΥΡΗΝΑ ΓΙΑ ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

### 4.1. Εισαγωγή

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την ανάλυση πολυμεταβλητών δεδομένων και τις εκτιμήτριες που προκύπτουν από αυτά. Ιδιαίτερη προσοχή θα δώσουμε και πάλι στην εκτιμήτρια με την μέθοδο του πυρήνα, όχι επειδή είναι η μοναδική, ή δεν υπάρχει καλύτερη, αλλά για τους λόγους που έχουμε προαναφέρει.

Στην περίπτωση των πολυμεταβλητών δεδομένων η διάκριση ανάμεσα στις ποικίλες εφαρμογές της εκτιμήτριας γίνεται πιο ισχυρή από ότι στις μονομεταβλητές περιπτώσεις. Είναι εύκολο να κατανοήσουμε ένα περίγραμμα ενός γραφήματος όταν είναι σε 2 διαστάσεις. Από την άλλη μεριά, μεγαλύτερες δυσκολίες προκύπτουν όταν παρουσιάζονται γραφήματα σε περισσότερες διαστάσεις.

Ένας έμπειρος χρηστής θα μπορούσε να αποκτήσει πολλές πληροφορίες για την εκτιμήτρια από μια απεικόνιση τριών διαστάσεων. Όμως, εάν δεν ενδιαφερόμαστε για την απεικόνιση της εκτίμησης αλλά θέλουμε να την χρησιμοποιήσουμε ως ενδιάμεση διαδικασία για άλλες στατιστικές μεθόδους, γίνεται αναγκαία προϋπόθεση η απεικόνιση σε περισσότερες διαστάσεις.

### 4.2. Η εκτιμήτρια με την μέθοδο του πυρήνα σε περισσότερες διαστάσεις

Σε αυτήν την παράγραφο θα δώσουμε μια περιγραφή της εκτιμήτριας με την μέθοδο του πυρήνα για την περίπτωση των πολυμεταβλητών δεδομένων και θα συγκρίνουμε αυτή τη μέθοδο με τα πολυμεταβλητά ιστογράμματα. Σε όλο το Κεφάλαιο θα συμβολίζουμε με έντονα γράμματα τις παρατηρήσεις $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$, στο χώρο πολλών διαστάσεων, των οποίων η συνάρτηση πρόκειται να εκτιμηθεί.

### 4.2.1. Ορισμός της εκτιμήτριας με την μέθοδο του πυρήνα για πολυμεταβλητά δεδομένα

Υποθέτουμε ότι έχουμε ένα δείγμα $\boldsymbol{X_1, \dots, X_n}$ από παρατηρήσεις d-διαστάσεων που ανήκουν στο $R^d$ των οποίων η σ.π.π. πρόκειται να εκτιμηθεί με την εκτιμήτρια πυρήνα. Ο ορισμός της εκτιμήτριας πυρήνα, ως άθροισμα των καμπυλών, μπορεί να γενικευτεί στην περίπτωση των πολυμεταβλητών δεδομένων. Η εκτιμήτρια σε αυτήν την περίπτωση ορίζεται ως εξής:

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{1}{h}(x - X_i)\right\}. \tag{4.1}$$

Η συνάρτηση πυρήνα $K$ είναι τώρα μια συνάρτηση ορισμένη σε d-διαστάσεις που ικανοποιεί την ακόλουθη συνθήκη:

$$\int_{R^d} K(x)dx = 1, \tag{4.2}$$

και συνήθως είναι μια συμμετρική μονοκόρυφη συνάρτηση πυκνότητας πιθανότητας, για παράδειγμα η τυποποιημένη πολυμεταβλητή κανονική συνάρτηση, η οποία ορίζεται ως εξής:

$$K(x) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}x^T x\right), \tag{4.3}$$

ή η πολυμεταβλητή Epanechnokov συνάρτηση πυρήνα που ορίζεται ως εξής:

$$K_e(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1 - x^T x), & \text{εάν } x^T x < 1 \\ 0 \quad , & \text{διαφορετικά} \end{cases}, \tag{4.4}$$

όπου το $c_d$ είναι ο όγκος της μοναδιαίας σφαίρας στο $R^d$.

Για την περίπτωση των 2 διαστάσεων, ακολουθούν κάποιες χρήσιμες συναρτήσεις πυρήνα:

$$K_e(x) = \begin{cases} 3\pi^{-1}(1 - x^T x)^2, & \text{εάν } x^T x < 1 \\ 0 \quad , & \text{διαφορετικά'} \end{cases} \tag{4.5}$$

$$K_e(x) = \begin{cases} 4\pi^{-1}(1 - x^T x)^3, & \text{εάν } x^T x < 1 \\ 0 \quad , & \text{διαφορετικά.} \end{cases} \tag{4.6}$$

Το πλεονέκτημα αυτών των συναρτήσεων είναι ότι έχουν υψηλότερης τάξης διαφορισιμότητα και επιπλέον, ότι μπορούν να υπολογιστούν πιο γρήγορα από την συνάρτηση πυρήνα που δίνεται από την σχέση (4.3).

Η χρήση της παραμέτρου h στη σχέση (4.1) συνεπάγεται ότι η συνάρτηση πυρήνα K που προκύπτει από κάθε παρατήρηση θα εξαπλώνεται σε όλες τις διευθύνσεις. Σε ορισμένες περιπτώσεις ίσως είναι χρήσιμο να χρησιμοποιούμε ένα διάνυσμα ή έναν πίνακα για να ορίσουμε τα πλάτη των κελιών. Μια τέτοια περίπτωση είναι όταν η μεταβλητότητα των παρατηρήσεων είναι μεγαλύτερη κατά μια διεύθυνση σε σχέση με τις υπόλοιπες.

Μόνο σε μερικές στατιστικές διαδικασίες είναι καλύτερα να εξαπλώνουμε τα δεδομένα για να αποφύγουμε σοβαρές διαφορές στη διάδοση των παρατηρήσεων ως προς τις άλλες διευθύνσεις. Εάν αυτό συμβεί, δεν θα είναι πλέον αναγκαίο να ορίσουμε πιο περίπλοκες σχέσεις για την συνάρτηση πυρήνα, από αυτή που δίνεται από τη σχέση (4.1).

Μια ελκυστική διαισθητική προσέγγιση είναι πρώτα να μετασχηματίσουμε τις παρατηρήσεις ώστε να έχουν πίνακα συνδιακύμανσης τη μονάδα, έπειτα να τις εξομαλύνουμε χρησιμοποιώντας μια ακτινωτή συμμετρική συνάρτηση πυρήνα K και στο τέλος να τις μετασχηματίσουμε πάλι. Αυτό είναι ισοδύναμο με το να χρησιμοποιήσουμε μια εκτιμήτρια της μορφής:

$$\hat{f}(x) = \frac{(\det S)^{-\frac{1}{2}}}{n h^d} \sum_{i=1}^{n} k\{h^{-2}(x - X_i)^T S^{-1}(x - X_i)\}, \qquad (4.7)$$

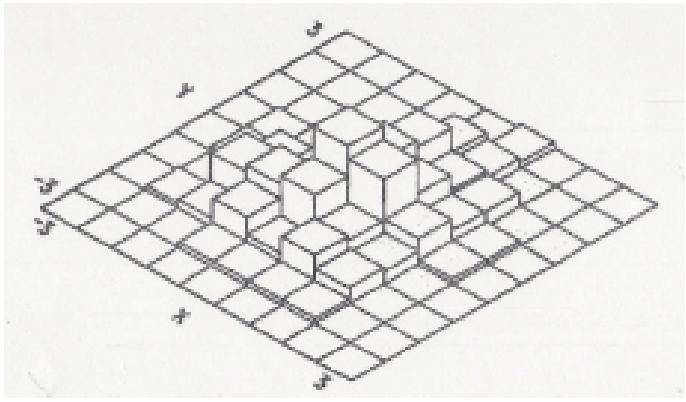
όπου η συνάρτηση k δίνεται από τη σχέση:

$$k(x^T x) = K(x),$$

και S είναι ο πίνακας συνδιακύμανσης των παρατηρήσεων. Εάν K είναι η κανονική συνάρτηση πυρήνα, τότε $k(u)$ είναι ίσο με $2\pi^{-\frac{d}{2}} \exp\left(-\frac{1}{2}u\right)$.

### 4.2.2. Ιστογράμματα πολυμεταβλητών παρατηρήσεων

Η κατασκευή ενός ιστογράμματος πολυμεταβλητών παρατηρήσεων απαιτεί τον προσδιορισμό όχι μόνο του μεγέθους των κελιών, αλλά και του προσανατολισμού τους. Στο Γράφημα 4.1 φαίνεται ένα ιστόγραμμα από διμεταβλητά δεδομένα. Κάποιος χωρίς εμπειρία είναι δύσκολο να βγάλει συμπεράσματα για τη δομή των δεδομένων εξαιτίας της ασυνέχειας που παρουσιάζουν τα ΄΄κουτιά΄΄.

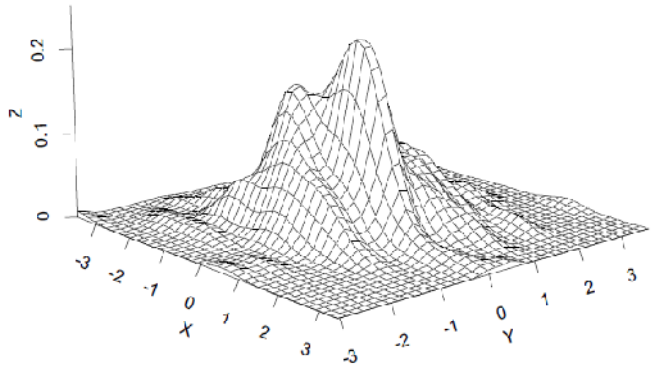


**Γράφημα 4.1: Ιστόγραμμα κατασκευασμένο από διμεταβλητές παρατηρήσεις.**

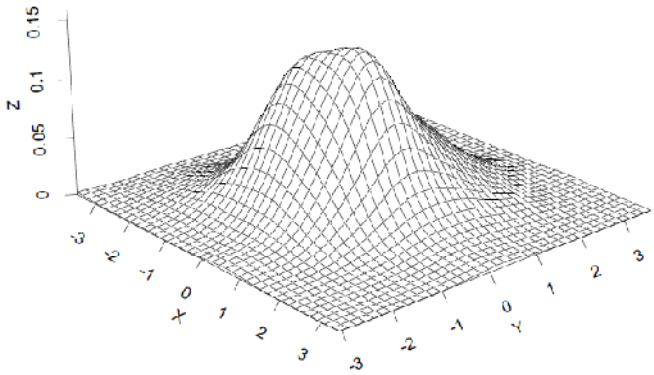
Μια επιπλέον δυσκολία με τα ιστογράμματα είναι ότι εάν το πλάτος του κελιού έχει επιλεγεί αρκετά μικρό ώστε να έχει την δυνατότητα να παρουσιάζει πληροφορίες προς όλες τις διευθύνσεις, τότε, ακόμα και στην περίπτωση των 2-διαστάσεων ο συνολικός αριθμός των κελιών γίνεται τόσο μεγάλος που θα έχει ως αποτέλεσμα τα τυχαία σφάλματα να κυριαρχήσουν. Αυτό συμβαίνει στο Γράφημα 4.1., το οποίο είναι κατασκευασμένο από 200 παρατηρήσεις. Ένα 9×9 πλέγμα κελιών που αποδίδουν σύνολο 81 κελιά, είναι ένας υπερβολικά μεγάλος αριθμός για το μέγεθος του δείγματος των 200 παρατηρήσεων. Ένας ρεαλιστικός αριθμός κελιών για την κατασκευή ενός τέτοιου ιστογράμματος θα μπορούσε να είναι 9 ή 16, αφού ένα 3×3 ή 4×4 ιστόγραμμα είναι ικανοποιητικό. Για περισσότερες από 2 διαστάσεις ο αριθμός των κελιών αυξάνεται δραματικά.

Ωστόσο, διμεταβλητές εκτιμήτριες που έχουν κατασκευαστεί χρησιμοποιώντας συνεχείς συναρτήσεις πυρήνα είναι πολύ πιο εύκολο να τις κατανοήσουμε. Για παράδειγμα στα Γραφήματα 4.2 και 4.3 που ακολουθούν, τα οποία έχουν κατασκευαστεί από διμεταβλητές παρατηρήσεις που προέρχονται από μια κανονική κατανομή, αποδίδουν

μια ξεκάθαρη εικόνα της κατανομής τους. Η δυσκολία να παρουσιάσουμε την ασυνέχεια των επιφανειών είναι προφανής, αν και κάθε υπολογιστικό λογισμικό πρόγραμμα γραφημάτων έχει την δυνατότητα απεικόνισης συνεχών επιφανειών, αλλά μόνο μερικά από αυτά μπορούν να υπολογίζουν Γραφήματα όπως το 4.1.



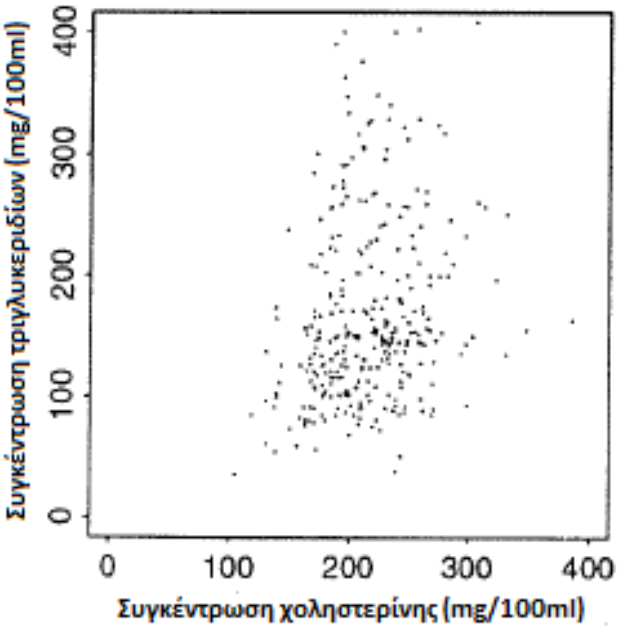

**Γράφημα 4.2: Εκτιμήτρια κατασκευασμένη από διμεταβλητές παρατηρήσεις με $h = 0.2$.**




**Γράφημα 4.3:Εκτιμήτρια κατασκευασμένη από διμεταβλητές παρατηρήσεις με $h = 0.4$.**

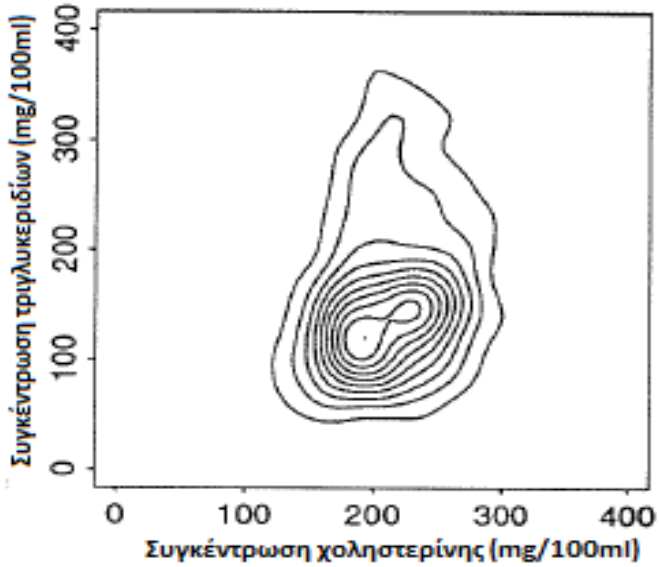
### 4.2.2. Γραφήματα Διασποράς

Ένα σημαντικό βήμα στην εξέταση των διμεταβλητών δεδομένων είναι να κατασκευάσουμε τα γραφήματα διασποράς τους. Ωστόσο, συχνά συμβαίνει να μην επισημαίνονται πολλά χαρακτηριστικά της εκτιμήτριας στα γραφήματα αυτά. Ένα παράδειγμα δίνεται στο Γράφημα 4.4. Τα δεδομένα αποτελούνται από ζευγάρια παρατηρήσεων που δείχνουν την συγκέντρωση πλάσματος λιπιδίων που πάρθηκαν από 320 ασθενείς σε μια έρευνα σχετικά με καρδιακές παθήσεις.




**Γράφημα 4.4: Γράφημα διασποράς για τα δεδομένα που δείχνουν την συγκέντρωση πλάσματος λιπιδίων.**

Το Γράφημα 4.5, που ακολουθεί, είναι ένα διάγραμμα περιγράμματος της εκτιμήτριας που έχει κατασκευαστεί χρησιμοποιώντας την συνάρτηση πυρήνα $K_3$ της σχέσης (4.6). Το πλάτος του κελιού εδώ είναι 40, αλλά η δομή είναι ίδια για διάφορα πλάτη κελιού. Ένα χαρακτηριστικό των παρατηρήσεων που γίνεται φανερό από την απεικόνιση είναι η δυο κορυφές που παρουσιάζει στο κεντρικό μέρος της η κατανομή, το οποίο είναι δύσκολο να το δει κανείς από το γράφημα διασποράς, ακόμη και αν γίνουν οι απαραίτητες διευκρινήσεις για την εκτιμήτρια.




**Γράφημα 4.5: Εκτιμήτρια πυρήνα για τα δεδομένα που δείχνουν την συγκέντρωση πλάσματος λιπιδίων.**

## 4.3. Η επιλογή της συνάρτησης πυρήνα και του πλάτους του κελιού

Αναλυτική περιγραφή της επιλογής της συνάρτησης πυρήνα $K$ και του πλάτους του κελιού $h$ έγινε στο τρίτο Κεφάλαιο. Με κάποιες κατάλληλες μετατροπές θα προσεγγίσουμε αυτές τις τιμές και για την περίπτωση των πολυμεταβλητών δεδομένων.

### 4.3.1. Ιδιότητες δειγματοληψίας

Όπως στην Παράγραφο 3.3.1, μπορούμε να αποκομίσουμε προσεγγιστικές εκφράσεις για την μεροληψία και την διακύμανση, οι οποίες μπορούν να χρησιμοποιηθούν για να βοηθήσουν στην επιλογή της συνάρτησης πυρήνα K και του πλάτους του κελιού h.

Υποθέτουμε ότι η συνάρτηση πυρήνα K είναι μια συμμετρική συνάρτηση πυκνότητας πιθανότητας και ότι η άγνωστη συνάρτηση f εχει περιορισμένες και συνεχείς δεύτερες παραγώγους.

Ορίζουμε τις σταθερές α και β ως εξης:

$$\mathrm{a} = \int \mathrm{t_1}^2 \mathrm{K(t)dt} \text{ και } \beta = \int \mathrm{K(t)}^2 \mathrm{dt}. \qquad (4.8)$$

Με την ίδια μεθόδευση όπως στο προηγούμενο Κεφάλαιο, χρησιμοποιώντας απλά το ανάπτυγμα Taylor σε πολλές διαστάσεις, αποκτούμε τις ακόλουθες προσεγγίσεις για την μεροληψία και την διακύμανση:

$$\text{bias}_h(x) \approx \frac{1}{2}h^2 a\nabla^2 f(x), \tag{4.9}$$

$$\text{var}\hat{f}(x) \approx n^{-1}h^{-d}\beta f(x). \tag{4.10}$$

Συνδυάζοντας αυτές τις σχέσεις, όπως στην Παράγραφο 3.3.2, παίρνουμε προσεγγιστική έκφραση για το μέσο ολοκληρώσιμο τετραγωνικό σφάλμα:

$$\text{MISE} = \frac{1}{4}h^4 a^2 \int\{\nabla^2 f(x)\}^2 dx + n^{-1}h^{-d}\beta. \tag{4.11}$$

Η βέλτιστη τιμή του $h$ που προκύπτει από την ελαχιστοποίηση του μέσου ολοκληρώσιμου τετραγωνικού σφάλματος, δίνεται από την σχέση:

$$h_{opt}{}^{d+4} = d\beta a^{-2}\{\int(\nabla^2 f)^2\}^{-1}n^{-1}, \tag{4.12}$$

η οποία συγκλίνει στο $0$ καθώς το $n$ αυξάνεται αλλά με πολύ αργό ρυθμό $n^{-\frac{1}{d+4}}$. Παρατηρούμε, όμως, ότι η βέλτιστη τιμή του $h$ εξαρτάται και από την άγνωστη συνάρτηση που πρόκειται να εκτιμηθεί.

Τέλος, το $h_{opt}$ μπορεί να αντικατασταθεί στην σχέση (4.11) ώστε να αποκτήσουμε το ελάχιστο μέσο ολοκληρώσιμο τετραγωνικό σφάλμα έτσι ώστε να οδηγηθούμε στην επιλογή της συνάρτησης πυρήνα $K$ .

### 4.3.2. Η επιλογή του πλάτους του κελιού για την κανονική κατανομή

Το πρώτο βήμα ώστε να επιλέξουμε το πλάτος του κελιού είναι να χρησιμοποιήσουμε την σχέση (4.12), όταν η $f$ είναι μια κανονική σ.π.π.. Εάν η $\varphi$ εναι η τυποποιημένη $d$-μεταβλητή κανονική σ.π.π., μπορεί να δειχθεί ότι:

$$\int(\nabla^2\varphi)^2 = \left(2\sqrt{\pi}\right)^{-d}\left(\frac{1}{2}d + \frac{1}{4}d^2\right). \tag{4.13}$$

Η τιμή που δίνεται από την σχέση (4.13) μπορεί να αντικατασταθεί στη σχέση (4.12) ώστε να αποκτήσουμε την βέλτιστη τιμή του πλάτους του κελιού, η οποία είναι:

$$h_{opt} = A(K)n^{-\frac{1}{d+4}}, \tag{4.14}$$

όπου η σταθερά A(κ) δίνεται από την σχέση :

$$A(K) = [d\beta a^{-2}\{\int(\nabla^2\varphi)^2\}^{-1}]^{1/(d+4)}, \qquad (4.15)$$

η οποία εξαρτάται από την συνάρτηση πυρήνα.

Εάν χρησιμοποιήσουμε την εκτίμηση της σχέσης (4.7), τότε το $h_{opt}$ που προκύπτει από την σχέση (4.14) δίνει άμεσα μια κατάλληλη τιμή για το πλάτος του κελιού. Από την άλλη πλευρά, εάν η συνάρτηση πυρήνα είναι ακτινωτά συμμετρική και οι παρατηρήσεις δεν έχουν μετατραπεί, η διαδικασία αποδεικνύει ότι η $\sigma^2$ είναι η μέση οριακή διακύμανση.

Η αναφορά που έχουμε κάνει στην Παράγραφο 3.4.2 σχετικά με την χρήση της τιμής $1.06\sigma n^{-1/5}$ για το πλάτος του κελιού στην περίπτωση των μονομεταβλητών δεδομένων, εφαρμόζεται και στη περίπτωση πολυμεταβλητών δεδομένων και συμπίπτει με την τιμή που δίνεται από την σχέση (4.14). Παρ' όλα αυτά, η μέθοδος που μόλις περιγράψαμε δίνει μια γρήγορη και εύκολη επιλογή για την τιμή του πλάτους του κελιού.

### 4.3.3. Πιο εξελιγμένοι τρόποι επιλογής του πλάτους του κελιού

Τόσο η least-square cross validation μέθοδος όσο και η likelihood cross validation μέθοδος μπορούν να χρησιμοποιηθούν και στην περίπτωση των πολυμεταβλητών δεδομένων. Η least-square cross validation συνάρτηση score $M_1(h)$ που δίνεται από την σχέση (3.33) σε αυτήν την περίπτωση δίνεται από την σχέση:

$$M_1(h) = n^{-2}h^{-d}\sum_i\sum_j K^*\{h^{-1}(X_i - X_j)\} + 2n^{-1}h^{-d}K(0). \qquad (4.16)$$

Πρέπει να τονίσουμε ότι οι υπολογιστικές προσπάθειες που απαιτούνται για τον υπολογισμό της συνάρτησης $M_1(h)$ εξαρτώνται από τις d-διαστάσεις μόνο την στιγμή που υπολογίζουμε τις τετραγωνικές αποστάσεις $(X_i - X_j)^T(X_i - X_j)$.

Η δυσκολία που παρουσιάζεται στην likelihood cross validation μέθοδο γίνεται ακόμη μεγαλύτερη στην περίπτωση των πολυμεταβλητών δεδομένων, επειδή είναι πιο δύσκολο να ανιχνεύσουμε τις απομακρυσμένες παρατηρήσεις όταν αυτές επεκτείνονται σε χώρο πολλών διαστάσεων.

Παρ' όλα αυτά, έχουν παρατηρηθεί ικανοποιητικά αποτελέσματα χρησιμοποιώντας συναρτήσεις πυρήνα με περιορισμένο στήριγμα, αλλά αυτές οι συναρτήσεις πυρήνα απαιτούν πολύ υπολογιστικό χρόνο. Για αυτό το λόγο είναι προτιμότερο να χρησιμοποιούμε την least-square cross validation μέθοδος.

Η test graph μέθοδος μπορεί και αυτή να επεκταθεί σε περισσότερες διαστάσεις, εάν και δεν είναι πιθανό να βρούμε εφικτή προσέγγιση σε περισσότερες από δυο διαστάσεις.

Τέλος, η επαναληπτική προσέγγιση των Scott-Tapia-Thompson που περιγράφτηκε στην Παράγραφο 3.4.6 μπορεί και αυτή με τη σειρά της να χρησιμοποιηθεί στην περίπτωση των πολυμεταβλητών δεδομένων, εάν και οι υπολογισμοί είναι πιθανό να είναι δυσκολότεροι.

## 4.4. Εκτιμήτρια πυρήνα μεταβλητού πλάτους κελιού

Σε αυτή την τελευταία παράγραφο αυτού του κεφαλαίου θα προσπαθήσουμε να βελτιώσουμε την εκτιμήτρια με την μέθοδο του πυρήνα για πολυμεταβλητές παρατηρήσεις δίνοντας την δυνατότητα στα πλάτη των κελιών να ποικίλουν. Μια προσέγγιση για αυτό το σκοπό γίνεται με την χρήση της εκτιμήτριας με την μέθοδο των K- κοντινότερων γειτόνων, όπως και στην Παράγραφο 2.6, για πολυμεταβλητές παρατηρήσεις, που δίνεται από την ακόλουθη σχέση:

$$\hat{f}(x) = \frac{k}{nV_d h_k(x)^d}, \tag{4.17}$$

όπου $h_k(x)$ είναι η απόσταση των $X_i$ από το $k^o$- κοντινότερο σημείο και $V_d$ είναι ο όγκος της μοναδιαίας σφαίρας $S_d$.

Έτσι, η εκτιμήτρια με την μέθοδο των K- κοντινότερων γειτόνων μπορεί να γραφεί με την μορφή της εκτιμήτριας πυρήνα, εάν η συνάρτηση πυρήνα K εχει επιλεγεί να είναι μια ομοιόμορφη συνάρτηση στην μοναδιαία σφαίρα $S_d$ ως εξής:

$$\hat{f}(x) = \frac{1}{nh_d(x)^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_k(x)}\right). \tag{4.18}$$

Στο τέλος, θα ορίσουμε αυτήν την εκτιμήτρια για διμεταβλητές παρατηρήσεις καθώς θα την χρησιμοποιήσουμε στο επόμενο Κεφάλαιο. Οπότε, έστω ότι έχουμε ένα δείγμα από

διμεταβλητές παρατηρήσεις $(x_i, y_i)$, η εκτιμήτρια με την μέθοδο του πυρήνα θα πάρει την μορφή:

$$\hat{f}(x, y) = \frac{1}{n h_k(x) h_k(y)} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_k(x)}\right) K\left(\frac{y - y_i}{h_k(y)}\right). \quad (4.19)$$

# ΚΕΦΑΛΑΙΟ 5

## 5. NADARAYA-WATSON ΕΚΤΙΜΗΤΡΙΑ

### 5.1. Εισαγωγή

Υποθέτουμε ότι έχουμε ένα δείγμα από πραγματικές διμεταβλητές παρατηρήσεις $(x_i, Y_i)$ που ανήκουν σε μια από κοινού συνάρτηση πυκνότητας πιθανότητας f. Το πρόβλημα που θα συζητήσουμε σε αυτή την Παράγραφο είναι η εύρεση μη παραμετρικών εκτιμητριών, της δεσμευμένης μέσης τιμής και της δεσμευμένης διακύμανσης $m(x) \equiv E(Y|X = x), V(x) \equiv Var(Y|X = x)$. Μέθοδοι που βασίζονται στην εκτιμήτρια πυρήνα είναι υψίστης σημασίας στη δουλειά των Nadaraya (1964) και Watson (1964). Ωστόσο, η έρευνα στις Nadaraya-Watson εκτιμήτριες $\hat{m}_{NW}(x)$ και $\hat{V}_{NW}(x)$ δείχνει ότι αυτές είναι, ως προς κάποιο βαθμό, ασυνεπείς σε σχέση με την εκτιμήτρια πυρήνα $\hat{f}$, η οποία δίνει καλύτερη προσέγγιση της f. Η ασυνέπεια είναι ως προς το γεγονός ότι αυτές σχηματίζουν εκτιμήτριες χρησιμοποιώντας την εμπειρική κατανομή των $y_i$ αντί της εξομαλυμένης κατανομής πυρήνα. Αυτή η έρευνα προτείνει ότι η χρήση εξομαλυμένων εκτιμητριών $\hat{m}_S(x)$ και $\hat{V}_S(x)$ θα είχε καλύτερα αποτελέσματα.

Ο Watson κατασκεύασε την εκτιμήτρια $\hat{m}_{NW}(x)$ ώστε να παρέχει μια απλή υπολογιστική μέθοδο για την απόκτηση ενός γραφήματος για μεγάλο αριθμό παρατηρήσεων, η οποία σχεδιάζει μια καμπύλη σε ένα γράφημα διασποράς, για να αποκαλύψει ένα πρότυπο της σχέσης που κρύβεται από τον αριθμό και την διακύμανση των σημείων του γραφήματος.

### 5.2. Ορισμός Nadaraya-Watson εκτιμήτριας

Υποθέτουμε ότι έχουμε ένα δείγμα από ανεξάρτητες διμεταβλητές παρατηρήσεις $(x_1, y_1), \dots, (x_n, y_n)$. Η Nadaraya-Watson εκτιμήτρια, η οποία θα μελετηθεί σε αυτή την παράγραφο, κατασκευάζεται από δεδομένα που προέρχονται από μια από κοινού σ.π.π. f. Το μοντέλο παλινδρόμησης είναι:

$$Y_i = m(x_i) + e_i, \quad i = 1,2, \dots, n, \tag{5.1}$$

όπου η συνάρτηση $m(x_i)$ είναι άγνωστη και τα σφάλματα $\{e_i\}$ είναι τυχαίες μεταβλητές.

Τα σφάλματα $\{e_i\}$, ακόμη, ικανοποιούν τις ακόλουθες συνθήκες:

$$E[e_i] = 0, \quad V[e_i] = \sigma_\varepsilon{}^2, \quad Cov[e_i, e_j] = 0 \ \text{για} \ i \neq j.$$

Μπορούμε, επίσης, να εκφράσουμε την $m(x)$ ως εξής:

$$m(x) = E[Y|X = x] = \int yf(y|x)dy = \frac{\int yf(x,y)dy}{\int f(x,y)dy}, \tag{5.2}$$

Για να βρούμε την εκτιμήτρια, πρέπει να εκτιμήσουμε τον αριθμητή και τον παρανομαστή αυτής της σχέσης χρησιμοποιώντας, όμως, εκτιμήτριες πυρήνα.

Αρχικά ορίζουμε την εκτιμήτρια πυρήνα για διμεταβλήτες παρατηρήσεις ως εξής:

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_x}\right) K\left(\frac{y - y_i}{h_y}\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i) K_{h_y}(y - y_i),$$

όπου $K_{h_x}(x - x_i) = \frac{1}{nh_x} K\left(\frac{x - x_i}{h_x}\right)$.

Για τον υπολογισμό του αριθμητή έχουμε ότι:

$$\int y\hat{f}(x, y)dy = \frac{1}{n} \int y \sum_{i=1}^{n} K_{h_x}(x - x_i) K_{h_y}(y - y_i),$$

και ότι

$$\int y\, K_{h_y}(y - y_i)dy = y_i.$$

Οπότε προκύπτει η σχέση:

$$\int y\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i) y_i.$$

Για τον υπολογισμό του παρανομαστή έχουμε:

$$\int \hat{f}(x, y)dy = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i) \int K_{h_y}(y - y_i)dy$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i) = \hat{f}(x),$$

καθώς το ολοκλήρωμα $\int K_{h_y}(y - y_i)dy$ είναι ίσο με την μονάδα.

Άρα, η Nadaraya-Watson εκτιμήτρια δίνεται από την ακόλουθη σχέση:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_{h_x}(x-x_i)y_i}{\sum_{i=1}^{n} K_{h_x}(x-x_i)}. \quad (5.3)$$

## 5.3. Ασυμπτωτικές ιδιότητες

Σε αυτή τη Παράγραφο θα προσπαθήσουμε να βρούμε προσεγγιστικές εκφράσεις για την μεροληψία και την διακύμανση. Αυτό είναι λίγο περίπλοκο καθώς η εκτιμήτρια ορίζεται ως μια αναλογία από δυο συσχετισμένες τυχαίες μεταβλητές. Για τον παρανομαστή ισχύουν οι σχέσεις:

$$E[\hat{f}(x)] \approx f(x) + \frac{1}{2}h^2 k_2 f''(x)$$

$$V[\hat{f}(x)] \approx \frac{1}{nh} f(x) \int K(t)^2 dt, \quad (5.4)$$

όπως προκύπτει από την Παράγραφο 3.3.1.

Για τον αριθμητή έχουμε:

$$E[\sum_{i=1}^{n} K_{h_x}(x - x_i)Y_i] = \iint v \frac{1}{n} K\left(\frac{x-u}{h_x}\right) f(u, v)dudv$$

$$= \iint vK(t)f(x - ht, v)dtdv, \quad (5.5)$$

χρησιμοποιώντας την αλλαγή μεταβλητών $t = \frac{x-u}{h_x}$.

Γνωρίζουμε όμως ότι:

$$f(x - ht, v) = f(v|x - ht)f(x - ht).$$

Αντικαθιστώντας αυτήν την σχέση στην σχέση (5.5) έχουμε:

$$E[\sum_{i=1}^{n} K_{h_x}(x - x_i)Y_i] = \iint vK(t)f(v|x - ht)f(x - ht)dtdv$$

$$= \int K(t)f(x - ht) \int vf(v|x - ht)\, dvdt$$

$$= \int K(t)f(x - ht)m(x - ht)dt$$

$$= f(x)m(x) + h_x^2 k_2[f'(x)m'(x) + \frac{1}{2}f''(x)m(x) + \frac{1}{2}f(x)m''(x) + o(h^2), \quad (5.6)$$

χρησιμοποιώντας το ανάπτυγμα της σειράς Taylor για τις συναρτήσεις $f(x-ht)$ και $m(x-ht)$.

Έτσι έχουμε ότι:

$$E[\hat{m}(x)] \approx \frac{E\int \hat{f}(x,y)ydy}{E\hat{f}(x)} \tag{5.7}$$

$$= \frac{f(x)m(x) + {h_x}^2 k_2[f'(x)m'(x) + \frac{1}{2}f''(x)m(x) + \frac{1}{2}f(x)m''(x)}{f(x) + \frac{1}{2}h^2 k_2 f''(x)}$$

$$= m(x) + \frac{{h_x}^2}{2}k_2\left[m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right], \tag{5.8}$$

χρησιμοποιώντας την προσεγγιστική σχέση $(1-h^2c)^{-1} \approx (1-h^2c)$ για μικρά h.

Επομένως, η μεροληψία δίνεται από την ακόλουθη σχέση:

$$\text{bias}\big(\hat{m}(x)\big) \approx \frac{{h_x}^2}{2}k_2\left[m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right]. \tag{5.9}$$

Για να αποκτήσουμε μια προσέγγιση της διακύμανσης χρησιμοποιούμε την παρακάτω σχέση της διακύμανσης ως αναλογία δυο τυχαίων μεταβλητών, της N και της D, όπου N είναι ο αριθμητής και D ο παρανομαστής της σχέσης (5.3). Δηλαδή:

$$V\left(\frac{N}{D}\right) \approx \left(\frac{EN}{ED}\right)^2\left[\frac{V(N)}{(EN)^2} + \frac{V(D)}{(ED)^2} - \frac{2Cov(N,D)}{(EN)(ED)}\right]. \tag{5.10}$$

Επίσης, έχουμε ότι:

$$V[\tfrac{1}{n}\textstyle\sum_{i=1}^{n} K_{h_x}(x-x_i)Y_i] = \frac{1}{n}E\big[K_{h_x}(x-x_i)Y_i\big]^2 - O(n^{-1})$$

$$= \iint v^2 \frac{1}{n}K\left(\frac{x-u}{h_x}\right)^2 f(u,v)dudv$$

$$= \iint \frac{1}{n}\frac{1}{h}v^2K(t)^2 f(x-ht,v)dtdv,$$

χρησιμοποιώντας την αλλαγή μεταβλητών $t = \frac{x-u}{h_x}$.

Γνωρίζουμε όμως ότι:

$$f(v|x-ht) = \frac{f(x-ht,v)}{f(x-ht)},$$

οπότε

$$f(x - ht, v) = f(v|x - ht)f(x - ht),$$

και ότι:

$$\int v^2 f(v|x - ht) = {\sigma_\varepsilon}^2(x - ht) + m(x - ht)^2.$$

Άρα έχουμε:

$$V\left[\frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x - x_i)Y_i\right] = \int \frac{1}{n}\frac{1}{h} K^2(t) f(x - ht) \int v^2 f(v|x - ht)$$

$$= \int \frac{1}{n}\frac{1}{h} K^2(t) f(x - ht)\left[{\sigma_\varepsilon}^2(x - ht) + m(x - ht)^2\right]dt$$

$$= \frac{1}{n}\frac{1}{h}\int K^2(t) f(x)\left[{\sigma_\varepsilon}^2 + m(x)^2\right]dt, \qquad (5.11)$$

χρησιμοποιώντας το ανάπτυγμα της σειράς Taylor.

Ακόμη, έχουμε ότι:

$$V\left[\hat{f}(x)\right] \approx \frac{1}{n}\frac{1}{h}\int K^2(t) f(x)\, dt.$$

Τέλος,

$$Cov(N, D) = \frac{1}{n} E\left[K_{h_x}(x - x_i)^2 Y_i\right] - O(n^{-1})$$

$$\approx \frac{1}{n}\frac{1}{h}\int K^2(t) f(x) m(x)\, dt,$$

χρησιμοποιώντας όλες τις παραπάνω προσεγγίσεις.

Άρα, αντικαθιστώντας όλες τις σχέσεις αυτές στην αρχική σχέση (5.10), προκύπτει η προσεγγιστική σχέση για την διακύμανση:

$$V[\hat{m}(x)] \approx \frac{1}{n}\frac{1}{h}\int K^2(t)\frac{{\sigma_\varepsilon}^2}{f(x)}dx. \qquad (5.12)$$

Γνωρίζοντας την μεροληψία και την διακύμανση, μπορούμε να βρούμε την τιμή του MISE, η όποια είναι η εξής:

$$MISE = \frac{{h_x}^2}{2} k_2 \left[m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right] + \frac{1}{n}\frac{1}{h}\int K^2(t)\frac{{\sigma_\varepsilon}^2}{f(x)}dx. \qquad (5.13)$$

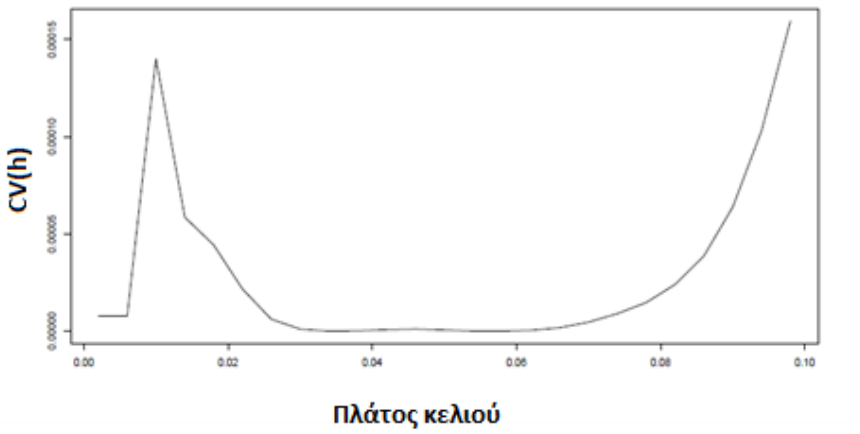
Τέλος, μπορούμε να υπολογίσουμε την βέλτιστη τιμή του πλάτους του κελιού, η οποία θα είναι που ελαχιστοποιεί το μέσο ολοκληρώσιμο τετραγωνικό σφάλμα.

## 5.4. Παράδειγμα

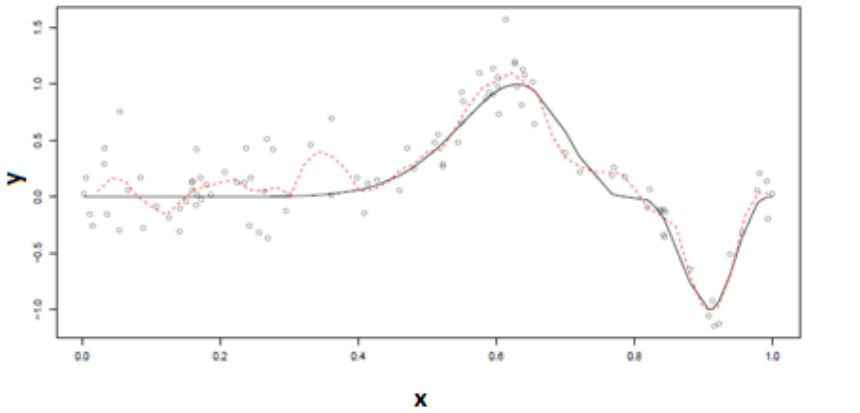
Θεωρούμε ότι έχουμε $n = 100$ διμεταβλητές παρατηρήσεις $(x_i, y_i)$, $i = 1, \dots, n$ από το μοντέλο παλινδρόμησης:

$$Y_i = m(x_i) + e_i, \quad i = 1,2, \dots, n$$

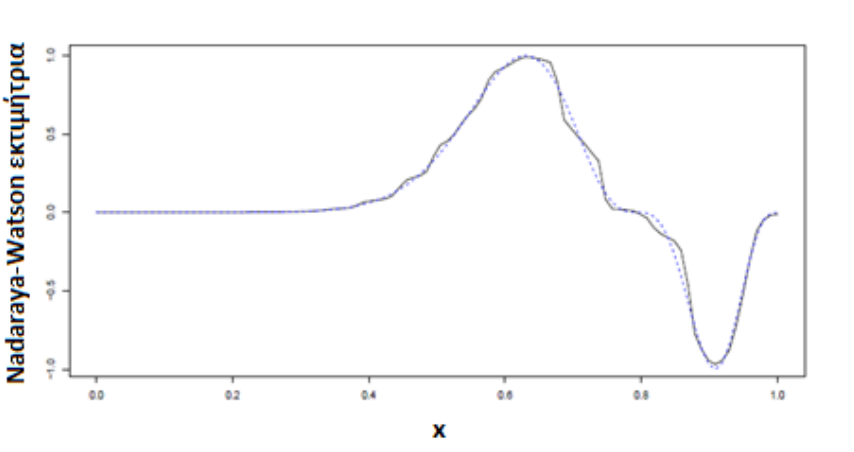
όπου η συνάρτηση $m(x_i)$ δίνεται από τον τύπο $m(x) = \sin(2 * \pi * x^3)^3$. Οι παρατηρήσεις ανήκουν στην τυποποιημένη κανονική κατανομή. Η least square cross-validation μέθοδος χρησιμοποιείται για την επιλογή μιας προσέγγισης του ποσοστού της εξομάλυνσης στο διάστημα (0,1) και ένα Γράφημα της συνάρτησης $CV(h)$ παρουσιάζεται στο Γράφημα 5.1 το οποίο κατασκευάστηκε από 25 διαφορετικές τιμές του h στο διάστημα (0,0.1). Η ελάχιστη τιμή του h είναι $h = 0.034$. Στο Γράφημα 5.2 παρουσιάζετε ένα Γράφημα διασποράς των παρατηρήσεων, με την Nadaraya-Watson εκτιμήτρια, με h που βρίσκεται από την least square cross-validation μέθοδο, μαζί με την πραγματική συνάρτηση παλινδρόμησης. Παρατηρούμε ότι η εκτιμήτρια ταιριάζει πολύ καλά με την πραγματική συνάρτηση στο διάστημα (0.4,1) ενώ παρουσιάζει θόρυβο στο διάστημα (0,0.4), όπου η πραγματική συνάρτηση είναι πιο επίπεδη. Το Γράφημα 5.3 δείχνει τη γραφική παράσταση των προβλεπτικών τιμών της εκτιμήτριας για $h = 0.034$ μαζί με την πραγματική συνάρτηση παλινδρόμησης. Για αυτήν την εκτιμήτρια και με το δοσμένο h, παρατηρούμε ότι οι προβλεπτικές τιμές είναι πολύ κοντά στις πραγματικές. Αυτό φαίνεται καλύτερα στο Γράφημα της μεροληψίας που απεικονίζεται στο Γράφημα 5.4. Τέλος, η πραγματική διακύμανση απεικονίζεται στο Γράφημα 5.5 και το μέσο τετραγωνικό σφάλμα (MSE) στο Γράφημα 5.6. Παρατηρούμε ότι τα δυο τελευταία Γραφήματα έχουν παραπλήσιο σχήμα, αυτό συμβαίνει γιατί η διακύμανση είναι λίγο μεγαλύτερη από το τετράγωνο της μεροληψίας, για τις δοθέντες παρατηρήσεις και την συγκεκριμένη εκτιμήτρια.



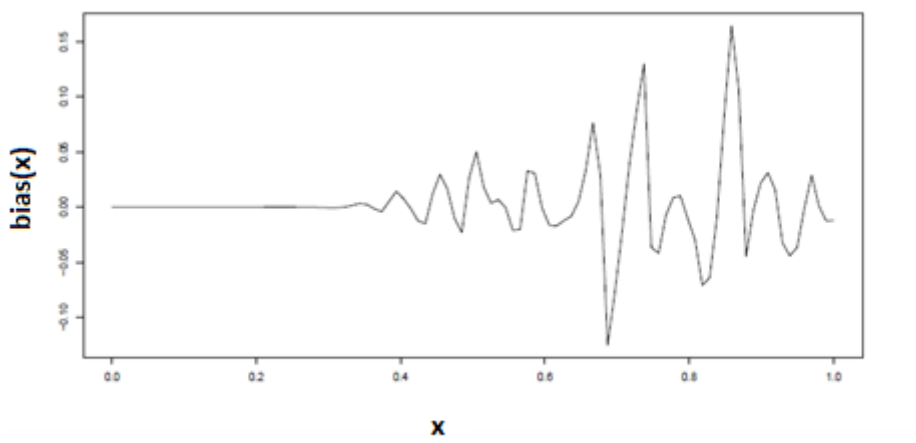

**Γράφημα 5.1: Η least square cross-validation μέθοδος για τις 100 παρατηρήσεις.**



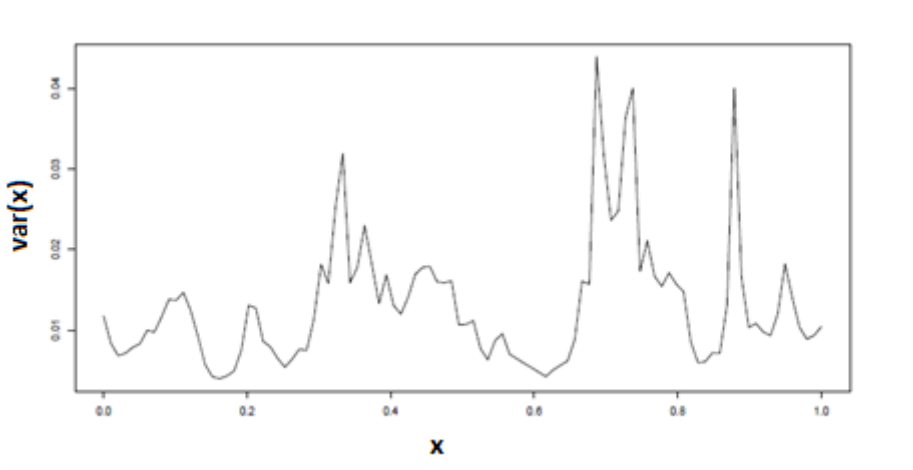

**Γράφημα 5.2: Ένα γράφημα διασποράς, η πραγματική συνάρτηση παλινδρόμησης (μαύρη γραμμή) και η Nadaraya-Watson εκτιμήτρια για πλάτος κελιού h=0,034 (κόκκινη διακεκομμένη γραμμή).**



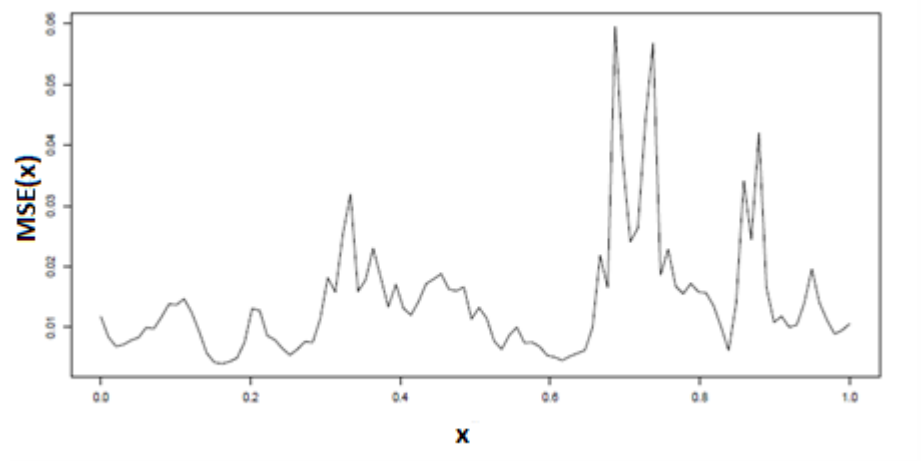

**Γράφημα 5.3: Οι προβλεπτικές τιμές από την Nadaraya-Watson εκτιμήτρια για πλάτος κελιού h=0,034 (μαύρη γραμμή) και η πραγματική συνάρτηση παλινδρόμησης (μπλε διακεκομμένη γραμμή).**




**Γράφημα 5.4: Η μεροληψία με την Nadaraya-Watson εκτιμήτρια για πλάτος κελιού h=0,034.**




**Γράφημα 5.5: Η διακύμανση με την Nadaraya-Watson εκτιμήτρια για πλάτος κελιού h=0,034.**




**Γράφημα 5.6: Το μέσο τετραγωνικό σφάλμα με την Nadaraya-Watson εκτιμήτρια για πλάτος κελιού h=0,034.**

# Επίλογος

Με την ολοκλήρωση της εργασίας είμαστε σε θέση να γνωρίζουμε κάποιους από τους πιο διαδεδομένους τρόπους κατασκευής μιας εκτιμήτριας και, επίσης, να κρίνουμε ποια εκτιμήτρια πρέπει να εφαρμοστεί κάθε φορά ανάλογα με τον σκοπό για τον οποίο θέλουμε να την χρησιμοποιήσουμε. Μεγαλύτερη έμφαση δώσαμε στην εκτιμήτρια με την μέθοδο του πυρήνα, και παρουσιάσαμε διάφορες μεθόδους που υπολογίζουν την βέλτιστη τιμή για το πλάτος των κελιών και της συνάρτησης πυρήνα, ώστε να έχουμε μείωση της ασυμφωνίας ανάμεσα στην εκτιμήτρια και την πραγματική συνάρτηση. Αυτό έχει ως αποτέλεσμα την απόκτηση μιας βελτιωμένης εκτιμήτριας που θα έχει την καταλληλότερη μορφή, ανάλογα με την στατιστική διαδικασία που θέλουμε να την χρησιμοποιήσουμε.

# Βιβλιογραφία

[1]. Bailley R.W. and Addison J.T. (2010). A Smoothed-Distribution Form of Nadaraya-Watson Estimation. Department of Economics Discussion Paper 10-30,1-12.
[2]. Breiman L., Meisel W. and Purcell E. (1977). Variable kernel estimates of multivariate densities. Technometrics, **19**, 135-144.
[3]. Hall P. (1983). Large sample optimality of least square cross-validation in density estimation. The Annals of Statistics, **11**, 1156-1174.
[4]. Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., **33**, 1065-1076.
[5]. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist., **27**, 832-837.
[6]. D.W.,Baggerly K.A. and Scott D.W. (1994). Cross- Validation of Multivariate Densities. J. Amer. Statis. Accos., **89**, 807-817.
[7]. Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel type density estimators. J. Amer. Statist. Assoc., **72**, 420-423.
[8]. Scott, D.W. (1979). On optimal and data-based histograms. Biometrica, **66**, 605-610.
[9]. Scott D.W., Gorry G.A., Hoffman R.G., Barboriak J.J. and Gotto A.M. (1980). A new approach for evaluating risk factors in coronary disease: a study of lipid concentrations and severity of disease in 1847 males. Circulation, J. Amer. Heart Assoc., **62**, 477-484.
[10]. Scott, D.W. and Factor, L.E. (1981). Monte Carlo study of three data-based nonparametric density estimators. J. Amer. Statist. Assoc., **76**, 9-15.
[11]. Silverman, B.W. (1978a). Choosing the window width when estimating a density. Biometrica, **65**, 1-11.
[12]. Silerman B.W.(1986). Density Estimation for Statistic and Data Analysis. Published in Monographs on Statistics and Applied Probability, London: Chapman and Hall.
[13]. Terrell G.R. and Scott D.W. (1992). Variable Kernel Density Estimation. Ann. Statist., **20**, 1236-1265.
[14]. Wand M.P. and Jones M.C. (1994). Multivariate plug-in bandwidth selection. Computational Statistic, **9**, 97-116.