ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

Πρώτη Έκδοση
Μεταπτυχιακή Εργασία Ειδίκευσης

Εκτιμητές Πολυδιάστατων Δειγμάτων : Γενική Παρουσίαση και Μελέτη του Τριγωνικού Εκτιμητή για Διδιάστατα Δείγματα.

Δημήτριος Ακριβός

1
2
Μ.Ε.

Επιβλέπων Καθηγητής Λεωνίδας Παληός

3

Φεβρουάριος 2003

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Σε αυτή την εργασία θα μας απασχολήσει το πρόβλημα του αντιπροσωπευτικού σημείου ενός συνόλου από δεδομένα τα οποία βρίσκονται σε κυρτή θέση στο επίπεδο. Δηλαδή θεωρούμε το $S = \{ X_1 , \dots , X_n \}$, ένα σύνολο από n δοθέντα σημεία σε κυρτή θέση στο επίπεδο. Να βρεθεί ένα σημείο p (ή σύνολο σημείων) που να περιγράφει (ή να αντιπροσωπεύει) με τον καλύτερο δυνατό τρόπο το σύνολο σημείων S. Αφού ο καλύτερός δυνατός τρόπος είναι αντικειμενικό ακόμα κριτήριο θα μελετήσουμε πολλούς από τους εκτιμητές του σημείου p, οι οποίοι στηρίζονται στην επέκταση της έννοιας του μεσοστοιχείου σε μεγαλύτερες διαστάσεις. Τελικά θα αναφερθούμε ειδικότερα σε έναν από αυτούς τους εκτιμητές, αυτόν του τριγωνικού εκτιμητή, καθώς θα μελετήσουμε διεξοδικότερα αλγορίθμους εύρεσης του τριγωνικού βάθους και του τριγωνικού εκτιμητή. Σε αυτήν την εργασία θα παρουσιάσουμε κάποιες καινούργιες παρατηρήσεις πάνω στο παραπάνω πρόβλημα σε σχέση με τον τριγωνικό εκτιμητή, ενώ παράλληλα θα προσπαθήσουμε να θέσουμε ένα θεωρητικό υπόβαθρο για την μελέτη του. Οι παρατηρήσεις έχουν γίνει πάνω σε μεγάλο πλήθος πειραματικών δεδομένων, ενώ μέρος του θεωρητικού υπόβαθρου περιλαμβάνει τον ορισμό μίας νέας σχετικής δομής του πίνακα των σημείων τομής, η οποία μας βοηθάει να κατανοήσουμε τα πειραματικά συμπεράσματα, αλλά και να τα αποδείξουμε για πολύ μικρό αριθμό σημείων.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1. Ορισμός του Προβλήματος.

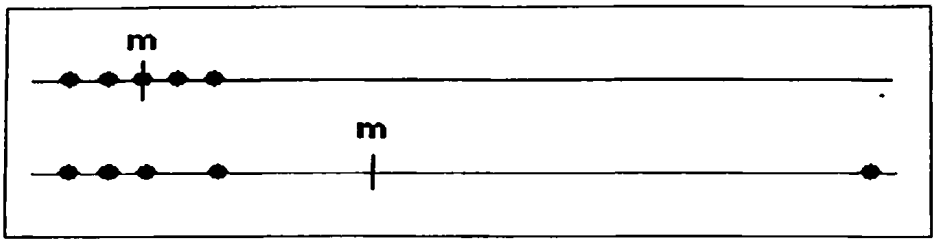
Θεωρούμε το ακόλουθο πρόβλημα : Έστω ότι μας δίνεται ένα σύνολο από σημεία με συνάρτηση κατανομής πιθανότητας (probability distribution) η οποία είτε είναι άγνωστη είτε είναι μη-παραμετρική και μονότονη. Μπορούμε κάτω από αυτές τις προϋποθέσεις να υπολογίσουμε ένα σημείο (ή περιοχή) που να περιγράφει κατά τον καλύτερο τρόπο το παραπάνω σύνολο από σημεία δεδομένων;

Ορισμός 1.1. *Έστω $S = \{X_1, ..., X_n\}$ ένα σύνολο από n δοθέντα σημεία στον R^d. Να βρεθεί ένα σημείο p (ή σύνολο σημείων) που να περιγράφει (ή να αντιπροσωπεύει) με τον καλύτερο δυνατό τρόπο το σύνολο σημείων S.*

Επειδή ο «καλύτερος» δυνατός τρόπος αντιπροσώπευσης ενός συνόλου είναι μία υποκειμενική έννοια, υπάρχουν πολλές διαφορετικές προσεγγίσεις για την εύρεση αυτού του σημείου p, το οποίο και ονομάζεται **εκτιμητής**. Κοινό γνώρισμα πάντως όλων των προσεγγίσεών είναι ότι ο εκτιμητής οφείλει να έχει κάποιες συγκεκριμένες ιδιότητες με κυριότερη αυτή της ευστάθειας (robustness).

1.2. Ευστάθεια του Προβλήματος.

Όταν εργαζόμαστε με τέτοια προβλήματα είναι επόμενο να λαμβάνουμε σοβαρά υπόψη μας τον παράγοντα της ευστάθειας, δηλαδή το κατά πόσο αλλάζει η απάντηση μας στο παραπάνω πρόβλημα αν μερικά από τα δοθέντα σημεία είτε μεταβληθούν (δηλαδή, αλλάξουν συντεταγμένες ή εφαρμοστούν πάνω τους γεωμετρικοί μετασχηματισμοί) είτε αντικατασταθούν από άλλα σημεία, ή το πόσο εύκολο είναι να κατασκευαστεί ένα σύνολο από σημεία δεδομένων που ο εκτιμητής μας να δίνει μία απάντηση (ένα σημείο) που να είναι εντελώς παράλογη (για παράδειγμα το αντιπροσωπευτικό σημείο να βρίσκεται έξω από το κυρτό περίβλημα των σημείων που καλείται να αντιπροσωπεύσει). Οι κυριότεροι παράγοντες ευστάθειας ορίζονται στο παράρτημα Α αυτής της εργασίας. Χαρακτηριστικό παράδειγμα μη-ευσταθούς εκτιμητή (γεγονός που καταδεικνύει και την πολυπλοκότητα του προβλήματος) είναι η περίπτωση του **γεωμετρικού μέσου (mean)** ενός συνόλου από σημεία. Είναι εύκολο να διαπιστώσουμε ότι αν ένα σημείο μετακινηθεί πάρα πολύ μακριά από τα υπόλοιπα, ο γεωμετρικός μέσος θα ακολουθήσει την μετακίνηση του σημείου αυτού. Για να κάνουμε το παράδειγμα πιο απτό μπορούμε εύκολα να δούμε ότι αν το σημείο αυτό μετακινηθεί στο άπειρο τότε και ο γεωμετρικός μέσος θα μετακινηθεί στο άπειρο. Το παραπάνω παράδειγμα καταδεικνύει την σημαίνουσα σπουδαιότητα που έχει η ευστάθεια του εκτιμητή όταν υπάρχει η πιθανότητα το σύνολο δεδομένων μας να έχει τέτοιου είδους «προβληματικά» σημεία. Είναι επίσης ουσιώδες να μην δίνει ο εκτιμητής μας περισσότερη βαρύτητα σε κάποια σημεία «αδικώντας» τα υπόλοιπα. Αν χρησιμοποιήσουμε τον γεωμετρικό μέσο ως εκτιμητή, τότε τα απομακρυσμένα σημεία έχουν μεγαλύτερη βαρύτητα απ' ότι έχουν τα σημεία που βρίσκονται κοντά στον μέσο. Στο σχήμα 1.1 βλέπουμε ότι αν η διαγραφή ενός απομακρυσμένου σημείου από το σύνολο δεδομένων μας αυτό θα έχει μεγαλύτερη επίδραση στην καινούργια θέση του μέσου από ότι η διαγραφή ενός σημείου από την πιο πυκνή περιοχή δεδομένων: στην πρώτη περίπτωση ο μέσος θα επιστρέψει (διανύοντας μεγάλη απόσταση) στο μέσο της πυκνής περιοχής δεδομένων ενώ στην δεύτερη περίπτωση θα μετακινηθεί λίγο ακόμα προς το απομακρυσμένο σημείο.



Σχήμα 1.1. Ο μέσος (**m**) δεν αποτελεί ευσταθή εκτιμητή.

Ορισμός 1.2. *Το μεσοστοιχείο (median) σε ένα σύνολο μονοδιάστατων σημείων είναι το μεσαίο στοιχείο του συνόλου.*

Για να βρούμε το μεσοστοιχείο ενός συνόλου n στοιχείων ταξινομούμε τα στοιχεία του συνόλου και διαλέγουμε είτε το $\frac{n+1}{2}$ *στοιχείο αν το σύνολο έχει περιττό πλήθος στοιχείων είτε την μέση τιμή των δύο κεντρικών ταξινομημένων στοιχείων αν το σύνολο έχει άρτιο πλήθος στοιχείων.*

1.3. Παράγοντες και Μέτρα Ευστάθειας.

Όπως αναφέρθηκε και παραπάνω, ένα σημείο είναι αρκετό για να επηρεάσει σημαντικά τον μέσο ενός συνόλου δεδομένων. Αντιθέτως τουλάχιστον τα μισά από τα σημεία του συνόλου δεδομένων πρέπει να μετακινηθούν στο άπειρο ώστε να αναγκάσουν το μεσοστοιχείο να τα ακολουθήσει στο άπειρο. Το παραπάνω γεγονός συνεπάγεται τον ακόλουθο παράγοντα ευστάθειας για τους διάφορους εκτιμητές μεσοστοιχείου:

- **Το σημείο κατάρρευσης (breakdown point)** του συνόλου δεδομένων αποτελεί το ποσοστό των σημείων του συνόλου που πρέπει να μετακινηθούν στο άπειρο ώστε και ο εκτιμητής να μετακινηθεί επίσης στο άπειρο [Sma90].

Στον R^1 (σε ένα μονοδιάστατο –univariate- σύνολο δεδομένων) το μεσοστοιχείο έχει σημείο κατάρρευσης κοντά στο ½ ενώ ο μέσος έχει σημείο κατάρρευσης στο $\frac{1}{n}$, όπου n είναι το πλήθος των σημείων του συνόλου δεδομένων. Έχει δειχτεί [RL91] ότι το μέγιστο σημείο κατάρρευσης για οποιονδήποτε εκτιμητή είναι το ½ και από αυτό καταλαβαίνουμε ότι στον R^1 το μεσοστοιχείο αριστεύει σε

αυτο το κριτήριο ευσταθειας. Ειδικότερα έχει αποδειχτεί [Bas91] για τον R^1 ότι το μεσοστοιχειο ειναι ο μόνος εκτιμητής που κατέχει τις ιδιότητες της ισο-μεταβλητότητας (ιδιότητας πιο γνωστης ως γραμμικη αμεταβλητότητα), του μέγιστου σημείου καταρρευσης και της μονοτονικότητας (δηλαδη, της ιδιοτητας που έχει το μεσοστοιχειο να μην μετακινειται σε κατευθυνση αντιθετη με αυτη μιας διαταραχής).

Μέτρα και κριτηρια ευστάθειας τα οποία τυπικά χρησιμοποιούνται στον R^2 ή σε μεγαλύτερους χώρους περιλαμβάνουν την αμεταβλητοτητα σε συγκεκριμένους τύπους μετασχηματισμών των δεδομένων. Για παράδειγμα, δοθέντος ενός συνόλου σημείων στο επιπεδο δεν θέλουμε με κανέναν τροπο ο εκτιμητής μας να επηρεάζεται από την επιλογή των αξόνων συντεταγμένων (ούτε επίσης και του σημείου αναφοράς, της κατεύθυνσης και της κλίμακας). Ένας εκτιμητής είναι αμετάβλητος σε γραμμικούς μετασχηματισμούς του συνόλου δεδομένων, εάν η θέση του σε σχέση με τα δεδομένα δεν επηρεαζεται απο την μετακινηση, την περιστροφή, την αλλαγη κλίμακας ή το τράβηγμα (shearing) του συστήματος συντεταγμένων.

Έχουμε ήδη δει ότι η υψηλή τιμή σημείου καταρρευσης των μεσοστοιχείων στον R^1 είναι ο λόγος που τους προτιμάμε σε σχέση με τον μέσο ως εκτιμητή του αντιπροσωπευτικού σημείου του συνόλου. Η τιμή του σημείου καταρρευσης του μέσου είναι $\frac{1}{n}$ σε όλες τις διαστάσεις, γεγονός το οποίο εύκολα οπτικοποιείται στον R^2 ή στον R^3.

1.4. Προτάσεις Επέκτασης του Μονοδιάστατου Μεσοστοιχείου στον R^d.

Στην προσπάθειά μας να κάνουμε μία παρόμοια εκτίμηση και για το μεσοστοιχείο, αμέσως ερχόμαστε αντιμέτωποι με το εξής πρόβλημα : **Τι είναι (ή πώς ορίζεται) το μεσοστοιχείο ενός πολυδιάστατου συνόλου δεδομένων;** Αυτή η ερώτηση απασχόλησε τους μαθηματικούς για τουλάχιστον έναν αιώνα και ως συνέπεια έχουν οριστεί πολλές επεκτάσεις. Υπάρχουν περισσότεροι από ένας τρόποι να περιγράψεις ένα μονοδιάστατο μεσοστοιχείο οι οποίοι έχουνε πολύ καλά και γνωστά αποτελέσματα στον R^1 και για αυτόν τον λόγο υπάρχουν πολλοί πολυδιάστατοι γενικευμένοι ορισμοί.

Είναι ξεκάθαρο ότι δεν μπορούμε να γενικεύσουμε το μονοδιάστατο μεσοστοιχείο σε υψηλότερες διαστάσεις επιλέγοντας εκ περιτροπής μία από τις

κατευθύνσεις του πολυδιάστατου χώρου. Πολλοί πρότειναν τον συνδυασμό χρήσης διαφόρων μεσοστοιχείων σε διαφορετικές κατευθύνσεις. Το 1902 ο Hayford πρότεινε το **διάνυσμα-των-μεσοστοιχείων** των ορθογώνιων συντεταγμένων [Sma90]. Αυτό το μεσοστοιχείο περιλαμβάνει την επιλογή ενός ορθογώνιου συστήματος συντεταγμένων και τον υπολογισμό του μονοδιάστατου μεσοστοιχείου κατά μήκος κάθε άξονα. Η μέθοδος αυτή δεν διαφοροποιείται αισθητά από την εύρεση του πολυδιάστατου μέσου και δουλεύει καλά και σε άλλα προβλήματα όπως η εύρεση παραγώγων συναρτήσεων. Δυστυχώς, όπως παρατήρησε και ο Hayford, το διάνυσματων- μεσοστοιχείων εξαρτάται από την επιλογή των ορθογώνιων κατευθύνσεων. Είναι πολύ εύκολο να διαπιστώσει κανείς το παραπάνω συμπέρασμα αν χρησιμοποιήσει ένα σύνολο τριών μη-συνευθειακών σημείων. Στην πραγματικότητα, όπως επεσήμανε και ο Rousseeuw [Rou85], αυτή η μέθοδος μπορεί να παράγει ένα μεσοστοιχείο που να βρίσκεται έξω από το κυρτό περίβλημα των δεδομένων σημείων. Ο Mood [Moo41] επίσης πρότεινε **μία κοινή κατανομή από μονοδιάστατα μεσοστοιχεία** και χρησιμοποίησε ολοκλήρωση για να βρει το πολυδιάστατο μεσοστοιχείο.

Αμέσως μετά τον ορισμό του Hayford για το μεσοστοιχείο, ο Weber πρότεινε έναν νέο ορισμό για τον εκτιμητή του αντιπροσωπευτικού σημείου ενός συνόλου δεδομένων. Ο εκτιμητής του Weber ορίζεται ως το σημείο που ελαχιστοποιεί το άθροισμα των αποστάσεων από όλα τα δεδομένα σημεία. Αυτός ο ορισμός ισχύει για το μονοδιάστατο μεσοστοιχείο και κατά συνέπεια μπορεί να χρησιμοποιηθεί και ως περιγραφή του πολυδιάστατου μεσοστοιχείου. Αυτός ο εκτιμητής μπορεί να βρεθεί στην βιβλιογραφία με μία ποικιλία ονομάτων όπως το **L_1-μεσοστοιχείο** [Sma90] ή **το κέντρο των μεσοστοιχείων (mediancentre)**. Μερικές ιδιότητες αυτού του πολυδιάστατου μεσοστοιχείου εκθέτονται στην **ενότητα 2.1.**

Το 1929 ο Hotelling [Hot29] περιέγραψε το μονοδιάστατο μεσοστοιχείο ως το σημείο το οποίο ελαχιστοποιεί το μέγιστο πλήθος των δεδομένων σημείων σε κάποια από τις πλευρές του. Η ιδέα του Hotelling επεκτάθηκε σε μεγαλύτερες διαστάσεις πολλά χρόνια αργότερα από τον Tukey [Tuk75]. **Το μεσοστοιχείο Tukey, ή το μεσοστοιχείο ημιχώρου,** είναι ίσως το περισσότερο μελετημένο και χρησιμοποιημένο πολυδιάστατο μεσοστοιχείο στην εποχή μας και περιγράφεται πιο λεπτομερώς στην **ενότητα 2.2.**

Ένας ενδιαφέρων τρόπος υπολογισμού του μονοδιάστατου μεσοστοιχείου είναι μέσω επαναληπτικής διαγραφής των πιο ακραίων σημείων από το σύνολο

δεδομένων μας. Μία γενίκευση αυτής της πρότασης εμφανίστηκε από τον Shamos [Sha76] και τον Barnett [Bar76], παρόλο που ο Shamos αναφέρει ότι η ιδέα αυτή αρχικά ανήκει στον Tukey. Η **αφαίρεση του κυρτού περιβλήματος (convex hull peeling)** επαναληπτικά απομακρύνει το κυρτό περίβλημα των δεδομένων σημείων μέχρι να απομείνει ένα κυρτό σύνολο (convex set). Το convex hull peeling και η σχετιζόμενη μέθοδος της **αφαίρεσης ελλειψοειδούς (ellipsoid peeling)** περιγράφονται στην **ενότητα 2.3.**

Το 1983 ο Oja [Oja83] εισήγαγε έναν ορισμό για το πολυδιάστατο μεσοστοιχείο, ο οποίος γενικεύει την ιδέα ότι το μονοδιάστατο μεσοστοιχείο είναι το σημείο με το ελάχιστο άθροισμα αποστάσεων από όλα τα δεδομένα σημεία. Ωστόσο, ο Oja ως απόσταση χρησιμοποίησε τον μονοδιάστατο όγκο. Έτσι το **μεσοστοιχείο Oja** είναι το σημείο για το οποίο ελαχιστοποιείται ο συνολικός όγκος των **στοιχειωδών πολυδιάστατων αντικειμένων πλήρους διάστασης**, τα οποία σχηματίζονται από το σημείο αυτό και κατάλληλα υποσύνολα του συνόλου δεδομένων. **Ένα στοιχειώδες πολυδιάστατο αντικείμενο πλήρους διάστασης (simplex)** είναι ένα ευθύγραμμο τμήμα στον R^1, ένα τρίγωνο στον R^2, ένα τετράεδρο στον R^3 κ.τ.λ. Η **ενότητα 2.4** περιέχει περισσότερες πληροφορίες για το μεσοστοιχείο Oja.

Το 1990, η Regina Liu [Liu90] πρότεινε έναν ακόμη ορισμό για την πολυδιάστατη περίπτωση, γενικεύοντας το γεγονός ότι το μονοδιάστατο μεσοστοιχείο είναι το σημείο το οποίο περιέχεται στα περισσότερα διαστήματα μεταξύ ζευγών σημείων δεδομένων, περιγράφοντας έτσι το **simplicial μεσοστοιχείο**. Σε μεγαλύτερες διαστάσεις τα διαστήματα αντικαθιστώνται στον ορισμό από στοιχειώδη πολυδιάστατα αντικείμενα πλήρους διάστασης. Η παραπάνω μέθοδος περιγράφεται καλύτερα στην **ενότητα 2.5.**

Οι Gil, Steiger και Wigderson [GSW92] σύγκριναν την ευστάθεια και υπολογιστικά στοιχεία συγκεκριμένων μεσοστοιχείων, έχοντας όμως θέσει τον περιορισμό ότι το μεσοστοιχείο πρέπει να είναι ένα από τα σημεία του συνόλου των δεδομένων. Πρότειναν επίσης έναν νέο ορισμό για το πολυδιάστατο μεσοστοιχείο: για κάθε σημείο από το σύνολο δεδομένων παίρνουμε το διανυσματικό άθροισμα όλων των διανυσμάτων που σχηματίζει με τα υπόλοιπα σημεία του συνόλου δεδομένων. Το μεσοστοιχείο είναι οποιοδήποτε σημείο από το σύνολο δεδομένων για το οποίο το μέτρο του διανυσματικού αθροίσματος του είναι μικρότερο ή ίσο με την μονάδα. Οι συγγραφείς σημειώνουν πως το μονοδιάστατο μεσοστοιχείο ικανοποιεί

την παραπάνω συνθήκη και υποστηρίζουν ότι το προτεινόμενο μεσοστοιχείο έχει την ιδιότητα της μοναδικότητας. Παρόλα αυτά η πρότασή τους μοιάζει παράξενη στον R^2, αφού μπορούμε εύκολα να πάρουμε ως σύνολο δεδομένων ένα συμμετρικό κυρτό σύνολο από σημεία δεδομένων για τα οποία το μήκος του διανυσματικού αθροίσματος είναι μεγαλύτερο από την μονάδα για όλα τα σημεία του συνόλου των δεδομένων. Θα ήταν πιο λογικό να θεωρήσουμε ως υποψήφια για μεσοστοιχείο όλα τα σημεία στον R^2.

Ένας παραπάνω λόγος που χρησιμοποιούμε το μεσοστοιχείο στον R^1 είναι ότι μας παρέχει μία μέθοδο για κατάταξη (ranking) των δεδομένων μας. Αυτή η παράμετρος μπορεί να γενικευτεί για τα μεσοστοιχεία Oja, Liu και Tukey, καθώς και για την αφαίρεση κυρτού περιβλήματος. Καθένα από τα παραπάνω μεσοστοιχεία μεγιστοποιεί ή αντίστοιχα ελαχιστοποιεί μία συγκεκριμένη συνάρτηση βάθους και κάθε σημείο στον R^d μπορεί να πάρει μία τιμή βάθους ανάλογη με την εκάστοτε συνάρτηση βάθους.

Πολύ πρόσφατα προτάθηκε μία καινούργια και ενδιαφέρουσα ιδέα για το πολυδιάστατο βάθος. Το **βάθος υπερεπιπέδου (hyperplane depth)** ενός σημείου ως προς ένα σύνολο από υπερεπίπεδα είναι το ελάχιστο πλήθος υπερεπιπέδων από το σύνολο τα οποία πρέπει να διασχίσει μία ακτίνα εκτεινόμενη από το σημείο αυτό. Στον R^1 ο παραπάνω ορισμός συμπίπτει με αυτόν του μεσοστοιχείου n σημείων, ενώ παράλληλα έχει προταθεί ένας ορισμός του πολυδιάστατου μεσοστοιχείου σχετικά με το βάθος υπερεπιπέδου από τον Aloupis [Alo01]. Περισσότερα για αυτόν τον εκτιμητή θα δούμε στην **ενότητα 2.6.**

1.5. Άλλοι Εκτιμητές.

Σε αυτό το σημείο θα αναφερθούμε σε ορισμένους άλλους εκτιμητές του αντιπροσωπευτικού σημείου ενός συνόλου δεδομένων, οι οποίοι δεν είναι απαραίτητα γενικεύσεις του μονοδιάστατου μεσοστοιχείου, όπως η **κάλυψη με ελλειψοειδές ελάχιστου όγκου (minimal volume ellipsoid covering)** που προτάθηκε από τον Rousseeuw [Rou85]. Αυτός ο εκτιμητής είναι το κέντρο ενός ελλειψοειδούς ελάχιστου όγκου το οποίο περιέχει τουλάχιστον τα μισά από τα δεδομένα σημεία μας. Ο εκτιμητής είναι αμετάβλητος σε γραμμικούς μετασχηματισμούς και έχει σημείο κατάρρευσης ίσο με 50% όταν τα σημεία

δεδομένων μας βρίσκονται σε γενική θέση (ένα σύνολο δεδομένων βρίσκεται σε γενική θέση όταν καμία τριάδα από σημεία δεν είναι συνευθειακά). Ο Rousseeuw περιέγραψε ακόμη έναν εκτιμητή με τις ίδιες ιδιότητες όπως ο παραπάνω ο οποίος προτάθηκε ανεξάρτητα από τον Stahel [Sta81] το 1981 και τον Donoho [Don82] το 1982. Η μέθοδος υπολογισμού του εκτιμητή περιλαμβάνει την εύρεση της προβολής για κάθε σημείο p του συνόλου δεδομένων ξεχωριστά για την οποία το p είναι το πιο απομακρυσμένο από το υπόλοιπο σύνολο δεδομένων. Τελικά το αποτέλεσμα προκύπτει με τον υπολογισμό ενός μέσου κάνοντας χρήση βαρών βασισμένων στα αποτελέσματα των προβολών. Ο Toussaint και ο Poulsen [TP79] πρότειναν το **διαδοχικό κλάδεμα του ελάχιστου επικαλυπτικού δέντρου (minimum spanning tree)** των δοθέντων σημείων ως μία μέθοδο καθορισμού του κέντρου τους. Για την κατασκευή του ελάχιστου επικαλυπτικού δέντρου, ενώνουμε συγκεκριμένα σημεία του συνόλου δεδομένων μας με ακμές έτσι ώστε να υπάρχει ένα μονοπάτι μεταξύ δύο οποιωνδήποτε κόμβων και το άθροισμα των μηκών όλων των ακμών να ελαχιστοποιείται. Αυτή η μέθοδος φαίνεται να δουλεύει καλά στην γενική περίπτωση, ωστόσο ορισμένες ακολουθίες από σημεία δεδομένων μπορεί να οδηγήσουν σε μή «αναμενόμενα» κέντρα του συνόλου των δεδομένων. Ακόμα μία προσέγγιση βασισμένη σε γράφους προτάθηκε από τον Green [Gre81]. Η πρόταση του Green περιλαμβάνει την κατασκευή ενός γράφου συνενώνοντας σημεία από το σύνολο δεδομένων τα οποία είναι γειτονικά στην κατά Delaunay τριγωνοποίηση του συνόλου δεδομένων αυτού. Δύο σημεία {p,q} του συνόλου στον R^2 είναι γειτονικά εάν υπάρχει ένα σημείο στον R^2 για το οποίο τα κοντινότερα σημεία του συνόλου είναι το p και το q. Το βάθος Delaunay ενός σημείου του συνόλου δεδομένων είναι το πλήθος των ακμών στο συντομότερο μονοπάτι από το σημείο αυτό μέχρι κάποιο σημείο του κυρτού περιβλήματος. Γενικά, ως κέντρο ενός γράφου θεωρείται το σύνολο των σημείων για τα οποία τα μέγιστα μήκη μονοπατιού για κάποια άλλα σημεία ελαχιστοποιείται [Har69].

Ακόμα υπάρχουν αρκετοί άλλοι εκτιμητές οι οποίοι βασίζονται περισσότερο σε στατιστικές μεθόδους. Κάποιοι από αυτούς, όπως οι **M-estimators, οι L-estimators και οι R-estimators** περιγράφονται αναλυτικά από τον Huber [Hub72].

1.6. Εφαρμογές Ευσταθών Εκτιμητών.

Οι ευσταθείς εκτιμητές ενός συνόλου δεδομένων έχουν χρησιμοποιηθεί κατά κόρον για περιγραφή δεδομένων (data description), πολυδιάστατες περιοχές εμπιστοσύνης (multivariate confidence regions), p-values, δείκτες ποιότητας (quality indices) και γραφήματα ελέγχου (control charts) [RR96]. Οι εφαρμογές των ορισμών βάθους περιλαμβάνουν έλεγχο υποθέσεων (hypothesis testing), απεικόνιση γραφικών (graphical display), ακόμα και θεωρίες ψηφοφορίας (voting theory) [RR99]. Το βάθος ημιχώρου, ημιεπιπέδου και το τριγωνικό βάθος συνδέονται στενά με την οπισθοδρόμηση (regression). Μία πολύ πρόσφατη μελέτη για τις στατιστικές χρήσεις του βάθους μπορεί να βρεθεί στην αναφορά [LPS99]. Μία πολύ πιο λεπτομερής εισαγωγή στους ευσταθείς εκτιμητές ενός συνόλου δεδομένων γίνεται στην εργασία του Small [Sma90] που αποτελεί σημείο αναφοράς. Τελειώνοντας, συσχετίσεις μεταξύ της υπολογιστικής γεωμετρίας και της στατιστικής έχουνε γίνει από τον Shamos [Sha76] .

ΚΕΦΑΛΑΙΟ 2

ΠΟΛΥΔΙΑΣΤΑΤΟΙ ΕΚΤΙΜΗΤΕΣ.

Σε αυτό το κεφάλαιο δίδουμε την αναλυτική περιγραφή μερικών από τους κυριότερους πολυδιάστατους εκτιμητές (multivariate medians), συμπεριλαμβανομένων και θεμάτων ευστάθειας και υπολογιστικής πολυπλοκότητας.

2.1. L1-μεσοστοιχείο (L1-median)

Ο ορισμός του εκτιμητή που δίνεται παρακάτω έχει προταθεί ανεξάρτητα από πολλούς επιστήμονες και για αυτό είναι γνωστός με πολλά ονόματα στην βιβλιογραφία. Στην εργασία αυτή θα χρησιμοποιούμε τον όρο L1-μεσοστοιχείο που χρησιμοποιείται και από τον Small [Sma90].

Το 1909 ο Weber [Web09] δημοσίευσε την εργασία του *Θεωρία πάνω στην θέση εγκατάστασης των βιομηχανιών*. Ως λύση στο πρόβλημα ελαχιστοποίησης του κόστους μεταφοράς, ο Weber θεώρησε το σημείο που ελαχιστοποιεί το άθροισμα των ευκλείδειων αποστάσεων προς όλα τα σημεία ενός συνόλου δεδομένων. Στον R^1 το μεσοστοιχείο ταυτίζεται με το σημείο αυτό, έτσι η λύση του Weber μπορεί χρησιμοποιηθεί ως γενίκευση του μονοδιάστατου μεσοστοιχείου. Ο εκτιμητής αυτός

του Weber είναι ακριβώς το L1-μεσοστοιχείο. Και ο Weber και ο Pick, ο οποίος έγραψε το μαθηματικό παράρτημα στο βιβλίο του Weber, δεν ήταν σε θέση να βρουν μία λύση για σύνολα δεδομένων με περισσότερα από τρία σημεία στο επίπεδο. Κατά την δεκαετία του 1920, ο ίδιος ορισμός επαναδιατυπώθηκε ανεξάρτητα από πολλούς άλλους ερευνητές, οι οποίοι ενδιαφέρονταν κυρίως για την εύρεση κέντρων οικιστικών πληθυσμών. Ο Eells [Eel30] επισήμανε πως για χρόνια ο αρμόδιος οργανισμός των Ηνωμένων Πολιτειών χρησιμοποιούσε τον μέσο για να υπολογίσει το κέντρο του πληθυσμού της χώρας. Προφανώς νόμιζαν πως ο μέσος δίνει το σημείο που ελαχιστοποιεί το άθροισμα των αποστάσεων των σημείων ενός συνόλου. Το πρόβλημα επισημάνθηκε και από άλλους ανεξάρτητα, όπως αναφέρεται στο [Alo01], και διαπιστώθηκε η ανωτερότητα αυτού του L1-μεσοστοιχείου σε σχέση με τις προσεγγίσεις που δίνει ο μέσος και το διάνυσμα-των-μεσοστοιχείων του Hayford.

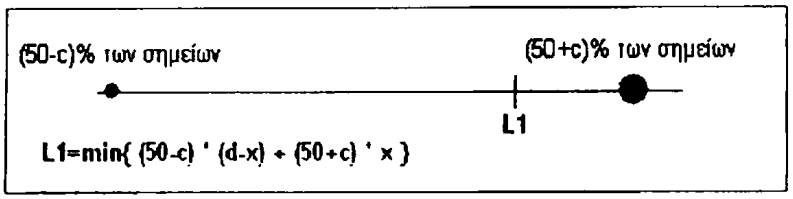
Ορισμός 2.1. *Έστω $\{X_1, X_2, \ldots, X_n\}$ είναι ένα σύνολο από n σημεία στον R^d. Τότε ορίζουμε το L1-μεσοστοιχείο του συνόλου αυτού να είναι το σημείο p του R^d το οποίο ελαχιστοποιεί την παρακάτω ποσότητα:*

$$\sum_{i=1}^{n} \|X_i - p\|$$

Κανένας όμως από τους συγγραφείς που πρότειναν ανεξάρτητα τον παραπάνω ορισμό δεν ήταν σε θέση να προτείνει μία μέθοδο υπολογισμού του L1-μεσοστοιχείου για σύνολα σημείων μεγαλύτερα των τριών σημείων παρόλο που η λύση για τρία σημεία ήτανε γνωστή από τον 17° αιώνα. Σύμφωνα με τους Gross και Strempel [GS98], ο Pierre de Fermat ήταν ο πρώτος που έθεσε το πρόβλημα του υπολογισμού του σημείου με το ελάχιστο άθροισμα των αποστάσεων από τις πλευρές ενός δοθέντος τριγώνου. Ο Fermat συζήτησε το πρόβλημα αυτό με τον Torriceli, ο οποίος αργότερα υπολόγισε την λύση του προβλήματος. Έτσι το L1-μεσοστοιχείο για σύνολα με περισσότερα από 3 σημεία συχνά αναφέρεται ως γενικευμένο Fermat-Torriceli σημείο.

Ο Scates [Sca33] εξέφρασε την άποψη ότι η ακριβής θέση του L1-μεσοστοιχείου για πλήθος σημείων μεγαλύτερο του 3 δεν μπορεί να υπολογιστεί. Ωστόσο, ο Galvani [Gal33] απέδειξε ότι η λύση του προβλήματος είναι ένα μοναδικό σημείο στον R^2 και σε μεγαλύτερες διαστάσεις. Όλοι οι αλγόριθμοι μέχρι και σήμερα περιλαμβάνουν επαναληπτικές μεθόδους ή χρήση παραγώγων και βρίσκουνε μόνο

μία προσέγγιση της πραγματικής λύσης. Οι Gross και Strempel [GS98] μελέτησαν επαναληπτικές και με χρήση παραγώγων μεθόδους για τον υπολογισμό τοῦ L1-μεσοστοιχείου.



Σχήμα 2.1. Το σημείο κατάρρευσης του L1-μεσοστοιχείου είναι το ½.

Είναι γνωστό ότι το L1-μεσοστοιχείο είναι αμετάβλητο σε περιστροφές των δεδομένων σημείων, αλλά όχι και σε αλλαγές κλίμακας. Το σημείο κατάρρευσης του L1-μεσοστοιχείου έχει βρεθεί ότι είναι το ½ [Rou85]. Αυτό είναι προφανές στον R^1 απλά παρατηρώντας ότι αν περισσότερα από τα μισά σημεία βρίσκονται στην ίδια θέση τότε το L1-μεσοστοιχείο θα συνεχίσει να βρίσκεται κοντά σε εκείνη την περιοχή (όπως φαίνεται και στο σχήμα 2.1). Αυτή η παρατήρηση μπορεί να επεκταθεί για οποιαδήποτε φραγμένη περιοχή όπου βρίσκονται τα δοθέντα σημεία. Όταν λιγότερα από τα μισά σημεία μετακινηθούν στο άπειρο, το μεσοστοιχείο παραμένει στη γειτονία της πλειοψηφίας των σημείων του συνόλου δεδομένων, αφού η φραγμένη περιοχή σε σχέση με το άπειρο προσεγγιστικά μπορεί να θεωρηθεί ως σημείο.

2.2. Μεσοστοιχείο Ημιχώρου (Halfspace Median)

Το 1929 ο Hotelling [Sma90] εισήγαγε την δική του εκδοχή του μονοδιάστατου μεσοστοιχείου, ενώ μελετούσε το πρόβλημα δύο ανταγωνιστών πωλητών παγωτών σε μία μονοδιάστατη παραλία. Ο Hotelling ισχυρίστηκε ότι η βέλτιστη λύση για τον πρώτο πωλητή είναι να βρεθεί στην τοποθεσία όπου ελαχιστοποιείται το μέγιστο πλήθος των ανθρώπων σε κάποια πλευρά της παραλίας. Αυτή η τοποθεσία αποδεικνύεται ότι συμπίπτει με την τοποθεσία του μονοδιάστατου μεσοστοιχείου. Ο Tukey [Tuk75] θεωρείται πως είναι ο βασικός συντελεστής για την γενίκευση της παραπάνω ιδέας σε μεγαλύτερες διαστάσεις. Το πολυδιάστατο μεσοστοιχείο που βασίστηκε σε αυτήν την ιδέα συχνά αναφέρεται στην βιβλιογραφία ως το μεσοστοιχείο Tukey ή το μεσοστοιχείο ημιχώρου. Οι Donoho και Gasko

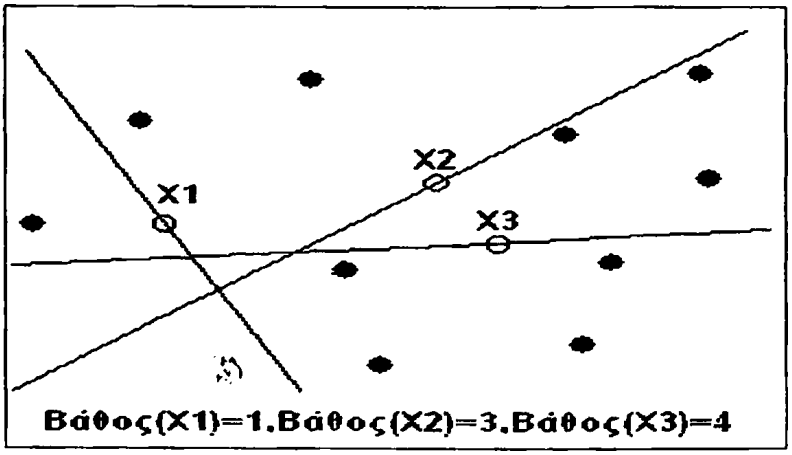
[DG92] είναι πραγματικά αυτοί που πρώτοι καθόρισαν το πολυδιάστατο μεσοστοιχείο ημιχώρου, αφού ο Tukey κύρια ασχολήθηκε με το βάθος ημιχώρου.

Για να μπορέσουμε να ορίσουμε το πολυδιάστατο μεσοστοιχείο ημιχώρου είναι πιο εύκολο να προσδιορίσουμε την έννοια του βάθους ημιχώρου για ένα σημείο p ως προς ένα δοθέν σύνολο από n σημεία στον R^d. Για κάθε κλειστό ημιχώρο ο οποίος περιέχει το p, μετράμε το πλήθος των σημείων του συνόλου δεδομένων που βρίσκονται μέσα στον ημιχώρο αυτόν. Τελικά θεωρούμε ως βάθος ημιχώρου του σημείου p το ελάχιστο πλήθος που βρήκαμε (για όλους τους δυνατούς ημιχώρους). Για παράδειγμα, στον R^2, τοποθετούμε μία γραμμή που να περνάει από το σημείο p έτσι ώστε το πλήθος των σημείων στην μία πλευρά της ευθείας να ελαχιστοποιείται. Το μεσοστοιχείο ημιχώρου ενός συνόλου από σημεία είναι το σημείο (ή τα σημεία) στον R^d με το μέγιστο βάθος ημιχώρου.

Ορισμός 2.2. *Έστω $\{X_1, X_2, ..., X_n\}$ ένα σύνολο από n σημεία στον R^d. Έστω ακόμα ότι H είναι η κλάση όλων των κλειστών ημιχώρων στον R^d. Τότε ορίζουμε την γενικευμένη στον R^d συνάρτηση κατανομής :*

$$F(H_k) = \frac{\left\{\sum_{i=1}^{n} I[X_i \in H_k]\right\}}{n}$$ *για κάθε H_k που ανήκει στο H.*

Σύμφωνα με τον Tukey, το βάθος D(p) ενός σημείου p στον R^d μέσα στο παραπάνω σύνολο n σημείων ορίζεται ως η ελάχιστη τιμή της $F(H_k)$, με το ελάχιστο να υπολογίζεται ως προς το σύνολο των H_k στα οποία ανήκει το p. Το μεσοστοιχείο ημιχώρου ενός συνόλου από σημεία είναι το σημείο (ή τα σημεία) στον R^d με το μέγιστο βάθος ημιχώρου.



Σχήμα 2.2. Τα βάθη ημιχώρου κάποιων σημείων στο επίπεδο.

Το μεσοστοιχείο ημιχώρου δεν είναι γενικά μοναδικό σημείο. Ωστόσο το σύνολο των σημείων που έχουν το μέγιστο βάθος είναι εγγυημένα ένα κλειστό και φραγμένο κυρτό σύνολο. Το μεσοστοιχείο είναι αμετάβλητο σε γραμμικούς μετασχηματισμούς και το σημείο κατάρρευσης του μπορεί να κυμαίνεται μεταξύ $\frac{1}{d+1}$ και $\frac{1}{3}$. Αν το σύνολο δεδομένων είναι σε γενική θέση, το μέγιστο βάθος ημιχώρου φράσσεται κάτω από το $\left\lceil \frac{n}{d+1} \right\rceil$ και άνω από το $\left\lceil \frac{n}{2} \right\rceil$, όπως αναφέρεται στο [DG92].

Παρατηρούμε εύκολα πως κάθε σημείο που βρίσκεται έξω από το κυρτό περίβλημα του συνόλου των δεδομένων έχει βάθος που ισούται με το μηδέν. Για να βρούμε την περιοχή με το μέγιστο βάθος στον R^2 μπορούμε να χρησιμοποιήσουμε το γεγονός ότι το σύνορό της αποτελείται από τμήματα των ευθειών που περνούν από ζεύγη σημείων του συνόλου των δεδομένων. Με άλλα λόγια, οι κορυφές της επιθυμητής περιοχής πρέπει να είναι σημεία τομής των ευθειών που ορίζονται από ζεύγη των σημείων του συνόλου των δεδομένων. Υπάρχουν $O(n^4)$ δυνατά σημεία τομής και είναι προφανής η μέθοδος εύρεσης του βάθους ενός σημείου ενός σημείου p σε $O(n^2)$ χρόνο (απλά για κάθε ευθεία που ορίζεται από το σημείο p και ένα σημείο του συνόλου των δεδομένων μετράμε το πλήθος των σημείων του συνόλου των δεδομένων που βρίσκονται από την μία και την άλλη μεριά της ευθείας). Έτσι σε $O(n^6)$ χρόνο μπορούμε να βρούμε το σύνολο των σημείων τομής με το μέγιστο βάθος.

Μία βελτίωση στον παραπάνω χρόνο έγινε από τους Rousseeuw και Ruts [RR96], οι οποίοι έδειξαν πως μπορούμε να υπολογίσουμε το βάθος ημιχώρου ενός σημείου σε $O(n \log n)$ χρόνο. Με τον αλγόριθμό τους το σημείο (ή το σύνολο σημείων) με μέγιστο βάθος μπορεί να υπολογιστεί σε $O(n^5 \log n)$ χρόνο. Αργότερα έδωσαν ένα πιο πολύπλοκο αλγόριθμο σε $O(n^2 \log n)$ χρόνο, μαζί με την υλοποίηση [RR98]. Προφανώς δεν έλαβαν υπόψη τους το γεγονός ότι ο Matousek [Mat91] είχε παρουσιάσει έναν αλγόριθμο $O(n \log^5 n)$ για τον υπολογισμό του μεσοστοιχείου ημιχώρου. Ο ίδιος πρότεινε έναν αλγόριθμο με τον οποίον είναι δυνατόν να υπολογίσουμε οποιοδήποτε σημείο με βάθος μεγαλύτερο από κάποια σταθερά k σε $O(n \log^4 n)$ χρόνο και μετά χρησιμοποιούσε δυαδική αναζήτηση στο k για να βρει το μεσοστοιχείο ημιχώρου. Πρόσφατα ο αλγόριθμος του Matousek βελτιώθηκε από τους Langerman και Steiger [LS00], ο αλγόριθμος των οποίων υπολογίζει το μεσοστοιχείο

ημιχώρου σε $O(n \log^4 n)$ χρόνο. Μία πιο πρόσφατη εργασία από τους Miller et al. [MRR01] σε $O(n^2)$ χρόνο και χώρο βρίσκει όλα τα σημεία ίδιου βάθους (depth contours) και μετά την εύρεση των οποίων είναι πολύ εύκολο να υπολογίσουμε το μεσοστοιχείο ημιχώρου. Στην εργασία τους υποστηρίζουν πως ο αλγόριθμος του Matousek είναι πολύ πολύπλοκος για πρακτικές εφαρμογές.

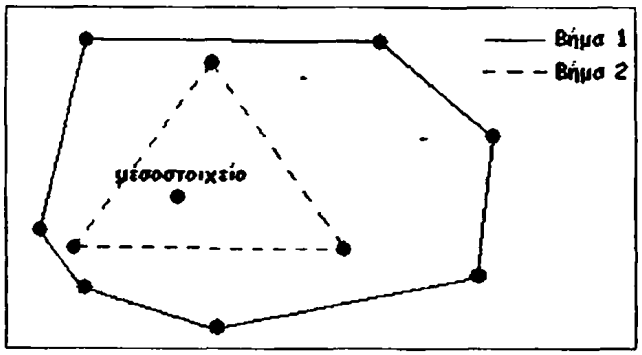
Άλλη ενδιαφέρουσα έννοια που βασίζεται στο βάθος ημιχώρου είναι το κεντρικό σημείο (centerpoint), το οποίο είναι το σημείο με βάθος τουλάχιστον $\left\lceil \frac{n}{d+1} \right\rceil$. Οι Gill, Steiger και Wigderson [GSW92] υποστήριξαν πως το κεντρικό σημείο μπορεί να χρησιμοποιηθεί ως πολυδιάστατος εκτιμητής αφού συμπίπτει με το μεσοστοιχείο στον R^1. Ο Edelsbrunner [Ede87] έδειξε ότι το κεντρικό σημείο μπορεί να βρεθεί για κάθε σύνολο δεδομένων. Ο αριθμός $\left\lceil \frac{n}{d+1} \right\rceil$ προκύπτει από το θεώρημα του Helly (περισσότερα στο [RH99a]). Οι Cole, Shahir και Yap μπόρεσαν να υπολογίσουν το κεντρικό σημείο σε $O(n \log^5 n)$ χρόνο. Ο Matousek βελτίωσε τον παραπάνω χρόνο θέτοντας το k ίσο με $\left\lceil \frac{n}{d+1} \right\rceil$ στον αλγόριθμό του και έτσι μπόρεσε να βρει το κεντρικό σημείο σε $O(n \log^4 n)$ χρόνο. Τελικά οι Jadhav και Mukhopadhyay [JM94] έδωσαν ένα αλγόριθμο $O(n)$ χρόνου για τον υπολογισμό ενός κεντρικού σημείου.

2.3. Αφαίρεση Κυρτού Περιβλήματος και συσχετιζόμενες μέθοδοι.

Ένας ενδιαφέρων τρόπος περιγραφής του μονοδιάστατου μεσοστοιχείου είναι μέσω της απομάκρυνσης των σημείων του συνόλου των δεδομένων που βρίσκονται στο εξωτερικό μέρος του συνόλου. Αγνοούμε δηλαδή την μικρότερη και την μεγαλύτερη τιμή επαναληπτικά μέχρι να μείνουμε με ένα ή δυο σημεία. Την παραπάνω ιδέα μπορούμε εύκολα να την επεκτείνουμε και σε μεγαλύτερες διαστάσεις. Ένας τρόπος είναι η επαναληπτική απομάκρυνση του κυρτού περιβλήματος του δοθέντος συνόλου σημείων, μέχρι να απομείνει ένα κυρτό σύνολο σημείων (όπως φαίνεται στο σχήμα 2.3). Όπως και στην περίπτωση του μονοδιάστατου μεσοστοιχείου, εάν μείνουν στο τέλος παραπάνω από ένα σημεία, υπολογίζουμε τον μέσο των σημείων αυτών. Αυτοί που μελέτησαν περισσότερο τον

παραπάνω πολυδιάστατο εκτιμητή είναι οι Shamos [Sha76] και Barnett [Bar76]. Ωστόσο, όπως σημειώνει ο Shamos, η ιδέα ανήκει αρχικά στον Tukey.

Ο πιο απλοϊκός αλγόριθμος για την αφαίρεση του κυρτού περιβλήματος απαιτεί $O(n^2 \log n)$ χρόνο στον R^2. Καθώς το κυρτό περίβλημα ενός συνόλου n σημείων στον R^2 μπορεί να υπολογιστεί σε $O(n \log n)$ χρόνο. Αντίστοιχα κάτω φράγματα έχουν υπολογιστεί σε διάφορα υπολογιστικά μοντέλα (όπως για παράδειγμα στο [Avi82]). Στην χειρότερη περίπτωση είναι πιθανό μόνο τρία σημεία δεδομένων να απομακρυνθούν σε κάθε επανάληψη, γεγονός που μας οδηγεί σε $O(n)$ υπολογισμούς κυρτού περιβλήματος. Ο ίδιος υπολογισμός μπορεί να γίνει σε $O(n^2)$ χρόνο απλά διαφοροποιώντας ελάχιστα τον αλγόριθμο περιτυλίγματος (gift-wrapping) του Jarvis [Jar73]. Αυτός ο αλγόριθμος χρειάζεται $O(h\ n)$ χρόνο για τον υπολογισμό του κυρτού περιβλήματος, όπου h είναι το πλήθος των σημείων του κυρτού περιβλήματος που επιστρέφει ο αλγόριθμος. Μετά τον υπολογισμό του κυρτού περιβλήματος, ο τροποποιημένος αλγόριθμος συνεχίζει με τον υπολογισμό του κυρτού περιβλήματος των εναπομεινάντων σημείων. Αυτή η τροποποίηση προτάθηκε πρώτα από τον Shamos [Sha76].



Σχήμα 2.3. Το διδιάστατο μεσοστοιχείο με χρήση επαναληπτικής αφαίρεσης κυρτού περιβλήματος

Αργότερα, οι Overmars και Van Leeuwen [OVL81] σχεδίασαν μία δομή δεδομένων η οποία κρατάει το κυρτό περίβλημα ενός συνόλου σημείων μετά την εισαγωγή ή την διαγραφή τυχαίων σημείων, με κόστος $O(\log^2 n)$ χρόνο για κάθε εισαγωγή ή διαγραφή. Αυτή η δομή άμεσα συνεπάγεται έναν αλγόριθμο $O(n \log^2 n)$ χρόνου για την επαναληπτική αφαίρεση του κυρτού περιβλήματος. Τελικά ο Chazelle [Cha85] βελτίωσε το παραπάνω αποτέλεσμα αγνοώντας τις εισαγωγές σημείων και λαμβάνοντας υπόψη του το πλεονέκτημα της παραπάνω δομής δεδομένων σε

διαδοχικές διαγραφές στοιχείων που λαμβάνουν χώρα κατά την εφαρμογή του αλγορίθμου. Ο αλγόριθμος του Chazelle χρειάζεται O(n log n) χρόνο για να υπολογίσει όλα τα επίπεδα κυρτών περιβλημάτων και τα βάθη κάθε σημείου του συνόλου των δεδομένων.

Μία τεχνική παρόμοια με αυτή της Αφαίρεσης του Κυρτού Περιβλήματος προτάθηκε από τον Titterington [Tit78]. Η πρόταση του περιλάμβανε την **επαναληπτική αφαίρεση του ελλειψοειδούς ελάχιστου όγκου (minimum volume ellipsoid peeling)**. Και οι δύο παραπάνω μέθοδοι αφαίρεσης δεδομένων έχουν πολύ χαμηλά σημεία κατάρρευσης. Οι Donoho και Gasko [DG92] απέδειξαν πως τα σημεία κατάρρευσης των παραπάνω μεθόδων δεν μπορούν να ξεπεράσουν το $\frac{1}{d+1}$ στον R^d και μάλιστα υποστήριξαν ότι το σημείο κατάρρευσης φαίνεται να τείνει προς το μηδέν καθώς το πλήθος των σημείων του συνόλου πλησιάζει το άπειρο. Υπάρχουν πολλά παραδείγματα που μπορούν να μας δείξουνε γιατί το σημείο κατάρρευσης πλησιάζει το μηδέν για ορισμένες κατανομές σημείων. Μία από αυτές φαίνεται στο σχήμα 2.4 όπου υπάρχει μόνο ένα σημείο στο εσωτερικό του κυρτού περιβλήματος και επομένως αποτελεί το μεσοστοιχείο. Τότε το σημείο αυτό παραμένει το μεσοστοιχείο ακόμη και αν μετακινηθεί οσοδήποτε μακριά προς τα δεξιά αρκεί να μετακινηθεί επίσης προς τα δεξιά κατά την ίδια απόσταση και το δεξιότερο σημείο (όπως φαίνεται στο σχήμα 2.4). Δηλαδή, η θέση του μεσοστοιχείου επηρεάζεται από πολύ λίγα σημεία, γεγονός που συνεπάγεται ότι δεν είναι ιδιαίτερα ευσταθής.

Σχήμα 2.4. Παράδειγμα που επιδεικνύει πως το σημείο κατάρρευσης της επαναληπτικής αφαίρεσης κυρτού περιβλήματος μπορεί να πλησιάζει το 0.

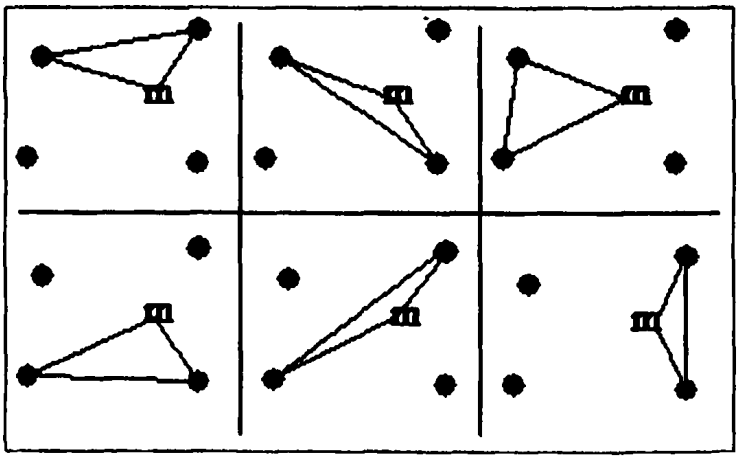
Μία άλλη μέθοδος βασισμένη σε ελλειψοειδές προτάθηκε από τον Rousseeuw [Rou85]. Ο εκτιμητής που πρότεινε βρίσκεται στο κέντρο του ελλειψοειδούς ελάχιστου όγκου το οποίο καλύπτει περίπου τα μισά σημεία του συνόλου των δεδομένων μας. Ο Rousseeuw απέδειξε ότι αυτός ο εκτιμητής είναι αμετάβλητος σε

γραμμικούς μετασχηματισμούς και έχει σημείο κατάρρευσης ίσο με 50% για δεδομένα σε γενική θέση.

2.4. Μεσοστοιχείο Στοιχειωδών d-διάστατων Αντικειμένων Πλήρους Διάστασης του Oja (Oja Simplicial Median).

Έστω ότι έχουμε d+1 σημεία σε γενική θέση στον R^d . Αυτά τα σημεία σχηματίζουν ένα στοιχειώδες d-διάστατο αντικείμενο πλήρους διάστασης (simplex). Για παράδειγμα, στον R^3 τέσσερα σημεία σχηματίζουν ένα τετράεδρο και στον R^2 τρία σημεία σχηματίζουνε ένα τρίγωνο. Έστω τώρα ότι έχουμε ένα σύνολο δεδομένων στον R^d για το οποίο αναζητούμε έναν εκτιμητή. Ο Oja [Oja83] πρότεινε το ακόλουθο μέτρο βάθους στον R^d για ένα σημείο p που δεν ανήκει αναγκαστικά στο σύνολο δεδομένων:

- Για κάθε υποσύνολο d σημείων από το σύνολο δεδομένων, σχημάτισε ένα στοιχειώδες d-διάστατο αντικείμενο πλήρους διάστασης μαζί με το σημείο p.
- Άθροισε το σύνολο των όγκων όλων των δυνατών στοιχειωδών d-διάστατων αντικειμένων πλήρους διάστασης του σημείου p.



Σχήμα 2.5. Τα στοιχειώδη διδιάστατα αντικείμενα πλήρους διάστασης στον R^2 το εμβαδόν των οποίων συνυπολογίζουμε για το μεσοστοιχείο Oja (για n = 4).

Το παραπάνω άθροισμα που προκύπτει ονομάζεται **βάθος Oja**. Το μεσοστοιχείο d-διάστατων αντικειμένων πλήρους διάστασης του Oja είναι κάθε σημείο p το οποίο στον R^d έχει το ελάχιστο βάθος Oja.

Ορισμός 2.3. *Σε ένα σύνολο $\{X_1, X_2, ..., X_n\}$ από n σημεία στον R^d ορίζουμε την συνάρτηση $c[Xi_1, Xi_2, ..., X_{id}; m]$ να είναι ο d-διάστατος όγκος του στοιχειώδους d-διάστατου αντικείμενου πλήρους διάστασης στον R^d του οποίου οι κορυφές είναι τα $Xi_1, Xi_2, ..., X_{id}$ και το m, με $i1 < i2 < ... < id$. Τότε το μεσοστοιχείο d-διάστατων αντικειμένων πλήρους διάστασης του Oja για το παραπάνω σύνολο σημείων είναι το σημείο p που ελαχιστοποιεί την ποσότητα:*

$$\sum_{i1<i2<...<id} c[X_{i1}, X_{i2}, ..., X_{id}; m]$$

όπου το άθροισμα λαμβάνεται για όλα τα υποσύνολα των ακεραίων δεικτών της μορφής $1 \le i1 < ... < id \le n$.

Θεωρώντας ότι ο μονοδιάστατος όγκος είναι το μήκος, το μεσοστοιχείο d-διάστατων αντικειμένων πλήρους διάστασης του Oja στον R^1 συμπίπτει με το μονοδιάστατο μεσοστοιχείο. Δηλαδή ελαχιστοποιεί το άθροισμα των αποστάσεων προς όλα τα σημεία του συνόλου των δεδομένων. Σε αντίθεση πάντως με το L1-μεσοστοιχείο, το μεσοστοιχείο d-διάστατων αντικειμένων πλήρους διάστασης του Oja δεν εγγυάται ότι θα είναι ένα μοναδικό σημείο σε μεγαλύτερες διαστάσεις. Ωστόσο ο Oja αναφέρει (χωρίς να το αποδεικνύει πάντως) πως τα σημεία που έχουνε το μέγιστο βάθος σχηματίζουν ένα κυρτό σύνολο και για να υπολογίσουμε αυτά τα σημεία στον R^2 αρκεί να λάβουμε υπόψη μας μόνο τις τομές των ευθειών που σχηματίζονται από ζεύγη σημείων του συνόλου των δεδομένων. Την απόδειξη των παραπάνω ιδιοτήτων μας την δίνουν οι Aloupis, Langerman, Soss και Toussaint στο [ALST01].

Ένα σπουδαίο χαρακτηριστικό του μεσοστοιχείου του Oja είναι το γεγονός ότι είναι αμετάβλητο σε γραμμικούς μετασχηματισμούς. Ωστόσο μπορούμε να κατασκευάσουμε σύνολα δεδομένων για τα οποία το σημείο κατάρρευσης του μεσοστοιχείου αυτού να πλησιάζει το μηδέν. Αυτό είναι εύκολο να το δούμε παρατηρώντας ότι για σύνολα δεδομένων με σημεία που δεν «απλώνονται» σε κάποια από τις διαστάσεις του χώρου που βρίσκονται (όπως δεδομένα σε μία ευθεία στον R^2

ή δεδομένα πάνω σε κάποιο επίπεδο στον R^3) είναι απλό να βρούμε d-διάστατα αντικείμενα πλήρους διάστασης με μηδενικό όγκο ακόμα και αν εκτείνονται μέχρι το άπειρο. Γενικά όμως στην πράξη είναι σχεδόν απίθανο να συναντήσουμε τέτοια σύνολα σε κάποια εφαρμογή.

Η πιο απλή μέθοδος υπολογισμού του μεσοστοιχείο d-διάστατων αντικειμένων πλήρους διάστασης του Oja είναι η εύρεση του βάθους κάθε σημείου τομής ξεχωριστά. Στον R^2 αυτό μπορεί να γίνει σε $O(n^6)$ χρόνο: για καθένα από τα $O(n^4)$ σημεία τομής των ευθειών που δημιουργούνται από ζεύγη σημείων του συνόλου των δεδομένων, υπάρχουν $O(n^2)$ τρίγωνα το εμβαδόν των οποίων πρέπει να υπολογίσουμε και κάθε τέτοιο εμβαδόν υπολογίζεται σε σταθερό χρόνο. Το ίδιο άνω φράγμα έχει και ο αλγόριθμος των Niinimaa, Oja και Nyblom [NON92]. Ο αλγόριθμος τους επιλέγει μία ευθεία που ορίζεται από δύο σημεία του συνόλου των δεδομένων και υπολογίζει την παράγωγο του βάθους Oja για κάθε σημείο τομής πάνω στην ευθεία μέχρι να βρεθεί ένα σημείο που να έχει την ελάχιστη τιμή. Τότε επιλέγεται μία νέα ευθεία που περνάει από αυτό το σημείο και επαναλαμβάνεται η ίδια διαδικασία. Ο υπολογισμός της παραγώγου γίνεται σε $O(n^2)$ χρόνο και είναι δυνατόν να επισκεφτούν όλα τα σημεία τομής. Οι Rousseeuw και Ruts [RR96] χρησιμοποιούν μία τεχνική για τον υπολογισμό της εφαπτομένης Oja σε $O(n \log n)$ χρόνο και υποστηρίζουν πως με την ίδια τεχνική μπορούν να βρούν το μεσοστοιχείο Oja σε $O(n^5 \log n)$ χρόνο. Τελικά οι Aloupis, Langerman, Soss και Toussaint [ALST01] μας δίνουν έναν αλγόριθμο υπολογισμού του μεσοστοιχείο του Oja σε $O(n \log^3 n)$ χρόνο.

2.5. Τριγωνικός Εκτιμητής (Simplicial Median)

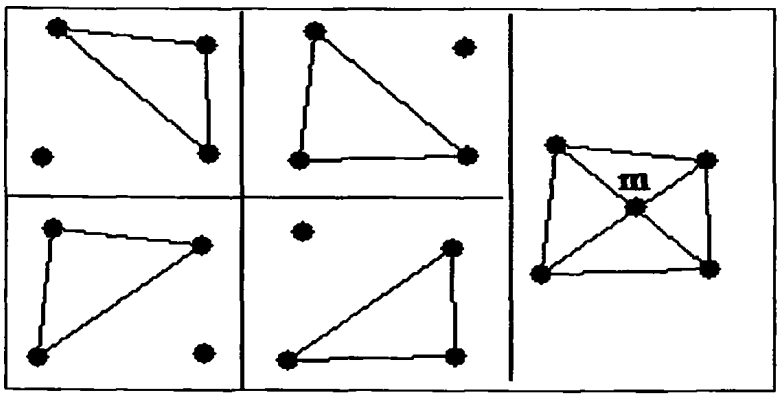
Μία ακόμα επέκταση της ιδέας του μονοδιάστατου μεσοστοιχείου στον πολυδιάστατο χώρο βασίζεται στο ότι το μονοδιάστατο μεσοστοιχείο βρίσκεται μέσα στο μεγαλύτερο πλήθος διαστημάτων στον R^1 που δημιουργούνται από σημεία του συνόλου των δεδομένων. Η Regina Liu γενίκευσε αυτή την ιδέα [Liu90]: ο τριγωνικός εκτιμητής στον R^d είναι ένα σημείο του χώρου αυτού το οποίο περιέχεται μέσα στα περισσότερα στοιχειώδη d-διάστατα αντικείμενα πλήρους διάστασης τα οποία σχηματίζονται από υποσύνολα d+1 σημείων του συνόλου των δεδομένων.

Ένα παράδειγμα φαίνεται στο σχήμα 2.6 για ένα σύνολο 4 σημείων στον R^2. Το τριγωνικό βάθος (simplicial depth) ενός σημείου στον R^d είναι το πλήθος των στοιχειωδών d-διάστατων αντικειμένων πλήρους διάστασης που περιέχουν το σημείο αυτό. Ο ακριβής ορισμός της Liu περιλαμβάνει κλειστά στοιχειώδη d-διάστατα αντικείμενα πλήρους διάστασης, παρόλο που αργότερα στο [Liu95] επαναλαμβάνει τον παραπάνω ορισμό χρησιμοποιώντας ανοιχτά στοιχειώδη d-διάστατα αντικείμενα πλήρους διάστασης. Πάντως όλες οι αναφορές μας στον τριγωνικό εκτιμητή λαμβάνουν υπόψη τους τον ορισμό με τα κλειστά στοιχειώδη d-διάστατα αντικείμενα πλήρους διάστασης (δηλαδή, θεωρούμε ότι ένα σημείο που βρίσκεται στο σύνορο του στοιχειώδη d-διάστατου αντικειμένου πλήρους διάστασης βρίσκεται μέσα στο αντικείμενο αυτό).

Ορισμός 2.4. *Σε ένα σύνολο $\{X_1, X_2, ..., X_n\}$ από n σημεία στον R^d ορίζουμε την συνάρτηση τριγωνικού βάθους ενός σημείου m ως προς το παραπάνω σύνολο σημείων ως εξής:*

$$SDF(m)=\binom{n}{d+1}^{-1}\sum_{1\le i_1<..<i_{d+1}\le n} I[m\in S(X_{i_1},...,X_{i_{d+1}})]$$

όπου $S(X_1, X_2, ..., X_{d+1})$ είναι το στοιχειώδες d-διάστατο αντικείμενο πλήρους διάστασης με κορυφές τις $X_1, X_2, ..., X_{d+1}$. Τότε ορίζουμε τον τριγωνικό εκτιμητή ως το σημείο (ή τα σημεία) όπου μεγιστοποιείται η συνάρτηση SDF(m).

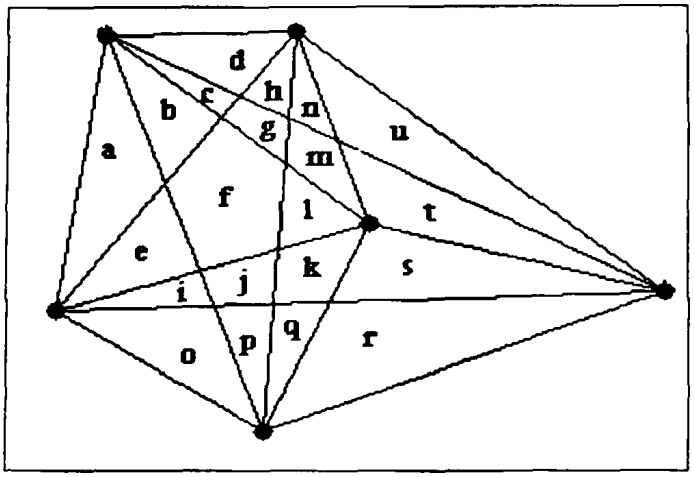


Σχήμα 2.6. Τα στοιχειώδη διδιάστατα αντικείμενα πλήρους διάστασης που λαμβάνουμε υπόψη για τον υπολογισμό του τριγωνικού εκτιμητή και η θέση του μεσοστοιχείου **m**.

Η Liu έδειξε πως ο τριγωνικός εκτιμητής είναι αμετάβλητος σε γραμμικούς μετασχηματισμούς. Ωστόσο δεν γνωρίζουμε και πολλά για το σημείο κατάρρευσης του εκτιμητή αυτού. Οι Gil, Steiger και Wigderson [GSW92] κατασκεύασαν ένα σύνολο σημείων για το οποίο το σημείο δεδομένων με το μέγιστο τριγωνικό βάθος μπορούσε να μετακινηθεί σε τυχαία κατεύθυνση με μεταβολή της θέσης λίγων μόνο σημείων του συνόλου δεδομένων. Όμως το παραπάνω αποτέλεσμα δεν αποδεικνύει τίποτα για τον εκτιμητή της Liu.

Παρόλο που η Liu ήταν η πρώτη που όρισε το τριγωνικό βάθος και εισήγαγε την έννοια στην επιστημονική κοινότητα, αυτή η ιδέα δεν ήταν καινούργια. Οι Boros και Furedi [BF84] απέδειξαν ότι για ένα σύνολο n σημείων σε γενική θέση στον R^2 υπάρχει πάντα ένα σημείο που περιέχεται σε τουλάχιστον $\left\{\frac{n^3}{27}+O\left(n^2\right)\right\}$ ανοιχτά τρίγωνα που δημιουργούνται από τα σημεία του συνόλου. Το παραπάνω αποτέλεσμα συνεπάγεται ότι το βάθος του τριγωνικού εκτιμητή είναι τουλάχιστον της τάξης του n^3. Ο Barany [Bar82] έδειξε πως στον R^d υπάρχει τουλάχιστον ένα σημείο το οποίο περιέχεται μέσα σε $\frac{1}{(d+1)^{d+1}}\binom{n}{d+1}+O\left(n^d\right)$ στοιχειώδη d-διάστατα αντικείμενα πλήρους διάστασης.

Μία απλή μέθοδος υπολογισμού του τριγωνικού εκτιμητή στον R^2 είναι η διαμέριση του επιπέδου σε περιοχές με σύνορα ευθύγραμμα τμήματα που σχηματίζονται από τα σημεία του συνόλου των δεδομένων. Σαν πρώτη παρατήρηση μπορούμε να δούμε ότι κάθε σημείο μέσα σε μία περιοχή έχει το ίδιο τριγωνικό βάθος. Ακόμα, ένα σημείο πάνω στα όρια μιας περιοχής πρέπει να έχει τιμή βάθους τουλάχιστον όσο και ένα σημείο που βρίσκεται στο εσωτερικό της γειτονικής αυτής περιοχής. Παρόμοια, ένα σημείο τομής (όπου περισσότερες από δύο περιοχές συναντιούνται) πρέπει να έχει τιμή τριγωνικού βάθους τουλάχιστον ίση με κάθε σημείο που βρίσκεται σε σύνορο γειτονικής περιοχής. Έτσι από τα παραπάνω συμπεραίνουμε πως ο τριγωνικός εκτιμητής βρίσκεται σίγουρα πάνω σε κάποιο σημείο τομής και πως καθορίζοντας το πλήθος των τριγώνων μέσα στα οποία περιέχεται κάθε σημείο τομής μπορούμε να βρούμε το τριγωνικό εκτιμητή.



Σχήμα 2.7. Τα ευθύγραμμα τμήματα με άκρα τα σημεία του συνόλου των δεδομένων χωρίζουν το κυρτό περίβλημα των σημείων στις περιοχές a-u .

Εάν ένα σύνολο από n ευθύγραμμα τμήματα ορίζει k σημεία τομής, τότε αυτά μπορούν να βρεθούν σε O(n log n + k) χρόνο και O(n) χώρο με την τεχνική της σάρωσης γραμμής (line sweeping) του Balaban [Bal95]. Στην περίπτωση του τριγωνικού βάθους, έχουμε $O(n^2)$ ευθύγραμμα τμήματα τα οποία σχηματίζονται από ζευγάρια σημείων του συνόλου των δεδομένων και δυστυχώς το k στην περίπτωση αυτή είναι $\Theta(n^4)$. Έτσι ο αλγόριθμος του Balaban απαιτεί $O(n^4)$ χρόνο και $O(n^2)$ χώρο, έτσι δεν υπάρχει διαφορά με τον προφανή τρόπο υπολογισμού όλων των σημείων τομής στην χειρότερη περίπτωση. Ο συνολικός χρόνος για αυτόν τον σκοπό είναι επίσης $O(n^4)$ και ο χώρος που απαιτείται είναι O(n). Εφόσον υπάρχουν συνολικά $O(n^3)$ τρίγωνα που σχηματίζονται από n σημεία, τότε ο υπολογισμός του τριγωνικού εκτιμητή με τον προφανή τρόπο απαιτεί $O(n^7)$ χρόνο και O(n) χώρο.

Οι Khuller και Mitchell [KM89] πρότειναν έναν O(n log n) αλγόριθμο για τον υπολογισμό των τριγώνων που σχηματίζονται από τριάδες σημείων του συνόλου δεδομένων που περιέχουν ένα σημείο p στον R^2 . Οι Gil, Steiger και Wigderson [GSW92] ανεξάρτητα πρότειναν τον ίδιο αλγόριθμο και θεώρησαν πως ο τριγωνικός εκτιμητής είναι το σημείο από το σύνολο των δεδομένων με το μέγιστο τριγωνικό βάθος. Μία τρίτη έκδοση του ίδιου αλγόριθμου εμφανίστηκε αργότερα, αλλά πάλι ανεξάρτητα από τους Rousseeuw και Ruts [RR96].

Οι Rousseeuw και Ruts ήταν οι πρώτοι που διαπίστωσαν ότι με τον υπολογισμό του βάθους κάθε σημείου τομής, ο τριγωνικός εκτιμητής μπορεί να βρεθεί σε $O(n^5 \log n)$ χρόνο. Τελικά οι Aloupis, Langerman, Soss και Toussaint [ALST01] πρότειναν έναν αλγόριθμο υπολογισμού του τριγωνικού εκτιμητή που

απαιτεί $O(n^4 \log n)$ χρόνο και $O(n^2)$ χώρο ή $O(n^4)$ χώρο και χρόνο. Ακόμα πρότειναν και χρήση κάποιων τεχνικών που βελτιώνουν τον απαιτούμενο χώρο σε $O(n^2)$ και βελτιώνουν την συνολική χρονική πολυπλοκότητα κατά κάποια σταθερά.

Στον R^3, οι Gil, Steiger και Wigderson [GSW92] πρότειναν έναν αλγόριθμο υπολογισμού του τριγωνικού βάθους ενός σημείου σε $O(n^2)$ χρόνο. Οι Cheng και Ouyang [CO98] ανακάλυψαν ένα μικρό σφάλμα στον παραπάνω αλγόριθμο και έδωσαν την διορθωμένη έκδοσή του. Οι ίδιοι πρότειναν έναν αλγόριθμο $O(n^4)$ χρόνου στον R^4 και επεσήμαναν πως στις μεγαλύτερες διαστάσεις ο αλγόριθμος που βασίζεται στον προφανή τρόπο υπολογισμού έχει καλύτερες επιδόσεις. Τέλος αναφέρουν ότι ο αλγόριθμος που προτάθηκε από τους Rousseeuw και Ruts [RR96] για μεγαλύτερες διαστάσεις φαίνεται να εμφανίζει κάποιες αποκλίσεις (discrepancies).

2.6. Βάθος Υπερεπιπέδου (Hyperplane Depth) και Ορισμός ενός Εκτιμητή βασισμένου σε αυτό.

Σχετικά πρόσφατα οι Rousseeuw και Hubert [RH99] πρότειναν μία καινούργια και πολύ ενδιαφέρουσα απόδοση της ιδέας του βάθους στον R^d.

Ορισμός 2.5. *Δεδομένου ενός συνόλου $S = \{h_1, \ldots, h_n\}$ από n υπερεπίπεδα (hyperplanes) στον R^d, το βάθος υπερεπιπέδου (hyperplane depth) ενός σημείου p του R^d ορίζεται να είναι :*

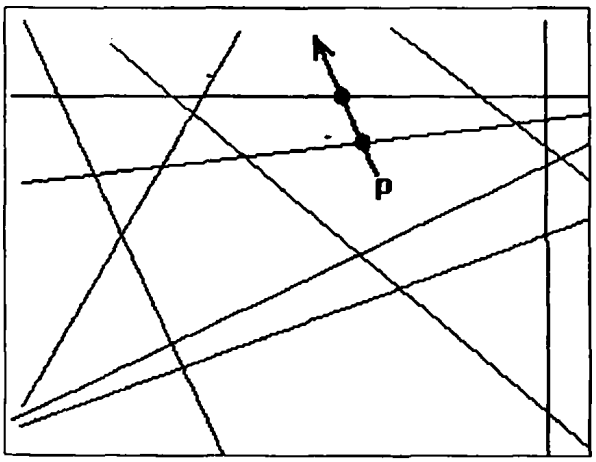
$$\delta(p) = \min_{u:\|u\|=1} (r(p,u))$$

όπου r(p,u) είναι το πλήθος των υπερεπιπέδων h_j του S που συναντούν την ακτίνα $\{p + tu , t \geq 0\}$ που περνάει από το p στην κατεύθυνση του u. Το μεσοστοιχείο είναι το σημείο που έχει το μέγιστο βάθος.

Όταν είμαστε στην μονοδιάστατη περίπτωση η παραπάνω εξίσωση του δ(p) συμφωνεί με τον συνήθη ορισμό του βάθους. Ο ορισμός του βάθους υπερεπιπέδου προέκυψε μετά από προβλήματα στην ευσταθή οπισθοδρόμηση (robust regression). Χρησιμοποιώντας τον γνωστό δυϊκό μετασχηματισμό σημείου / υπερεπιπέδου, τα

υπερεπίπεδα h_j αντιστοιχίζονται σε n σημεία δεδομένων στον R^d, ενώ το σημείο p του R^d αντιστοιχίζεται σε ένα υπερεπίπεδο. Το βάθος οπισθοδρόμησης (regression depth) του σημείου p υπολογίζεται από το ελάχιστο πλήθος από σημεία δεδομένων που το p συναντάει σε μία **περιστροφή-προς-κατακόρυφη-θέση (rotation-to-vertical)**, το οποίο συμπίπτει με το βάθος υπερεπιπέδου του χ στο αρχικό πρόβλημα.

Το βάθος υπερεπιπέδου έχει προκαλέσει πρόσφατα το ενδιαφέρον της ερευνητικής κοινότητας. Από την πλευρά της συνδυαστικής, έχει αποδειχτεί [ABET00] ότι -όπως και το Tukey μεσοστοιχείο- το βάθος υπερεπιπέδου πρέπει να είναι τουλάχιστον $\left\lceil \frac{n}{d+1} \right\rceil$. Ακόμα οι Rousseeuw και Hubert έδειξαν πως το μέγιστο βάθος υπερεπιπέδου δεν μπορεί να είναι μεγαλύτερο από $\left\lfloor \frac{n+d}{2} \right\rfloor$, εάν φυσικά τα υπερεπίπεδα είναι σε γενική θέση (δηλαδή κανένα ζεύγος υπερεπιπέδων δεν είναι παράλληλα μεταξύ τους και δεν υπάρχουν d+1 υπερεπίπεδα που να μην έχουν κοινό σημείο) και ότι το μεσοστοιχείο υπερεπιπέδου είναι αμετάβλητο σε γραμμικούς μετασχηματισμούς.



Σχήμα 2.8. Το βάθος υπερεπιπέδου στον R^2 ενός σημείου p είναι το ελάχιστο πλήθος ευθειών που τέμνονται από μία ακτίνα με αρχή το p.

Στην υπολογιστική πλευρά του προβλήματος, το κύριο θέμα αποτελεί η εύρεση των φραγμάτων της πολυπλοκότητας χρόνου για τον υπολογισμό του μέγιστου βάθους υπερεπιπέδου. Τα υπερεπίπεδα του συνόλου S στον R^d διαμερίζουν τον χώρο αυτό σε $O(n^d)$ κυρτές περιοχές και η διαμέριση αυτή ονομάζεται διάταξη (arrangement) του S, που συμβολίζεται με A(S). Είναι προφανές πως κάθε σημείο

μέσα σε κάποια περιοχή έχει το ίδιο βάθος. Το βάθος της διάταξης δ(A(S)) ορίζεται να είναι το μέγιστο βάθος μεταξύ των περιοχών της A(S). Αυτό που μας απασχολεί λοιπόν είναι ο υπολογισμός του δ(A(S)) και ενός σημείου-μάρτυρα αυτού του βάθους ή η εύρεση μίας περιοχής δεδομένου βάθους k = δ(A(S)).

Οι Rousseeuw και Hubert [RH99] παρατήρησαν πως για κάθε p το δ(p) μπορεί να υπολογιστεί σε $O(n^{d-1} \log n)$ χρόνο και αφού υπάρχουν $O(n^d)$ κορυφές στην διάταξη A(S) το συνολικό κόστος σε χρόνο θα είναι $O(n^{2d-1} \log n)$. Για την διδιάστατη περίπτωση αναφέρουν έναν $O(n^3)$ αλγόριθμο. Οι Amenta et al. [ABET00] παρατήρησαν πως η διάταξη A(S) μπορεί να κατασκευαστεί σε $O(n^d)$ χρόνο και τότε με την χρήση BFS (Breadth First Search) στον γράφο των γειτονικών περιοχών το βάθος κάθε περιοχής μπορεί να βρεθεί στον ίδιο $O(n^d)$ χρόνο. Το παραπάνω είναι ανάλογο της χρήσης ταξινόμησης για την εύρεση του μονοδιάστατου μεσοστοιχείου. Για την περίπτωση των δύο διαστάσεων οι van Kreveld et al. [vKMR99] πρόσφατα περιέγραψαν έναν ευρηματικό $O(n \log^2 n)$ αλγόριθμο για το βάθος υπερεπιπέδου. Ο παραπάνω αλγόριθμος αποδεικνύει πως δεν είναι απαραίτητη η ταξινόμηση και δεν είναι η βέλτιστη λύση για την εύρεση του μονοδιάστατου μεσοστοιχείου. Ο αλγόριθμος στηρίζεται σε μία δυαδική αναζήτηση, η οποία βασίζεται στην δυνατότητα επιλογής της k-οστής κορυφής, ταξινομημένης κατά την x συντεταγμένη, σε ένα σύνολο από διαδοχικές, υποψήφιες προς επεξεργασία, κορυφές. Αυτό μπορεί να γίνει σε O(n logn) χρόνο. Η κατακόρυφη γραμμή που περνάει από την επιλεγμένη κορυφή τότε ελέγχεται ώστε να καθορίσει το ένα από τα δύο ημιεπίπεδα που ορίζονται από την ευθεία και το οποίο τέμνει μία περιοχή μέγιστου βάθους. Ο έλεγχος αυτός γίνεται σε O(n log n) χρόνο, οπότε αφού υπάρχουν $O\left(\log\binom{n}{2}\right)$ βήματα δυαδικής αναζήτησης, η συνολική πολυπλοκότητα του αλγόριθμου είναι $O(n \log^2 n)$.

Παρόλο που το βάθος υπερεπιπέδου είναι αντίστοιχο του μεσοστοιχείου στον R^1, δεν έχουν πραγματοποιηθεί γενικεύσεις για το πολυδιάστατο μεσοστοιχείο ενός συνόλου σημείων στον R^d. Η ιδέα της ακτίνας που διαπερνάει το ελάχιστο πλήθος από υπερεπίπεδα μπορεί να χρησιμοποιηθεί ως εξής: Δεδομένου ενός συνόλου n σημείων στον R^d, κατασκευάζουμε το σύνολο S των υπερεπιπέδων από υποσύνολα d σημείων του συνόλου των δεδομένων. Μετά βρίσκουμε το σημείο p με το μέγιστο βάθος υπερεπιπέδου ως προς το σύνολο S των υπερεπιπέδων και το **ονομάζουμε H-μεσοστοιχείο (H-median)** του S. Το διδιάστατο H-μεσοστοιχείο μπορεί να

υπολογιστεί σε $O(n^2 \log n)$ χρόνο με χρήση του αλγορίθμου των Langerman και Steiger, αφού υπάρχουν $O(n^2)$ υπερεπίπεδα στον R^2 για ένα σύνολο n σημείων.

ΚΕΦΑΛΑΙΟ 3

ΑΛΓΟΡΙΘΜΟΙ ΚΑΙ ΚΑΤΩ ΟΡΙΑ ΓΙΑ ΤΟ ΤΡΙΓΩΝΙΚΟ ΒΑΘΟΣ ΣΤΟΝ R^2.

3.1. Εισαγωγή.

Σε αυτό το κεφάλαιο θα αναφερθούμε αναλυτικότερα σε κάποια αποτελέσματα και αλγορίθμους σχετικά με την εύρεση της τιμής του τριγωνικού βάθους ενός σημείου στον R^2 σε συνάρτηση με ένα σύνολο σημείων S στο επίπεδο. Στην ενότητα 3.2. περιγράφουμε τον αλγόριθμο εύρεσης του τριγωνικού βάθους που πρότειναν οι Rousseeuw και Ruts [RR96] για τον υπολογισμό του τριγωνικού βάθους ενός σημείου στο επίπεδο σε O(n log n) χρόνο. Ο αλγόριθμός τους για το τριγωνικό βάθος είναι πρακτικά όμοιος με αυτόν των Khuller και Mitchell [KM89] και αυτόν των Gil, Steiger και Wigderson [GSW92]. Στην παράγραφο 3.3. αποδεικνύουμε ότι ο υπολογισμός του τριγωνικού βάθους απαιτεί Ω(n log n) χρόνο, ο οποίος ταυτίζεται με το άνω φράγμα αυτών των αλγορίθμων. Στο κεφάλαιο 3.4. σχολιάζουμε το κάτω φράγμα για τον έλεγχο προσήμου (sign test) των Oja και Nyblom [ON89] και την σχέση του με τον αλγόριθμο εύρεσης του simplicial βάθους.

3.2. Υπολογισμός Τριγωνικού Βάθους στον R^2.

Το τριγωνικό βάθος ενός σημείου p σε σχέση με ένα σύνολο σημείων S = $\{X_1, X_2, \ldots, X_n\}$ στο επίπεδο μας δίνεται από το πλήθος των τριγώνων $\Delta(X_i, X_j, X_k)$ που σχηματίζονται από τριάδες σημείων του συνόλου δεδομένων τα οποία περιέχουν το p. Αυτός ο υπολογισμός φαίνεται να απαιτεί (με τον προφανή τρόπο) $O(n^3)$ βήματα. Ο αλγόριθμος των Rousseeuw και Ruts [RR96] υπολογίζει το τριγωνικό βάθος σε O(n log n) χρόνο συνδυάζοντας γεωμετρικές ιδιότητες με κάποια ταξινόμηση και μηχανισμούς ανανέωσης πληροφορίας.

Αλγόριθμος 3.1. Έστω S = $\{X_1, X_2, \ldots, X_n\}$ ένα σύνολο από δοθέντα σημεία στο επίπεδο και p ένα σημείο στο επίπεδο που δεν ανήκει αναγκαστικά στο σύνολο S.

1. Ο αλγόριθμος πρώτα υπολογίζει για κάθε σημείο X_i του συνόλου των δεδομένων το διάνυσμα $u_i = (X_i - p) / \| X_i - p \|$.
2. Μετά υπολογίζουμε την κλίση α_i κάθε διανύσματος u_i και αποθηκεύουμε τις κλίσεις αυτές στον πίνακα A.
3. Στο επόμενο βήμα ταξινομούμε τον πίνακα A με κάποιον από τους συνήθεις O(n log n) αλγορίθμους.
4. Μετά υπολογίζουμε την μέγιστη διαφορά μεταξύ διαδοχικών α_i. Αν αυτή είναι μεγαλύτερη από π τότε το σημείο p βρίσκεται έξω από το κυρτό περίβλημα του συνόλου των δεδομένων και συνεπώς η τιμή του βάθους του είναι ίση με το μηδέν. Αν όχι αφαιρούμε το α_1 από κάθε α_i. Έτσι από κατασκευής μπορούμε να θεωρήσουμε από εδώ και πέρα ότι: $0 = \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n < 2\pi$.
5. Στο επόμενο βήμα υπολογίζουμε τον μεγαλύτερο δείκτη (που θα τον ονομάζουμε k) τέτοιον ώστε να ισχύει $\alpha_k < \pi$.
6. Τότε για κάθε i = 1, ..., n ορίζουμε το h_i να είναι ο μικρότερος ακέραιος αριθμός ώστε να ισχύει : $\alpha_i \leq \alpha_{i+1} \leq \ldots \leq \alpha_{i+hi} < \alpha_i + \pi$. Η μέθοδος υπολογισμού των h_i είναι βασισμένη στο γεγονός ότι $h_i = F(i) - i$, όπου $F(i) = \sum_j I[j; 0 \leq a_j < a_i + \pi]$ με j=1, ..., n. Σκοπός μας είναι τώρα ο

υπολογισμός του μονοδιάστατου πίνακα F = (F(1) , F(2) , ... , F(n)) σε O(n) χρόνο από τον οποίο μπορούμε σε σταθερό χρόνο να υπολογίσουμε το κάθε h_i. Αυτό γίνεται με την χρήση των αντίποδων $(\alpha_i + \pi)$ των γωνιών α_i. Έτσι το αρχικό μας σύνολο γωνιών έχει τώρα μέγεθος 2n (n αρχικές κλίσεις α_i και n αντίποδες κλίσεις β_i).

7. Για κάθε κλίση α_i υπολογίζουμε την αντίποδη κλίση β_i ως εξής :

   $\beta_i = \alpha_i + \pi$, αν $0 \leq \alpha_i < \pi$

   και $\beta_i = \alpha_i - \pi$, αν $\pi \leq \alpha_i < 2\pi$

8. Αποθηκεύουμε τα α_i και τα β_i σε ένα μονοδιάστατο πίνακα $\Gamma = (\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n)$ μήκους 2n. Ακόμα δημιουργούμε και ένα πίνακα W = (1, ..., 1, -1, ..., -1) ίδιου μήκους, τα στοιχεία του οποίου είναι 1 αν στην αντίστοιχη θέση του πίνακα Γ έχω αρχική κλίση α_i ή −1 αν στην αντίστοιχη θέση του πίνακα Γ έχω αντίποδη κλίση β_i.
9. Μετά ταξινομούμε τον πίνακα Γ με κάποιον συνήθη αλγόριθμο ταξινόμησης σε O(2n log 2n) = O(n log n) χρόνο, ενώ παράλληλα εφαρμόζουμε τις ίδιες μεταθέσεις στοιχείων και στον πίνακα W.
10. Γνωρίζουμε αρχικά ότι F(1) = k (που έχουμε υπολογίσει σε προηγούμενο βήμα). Για να υπολογίσουμε τα υπόλοιπα στοιχεία του πίνακα F, βρίσκουμε τον μικρότερο δείκτη s ώστε το στοιχείο γ_s να είναι μεγαλύτερο του π. Τότε διασχίζουμε την ακολουθία $\gamma_s, \gamma_{s+1}, \ldots, \gamma_{2n}, \gamma_1, \ldots, \gamma_{s-1}$. Σε κάθε κλίση α_i (που το καταλαβαίνουμε από την τιμή της αντίστοιχης θέσης του πίνακα W) αυξάνουμε τον δείκτη NF (ο οποίος αρχικά έχει τιμή k) κατά ένα αλλά δεν αλλάζουμε τιμή στο F(i). Αντίθετα, σε κάθε αντίποδη κλίση β_i αυξάνουμε το i κατά ένα και τότε αναθέτουμε στο F(i) = NF.
11. Έχοντας υπολογίσει τα h_i από τα F(i) είμαστε σε θέση να υπολογίσουμε το πλήθος των τριγώνων που σχηματίζονται από σημεία του συνόλου S τα οποία περιέχουν το σημείο p παίρνοντας το συμπλήρωμα του πλήθους των τριγώνων που δεν περιέχουν το p σύμφωνα με την παρακάτω έκφραση:

$$depth(p;S) = \binom{n}{3} - \sum_{i=1}^{n} \binom{h_i}{2}$$

θεωρώντας την σύμβαση ότι $\binom{p}{q} = 0$ όταν p < q.

Σαν παρατήρηση μπορούμε να πούμε πως η παραπάνω έκφραση δίνει σωστό αποτέλεσμα ακόμα και όταν μερικά α_i συμπίπτουν σε τιμή, αφού έχουν διαφορετικούς δείκτες στην ταξινομημένη ακολουθία. Για παράδειγμα, αν $\alpha_3 = \alpha_4$ τότε θα ισχύει $h_4 = h_3 - 1$, όπως θα έπρεπε ώστε να αποφεύγουμε την περίπτωση κάποια τρίγωνα να μετρηθούν δύο φορές στην παραπάνω έκφραση. Ολόκληρος ο αλγόριθμος όπως λένε και οι συγγραφείς του [RR96] είναι σχεδιασμένος έτσι ώστε να λειτουργεί σωστά ακόμα και αν τα δεδομένα δεν είναι σε γενική θέση, π.χ. όταν τρία σημεία είναι συνευθειακά. Ακόμα, μπορούμε να παρατηρήσουμε πως τα α_i δεν παραμένουν αμετάβλητα σε γραμμικούς μετασχηματισμούς, αλλά τα h_i όπως και η παραπάνω έκφραση είναι.

Σε αυτό το σημείο θα δούμε κάποιες επιπλέον παρατηρήσεις που μας δίνουν οι Rousseeuw και Ruts στο [RR96] για τα βήματα του αλγορίθμου. Στο πρώτο βήμα του αλγορίθμου μετράμε τα σημεία τα οποία συμπίπτουν με το p και τα αφαιρούμε από το σύνολο των δεδομένων μας, αφού θα τα λάβουμε υπόψη μας στο τέλος του αλγορίθμου. Στο βήμα 2 ο υπολογισμός των α_i γίνεται χρησιμοποιώντας την συνάρτηση ASIN, όπου είναι πιο ακριβής και την συνάρτηση ACOS στις υπόλοιπες περιπτώσεις. Ακόμη είναι ενδιαφέρον το γεγονός ότι από το σημείο αυτό και πέρα δεν χρειαζόμαστε πλέον τις συντεταγμένες των δοθέντων σημείων και οι υπόλοιποι υπολογισμοί βασίζονται μόνο στις κλίσεις που είναι αποθηκευμένες στον πίνακα A. Στο βήμα 4 η αφαίρεση του α_1 από κάθε α_i ισοδυναμεί με περιστροφή των αρχικών σημείων του συνόλου των δεδομένων γύρω από το p, πράξη που μας επιτρέπει η αμεταβλητότητα σε γραμμικούς μετασχηματισμούς του τριγωνικού βάθους. Στο επόμενο βήμα του αλγορίθμου παρατηρούμε πως ο δείκτης k συμπίπτει με το πλήθος των α_i που βρίσκονται στο άνω ημικύκλιο, συμπεριλαμβανομένου και του σημείου 0, αλλά χωρίς το σημείο π. Στο έκτο βήμα λαμβάνεται υπόψη ότι το $\alpha_i + \pi$ μπορεί να υπερβαίνει τον πρώτο κύκλο και να βρίσκεται στο $[2\pi , 3\pi)$. Στο ίδιο βήμα, εάν υπολογίσουμε το κάθε h_i με τον τετριμμένο τρόπο θα χρειαζόμασταν $O(n)$ λειτουργίες για κάθε h_i και συνολικά $O(n^2)$ λειτουργίες. Όμως αυτό το αποφεύγουμε με την μέθοδο που προτείνουν οι συγγραφείς του αλγορίθμου από το βήμα 7 και μετά με την τεχνική χρήσης των αντίποδων των γωνιών.

Λαμβάνοντας υπόψη μας τα βήματα του παραπάνω αλγόριθμου παρατηρούμε πως ο υπολογισμός όλων των u_i από τα X_i στο βήμα 1 γίνεται σε $O(n)$ χρόνο. Ακόμη ο υπολογισμός των κλίσεων α_i από τα u_i στο βήμα 2 γίνεται και αυτός σε $O(n)$ χρόνο. Το βήμα 3 της ταξινόμησης του μονοδιάστατου πίνακα A απαιτεί $O(n \log n)$ χρόνο.

Το βήμα 4 που περιλαμβάνει τον υπολογισμό n-1 διαφορών και αφαίρεση του στοιχείου α_1 από κάθε στοιχείο του πίνακα A (που έχει μέγεθος n) μας κοστίζει συνολικά O(n) χρόνο. Το βήμα 5 κάνοντας απλή διάσχιση του πίνακα A στην χειρότερη περίπτωση μπορεί να μας κοστίσει O(n) χρόνο, ενώ με δυαδική αναζήτηση O(log n) χρόνο. Στο βήμα 7 ο υπολογισμός των n αντίποδων κλίσεων β_i από τα α_i γίνεται με πολυπλοκότητα O(n). Η δημιουργία των δύο πινάκων Γ και W, μεγέθους 2n και οι δύο, στο βήμα 8 του αλγορίθμου απαιτεί O(n) λειτουργίες. Η ταξινόμηση του πίνακα Γ και οι ταυτόχρονες μεταθέσεις στον πίνακα W απαιτούν, όπως αναφέραμε στο βήμα 9, O(n log n) χρόνο. Οι λειτουργίες που εκτελούνται στο βήμα 10 και απαιτούν την διάσχιση του πίνακα Γ, τον έλεγχο του πίνακα W και την ενημέρωση του πίνακα F απαιτούν χρόνο ανάλογο του μεγέθους των πινάκων αυτών, δηλαδή O(n) χρόνο. Ο υπολογισμός των h_i από τα F(i) στο βήμα 11 καθώς και ο υπολογισμός του αθροίσματος απαιτεί γραμμικό χρόνο στο σύνολο των h_i, δηλαδή O(n) χρόνο. Στα βήματα του παραπάνω αλγόριθμου βλέπουμε πως κυριαρχούν οι πολυπλοκότητες των βημάτων 3 και 9 που χρησιμοποιούν ταξινόμηση O(n) στοιχείων και απαιτούν O(n log n) χρόνο. Έτσι η συνολική πολυπλοκότητα του αλγορίθμου εύρεσης του τριγωνικού βάθους ενός σημείου p σε σχέση με ένα σύνολο σημείων S μεγέθους n είναι της τάξης του O(n log n).

3.3. Κάτω φράγμα για τον υπολογισμό του τριγωνικού βάθους ενός σημείου στον R^2.

Σε αυτήν την ενότητα θα δείξουμε πως η εύρεση του τριγωνικού βάθους ενός σημείου p στον R^2 έχει κάτω φράγμα Ω(n log n) χρόνου, το οποίο μας δίνεται από την αναγωγή του προβλήματος μας στο πρόβλημα της μοναδικότητας σημείου (element distinctness), δηλαδή του προβλήματος εξακρίβωσης εάν κάθε ένας από n δοθέντες αριθμούς είναι διαφορετικός από όλους τους υπολοίπους, το οποίο έχει ένα Ω(n log n) κάτω φράγμα πράξεων στο μοντέλο αλγεβρικού δέντρου απόφασης. Είναι λοιπόν φυσικό όλοι οι αλγόριθμοι εύρεσης τριγωνικού βάθους να συγκρίνονται με αυτό το κάτω φράγμα ως προς την αποτελεσματικότητα της πολυπλοκότητας τους.

Ορισμός 3.1. *Πρόβλημα μοναδικότητας στοιχείου: Δοθέντος ενός συνόλου* $E = \{e_1, e_2, \ldots, e_n\}$, *υπάρχει ένα ζευγάρι στοιχείων* e_i, e_j *(με* i *διάφορο του* j*) τέτοιο ώστε* $e_1 = e_2$;

Στο σημείο αυτό θα αποδείξουμε τον ισχυρισμό μας ανάγοντας το πρόβλημα της μοναδικότητας σημείου σε αυτό της εύρεσης του τριγωνικού βάθους.

Απόδειξη. Έστω $E = \{e_1, e_2, \ldots, e_n\}$ ένα σύνολο από πραγματικούς αριθμούς με μέγεθος μεγαλύτερο ή ίσο του τρία. Για κάθε e_i (με $1 \leq i \leq n$) δημιουργούμε το σύνολο σημείων S στο επίπεδο μεγέθους 2n με τις τιμές των στοιχείων του να δίνονται ως εξής:

$s_i = (e_i, 1)$

και $s_{n+i} = (-e_i, -1)$

Από την παραπάνω κατασκευή είναι σαφές ότι τα s_i και s_{n+i} είναι συμμετρικά μεταξύ τους ως προς την αρχή των αξόνων (0,0), την οποία και διαλέγουμε ως το σημείο p του οποίου θέλουμε να υπολογίσουμε το τριγωνικό βάθος.

Έστω τώρα ότι το s_i είναι μοναδικό στοιχείο στο σύνολο σημείων στο επίπεδο S. Τότε η ποσότητα h_i (η οποία ορίστηκε παραπάνω κατά την περιγραφή του αλγορίθμου 3.1) θα πρέπει να είναι ίση με n-1, αφού το h_i θα περιλαμβάνει τα σημεία $s_{i+1}, s_{i+2}, \ldots, s_{n+i-1}$ (αυτό συμβαίνει επειδή στο ημικύκλιο $[\alpha_i, \alpha_i + \pi)$ θα υπάρχει για κάθε στοιχείο e_j διάφορο του e_i είτε η κλίση του σημείου s_j είτε η κλίση του σημείου s_{n+j}). Τότε εάν κανένα στοιχείο του S δεν είναι διπλό με κάποιο άλλο στοιχείο του συνόλου αυτού τότε το τριγωνικό βάθος του σημείου p = (0,0) ως προς το σύνολο S θα είναι:

$$D_c = SD(p;S) = \binom{2n}{3} - \sum_{i=1}^{2n} \binom{n-1}{2}$$

Είναι εύκολο να δούμε πως η παραπάνω τιμή αθροίσματος D_c για δεδομένο n είναι σταθερή και ανεξάρτητή των σημείων του συνόλου των δεδομένων αν αυτά είναι μοναδικά μεταξύ τους.

Αν υποθέσουμε τώρα ότι $s_i = s_{i+1}$ για κάποιο i. Τότε το h_{i+1} θα έχει τιμή μικρότερη από n – 1, γιατί το h_{i+1} θα περιλαμβάνει το πολύ τα σημεία $s_{i+2}, s_{i+3}, \ldots, s_{n+i-1}$ αλλά όχι και το συμμετρικό σημείο s_{n+i} του σημείου s_i το οποίο συμπίπτει με το δικό του συμμετρικό σημείο s_{n+i+1}. Με την παραπάνω παρατήρηση είναι πλέον εύκολη η διαπίστωση πως αν τουλάχιστον ένα στοιχείο του συνόλου E είναι ίδιο με

κάποιο άλλο στοιχείο του ίδιου συνόλου τότε το ίδιο θα συμβαίνει για τουλάχιστον δυο σημεία του συνόλου S (αφού από ένα στοιχείο του συνόλου E δημιουργούνται δύο σημεία του συνόλου S) και στην περίπτωση αυτή το τριγωνικό βάθος του σημείου p = (0,0) ως προς το σύνολο σημείων S θα είναι αυστηρά μεγαλύτερο από το τριγωνικό βάθος του ίδιου σημείου στην περίπτωση που κάθε σημείο του συνόλου S (επομένως και του συνόλου E) είναι μοναδικό. Έτσι με την εύρεση του τριγωνικού βάθους του σημείου μπορούμε να απαντήσουμε στο πρόβλημα μοναδικότητας σημείου: τα στοιχεία του συνόλου E είναι μοναδικά αν και μόνο αν το τριγωνικό βάθος του σημείου p = (0,0) ως προς το σύνολο σημείων S είναι ίσο με D_c.

Οι υπολογισμοί και οι μετατροπές που απαιτούνται για την αναγωγή του προβλήματος της μοναδικότητας στοιχείου είναι η κατασκευή του συνόλου σημείων S και ο υπολογισμός του αθροίσματος D_c, οι οποίοι μπορούν να γίνουν σε γραμμικό χρόνο σε σχέση με το μέγεθος του προβλήματος (δηλαδή το μέγεθος του συνόλου των στοιχείων E). ▣

3.4. Φράγμα για τον έλεγχο προσήμου (sign test) των Oja και Nyblom και η σχέση του με την πολυπλοκότητα εύρεσης του τριγωνικού βάθους.

Στην ενότητα αυτή θα εξετάσουμε την σχέση που έχει ο διδιάστατος έλεγχος προσήμου των Oja και Nyblom [ON89] με το τριγωνικό βάθος. Ο έλεγχος προσήμου χρησιμεύει στην περίπτωση που θέλουμε να διαπιστώσουμε αν υπάρχει, από στατιστική σκοπιά, σημαντική διαφορά μεταξύ δύο κατανομών από n ζευγάρια σημείων δεδομένων. Στην περίπτωση των διδιάστατων ελέγχων προσήμου ελέγχουμε όπως είναι φυσικό κατανομές σημείων στο επίπεδο.

Ορισμός 3.2. *Έστω $s_1, s_2, \ldots, s_n$ είναι ένα τυχαίο δείγμα μίας διδιάστατης κατανομής. Τότε ο διδιάστατος έλεγχος προσήμου μπορεί να βασιστεί στην παρακάτω στατιστική ποσότητα:*

$$\sum\sum\sum_{1 \le i < j < k \le n} I_{ijk}$$

όπου $I_{ijk} = 1$ *αν υπάρχει μία ευθεία που περνάει από την αρχή των αξόνων τέτοια ώστε τα* s_i, s_j *και* s_k *να βρίσκονται από την ίδια πλευρά της ευθείας αυτής. Αλλιώς* $I_{ijk} = 0$.

Από τον παραπάνω ορισμό μπορούμε να παρατηρήσουμε πως ο έλεγχος προσήμου περιλαμβάνει τον υπολογισμό τριάδων σημείων που έχουν την ιδιότητα να βρίσκονται στην ίδια πλευρά μίας ευθείας που διέρχεται από ένα συγκεκριμένο σημείο, το οποίο στην περίπτωση του ελέγχου προσήμου είναι η αρχή των αξόνων. Οι Rousseeuw και Ruts [RR96] αναφέρουν πως η μέθοδός τους εύρεσης του τριγωνικού βάθους (την οποία περιγράψαμε σε προηγούμενη ενότητα αυτού του κεφαλαίου) μπορεί να χρησιμοποιηθεί για τον υπολογισμό του προσημασμένου ελέγχου των Oja και Nyblom. Έτσι είναι προφανές ότι ο έλεγχος προσήμου των Oja και Nyblom απαιτεί O(n log n) χρόνο στην χειρότερη περίπτωση.

Αντίθετα, μπορούμε να ανάγουμε το πρόβλημα εύρεσης του τριγωνικού βάθους ενός σημείου p ως προς ένα σύνολο n σημείων δεδομένων στο πρόβλημα του ελέγχου προσήμου. Αυτό μπορεί να γίνει απλά αν για τον υπολογισμό του τριγωνικού βάθους του σημείου p εφαρμόσουμε τον έλεγχο προσήμου των Oja και Nyblom πάνω στο σύνολο των σημείων δεδομένων θεωρώντας ως αρχή των αξόνων το σημείο p. Ο μετασχηματισμός που απαιτείται για την αναγωγή του προβλήματος προφανώς είναι γραμμικός στο σύνολο των n δοθέντων σημείων.

ΚΕΦΑΛΑΙΟ 4

ΑΛΓΟΡΙΘΜΟΙ ΓΙΑ ΤΟΝ ΤΡΙΓΩΝΙΚΟ ΕΚΤΙΜΗΤΗ ΣΤΟΝ R^2.

Σε αυτό το κεφάλαιο θα αναφερθούμε στον αλγόριθμο για την εύρεση του simplicial μεσοστοιχείου που πρότειναν οι Aloupis, Langerman, Soss και Toussaint [ALST01]. Έστω ένα σύνολο δεδομένων S που περιέχει n σημείων στον R^2 και I ένα σύνολο από ευθύγραμμα τμήματα τα οποία σχηματίζονται μεταξύ ζευγαριών σημείων του S. Ο αλγόριθμος **Simp-Med** [ALST01] που θα παρουσιάσουμε υπολογίζει το simplicial μεσοστοιχείο του συνόλου S των n σημείων και απαιτεί $O(n^4 \log n)$ χρόνο και $O(n^2)$ χώρο ή εναλλακτικά $O(n^4)$ χώρο και χρόνο. Στην πραγματικότητα ο αλγόριθμος κατά κύριο λόγο επεξεργάζεται κάθε ευθύγραμμο τμήμα του συνόλου I, σαρώνοντας τα ταξινομημένα σημεία τομής που δημιουργούνται από αυτό και τα άλλα ευθύγραμμα τμήματα του ίδιου συνόλου. Εάν εφαρμόσουμε τοπολογική σάρωση (topological sweep), αντί να ασχολούμαστε με την επεξεργασία του κάθε ευθύγραμμου τμήματος ξεχωριστά, μπορούμε να μειώσουμε την πολυπλοκότητα του υπολογισμού σε $O(n^4)$ χρόνο και $O(n^2)$ χώρο.

Πρόταση 4.1. *Για να βρούμε ένα σημείο με μέγιστο simplicial βάθος αρκεί να επεξεργαστούμε τα σημεία τομής των ευθυγράμμων τμημάτων του I .*

Απόδειξη. Αν υπολογίσουμε και μελετήσουμε την διάταξη (arrangement) που ορίζουν τα ευθύγραμμα τμήματα του συνόλου I στο επίπεδο, είναι απλό να παρατηρήσουμε πώς όλα τα σημεία που βρίσκονται μέσα στο εσωτερικό μίας δοθείσης περιοχής πρέπει να έχουνε την ίδια τιμή τριγωνικού βάθους. Κάθε σημείο που βρίσκεται πάνω στο σύνορο της περιοχής (δηλαδή πάνω σε κάποιο από τα ευθύγραμμα τμήματα του συνόλου I) πρέπει να έχει τιμή τριγωνικού βάθους τουλάχιστον ίση ή και μεγαλύτερη από αυτήν που έχουν τα σημεία στο εσωτερικό της περιοχής. Τελικά, κάθε κορυφή του συνόρου της περιοχής (δηλαδή κάθε σημείο τομής δύο ευθυγράμμων τμημάτων του συνόλου I) πρέπει να έχει τιμή τριγωνικού βάθους τουλάχιστον ίση ή και μεγαλύτερη από αυτήν που έχει κάθε σημείο που βρίσκεται πάνω στα γειτνιάζοντα σύνορα των περιοχών. Από τους παραπάνω περιορισμούς μπορούμε να δούμε πως οι κορυφές των συνόρων των περιοχών έχουν τιμή τριγωνικού βάθους τουλάχιστον ίση ή και μεγαλύτερη από κάθε άλλο γειτονικό τους σημείο. Επομένως τουλάχιστον ένα σημείο με την μέγιστη τιμή τριγωνικού βάθους θα βρίσκεται πάνω σε κάποια τομή δύο ευθυγράμμων τμημάτων του συνόλου I. ▣

Για να απλοποιήσουμε την περιγραφή του αλγορίθμου Simp-Med για τον υπολογισμό του τριγωνικού εκτιμητή θεωρούμε πως το σύνολο σημείων S βρίσκεται σε γενική θέση, με την έννοια ότι καμία τριάδα σημείων του S δεν είναι συνευθειακά. Αργότερα θα δούμε πως ο παραπάνω περιορισμός μπορεί να αρθεί αν χρειαστεί, χωρίς να επηρεάζονται οι πολυπλοκότητες χώρου και χρόνου.

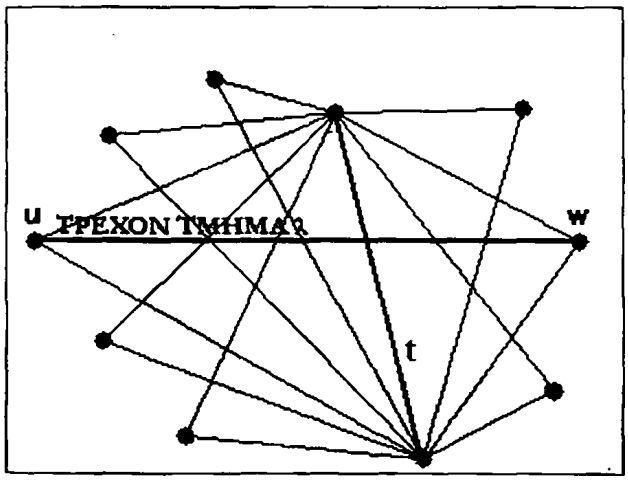
Αλγόριθμος Simp-Med

1. Υπολόγισε το τριγωνικό βάθος κάθε σημείου του συνόλου δεδομένων του S.
2. Υπολόγισε το πλήθος των σημείων τα οποία βρίσκονται αυστηρά από την μία πλευρά κάθε ευθύγραμμου τμήματος του I (και για τις δύο πλευρές κάθε ευθύγραμμου τμήματος ξεχωριστά).
3. Για κάθε ευθύγραμμο τμήμα λ που ανήκει στο I :
   a) Υπολόγισε και ταξινόμησε τα σημεία τομής όλων των άλλων ευθυγράμμων τμημάτων με το λ, αν αυτά φυσικά υπάρχουν.
   b) Θέσε ως τιμή της μεταβλητής MAX την τιμή του τριγωνικού βάθους του σημείου-άκρου (από τα δύο που ορίζουν τα άκρα του ευθύγραμμου τμήματος λ) με την μεγαλύτερη τιμή βάθους που υπολογίστηκε στο βήμα 1.

c) Αν υπάρχουν σημεία τομής πάνω στο ευθύγραμμο τμήμα λ, τότε υπολόγισε το τριγωνικό βάθος d του σημείου τομής που βρίσκεται πλησιέστερα στο παραπάνω άκρο του λ. Ενημέρωσε την τιμή του MAX (αν η τιμή του MAX είναι μικρότερη από αυτή του d, θέσε την τιμή του MAX ίση με d).

d) Συνέχισε την διάσχιση της ταξινομημένης λίστας. Κάθε φορά που αφήνουμε πίσω ένα ευθύγραμμο τμήμα που τέμνει το λ αφαιρούμε από την τιμή του d το πλήθος των σημείων που βρίσκονται αυστηρά πίσω (κατά την φορά διάσχισης του λ) από το ευθύγραμμο τμήμα αυτό (και έχουμε υπολογίσει στο βήμα 2). Κάθε φορά που συναντούμε ένα ευθύγραμμο τμήμα που τέμνει το λ προσθέτουμε στην τιμή του d το πλήθος των σημείων που βρίσκονται αυστηρά μπροστά (κατά την φορά διάσχισης του λ) από αυτό το ευθύγραμμο τμήμα. Ενημέρωσε την τιμή του MAX μετά από κάθε επεξεργασία σημείου τομής.

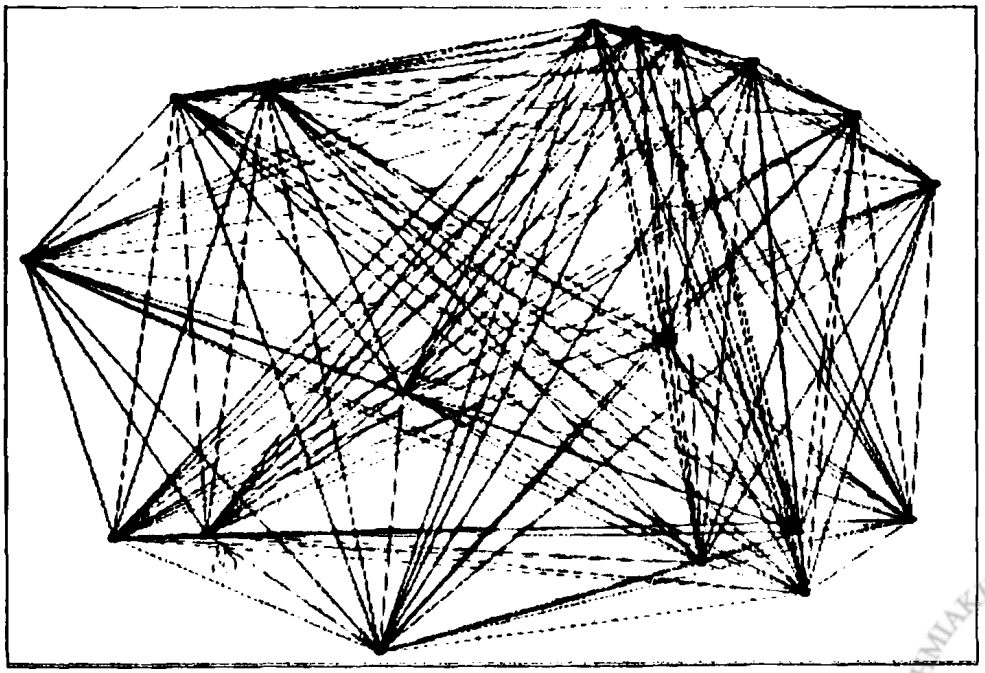
4. Επίστρεψε την μέγιστη τιμή MAX που βρέθηκε στις διασχίσεις όλων των ευθυγράμμων τμημάτων και το συσχετιζόμενο με αυτή την τιμή σημείο τομής.



Σχήμα 4.1. Στιγμιότυπο του Simp-Med κατά την διάσχιση του τρέχοντος τμήματος λ και του τέμνοντος ευθύγραμμου τμήματος t.

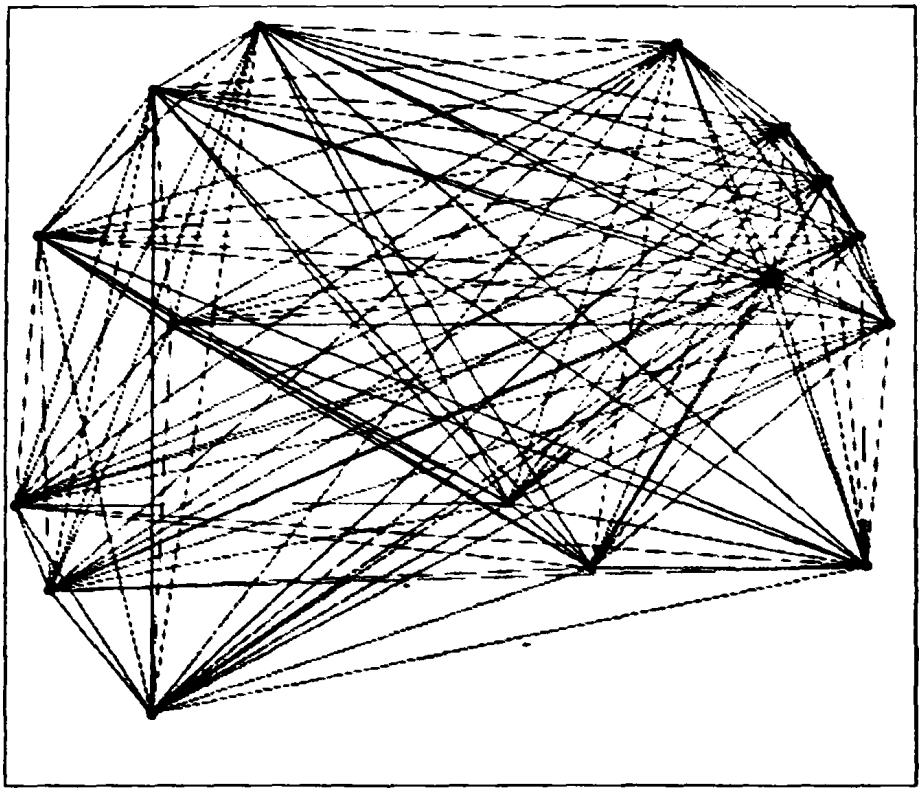
Το σχήμα 4.1. επιδεικνύει την λογική γύρω από το βήμα 3 του αλγορίθμου Simp-Med. Από την στιγμή που ξέρουμε την τιμή τριγωνικού βάθους ενός από τα σημεία τομής (τιμή που υπολογίζουμε στο βήμα 3c) του ευθύγραμμου τμήματος λ, είμαστε σε θέση να υπολογίσουμε όλες τις υπόλοιπες τιμές τριγωνικού βάθους κατά μήκος του δοθέντος τμήματος λ, την καθεμιά σε σταθερό χρόνο. Κάθε τέμνον

ευθύγραμμο τμήμα σχηματίζει ένα τρίγωνο με κάθε σημείο που βρίσκεται αυστηρά σε κάποια από τις πλευρές του. Έτσι κάθε φορά που κατά την διάσχιση του ευθύγραμμου τμήματος λ από το ένα άκρο του u στο άλλο του άκρο w συναντάμε ένα τέμνον ευθύγραμμο τμήμα t, εισερχόμαστε σε τόσα τρίγωνα όσα είναι και τα σημεία που βρίσκονται αυστηρά στην ίδια μεριά με το w ως προς το τμήμα t. Αντίστοιχα κάθε φορά που κατά την διάσχιση αφήνουμε πίσω μας ένα τέμνον ευθύγραμμο τμήμα t, εξερχόμαστε και από τόσα τρίγωνα όσα είναι τα σημεία που βρίσκονται από την ίδια πλευρά με το u ως προς το ευθύγραμμο τμήμα t. Έτσι στο σχήμα 4.1, αν υποθέσουμε ότι μετακινούμαστε από αριστερά προς τα δεξιά πάνω στο τρέχον ευθύγραμμο τμήμα λ, τότε ακριβώς πριν συναντήσουμε το ευθύγραμμο τμήμα t η τιμή που έχει η μεταβλητή d είναι 10. Όταν συναντήσουμε το ευθύγραμμο τμήμα t (δηλαδή όταν φτάσουμε στο σημείο τομής του λ με το t), είναι πολύ εύκολο να δούμε ότι βρισκόμαστε μέσα σε 3 επιπλέον τρίγωνα, ένα για κάθε σημείο στα δεξιά του t. Αμέσως μόλις κινηθούμε λίγο δεξιότερα και αφήσουμε πίσω μας το t, παρατηρούμε αμέσως ότι δεν βρισκόμαστε πλέον μέσα στα 5 τρίγωνα που σχηματίζει το t με τα 5 σημεία στα αριστερά του. Συνεπώς όσο βρισκόμαστε πάνω στο ευθύγραμμο τμήμα t έχουμε τιμή τριγωνικού βάθους d = 10 + 3 = 13 και αμέσως μετά d = 13 –5 = 8. Δύο παραδείγματα της υλοποίησης του αλγορίθμου Simp-Med φαίνονται στα σχήματα 4.2 (για 18 σημεία στο επίπεδο) και 4.3 (για 16 σημεία στο επίπεδο). Οι μικρές τελείες αναπαριστούν τα δοθέντα σημεία του συνόλου, ενώ οι μεγάλες τελείες τα σημεία με μέγιστο τριγωνικό βάθος.



Σχήμα 4.2. Αποτέλεσμα υλοποίησης αλγορίθμου Simp-Med για 18 τυχαία σημεία.

Στην περίπτωση που έχουμε πολλαπλά ευθύγραμμα τμήματα που τέμνουν το τρέχον ευθύγραμμο τμήμα στο ίδιο ακριβώς σημείο, μπορούμε να φτιάξουμε μία προσωρινή λίστα με αυτά τα ευθύγραμμα τμήματα. Όταν συναντήσουμε και το τελευταίο από τα τμήματα αυτά τότε επεξεργαζόμαστε την λίστα αυτή. Πρώτα προσθέτουμε το άθροισμα των σημείων που βρίσκονται στα δεξιά κάθε ευθύγραμμου τμήματος που τέμνει το ευθύγραμμο τμήμα στο σημείο αυτό και μετά αφαιρούμε το άθροισμα των σημείων που βρίσκονται στα αριστερά των τμημάτων αυτών.



Σχήμα 4.3. Αποτέλεσμα υλοποίησης αλγορίθμου Simp-Med για 16 τυχαία σημεία.

Πρόταση 4.2. *Ο αλγόριθμος Simp-Med υπολογίζει το διδιάστατο simplicial μεσοστοιχείο ενός συνόλου n σημείων σε $O(n^4 \log n)$ χρόνο και $O(n^2)$ χώρο ή εναλλακτικά σε $O(n^4)$ χρόνο και χώρο .*

Απόδειξη. Αφού όλα τα σημεία τομής των ευθύγραμμων τμημάτων που σχηματίζονται από ζευγάρια σημείων δεδομένων ελέγχονται από τον αλγόριθμο, τότε σύμφωνα με την πρόταση 4.1. θα βρούμε σίγουρα ένα σημείο με μέγιστη τιμή simplicial βάθους.

Το βήμα 1 του αλγορίθμου απαιτεί $O(n^2 \log n)$ χρόνο συνολικά εφόσον ο υπολογισμός του τριγωνικού βάθους ενός σημείου σύμφωνα με τον αλγόριθμου των

Rousseeuw και Ruts [RR96] που παρουσιάσαμε στο προηγούμενο κεφάλαιο απαιτεί $O(n \log n)$ χρόνο για ένα σημείο.

Το βήμα 2 είναι απλό να υπολογιστεί με τον προφανή τρόπο σε $O(n^3)$ χρόνο, απλώς ελέγχοντας για καθένα από τα $O(n^2)$ ευθύγραμμα τμήματα σε ποια μεριά του τμήματος αυτού βρίσκεται καθένα από τα n σημεία του συνόλου των δεδομένων.

Το βήμα 3a απαιτεί χρόνο $O(n^2 \log n)$, γιατί η εύρεση των σημείων τομής (το πλήθος των οποίων στην χειρότερη περίπτωση είναι της τάξης $\Theta(n^2)$) και η ταξινόμησή τους απαιτούν χρόνο $O(n^2)$ και $O(n^2 \log n)$, στην χειρότερη περίπτωση, αντίστοιχα. Το βήμα 3b είναι προφανές ότι απαιτεί σταθερό χρόνο. Το βήμα 3c απαιτεί $O(n \log n)$ χρόνο, αφού έχουμε τον υπολογισμό του τριγωνικού βάθους ενός σημείου. Το βήμα 3d απαιτεί χρόνο $O(n^2)$, αφού η ενημέρωση των τιμών d και MAX για κάθε σημείο τομής γίνεται σε σταθερό χρόνο και έχουμε το πολύ $O(n^2)$ σημεία τομής σε κάθε ευθύγραμμο τμήμα. Επομένως συνολικά το βήμα 3 απαιτεί $O(n^4 \log n)$ χρόνο, αφού εκτελείται για καθένα από τα $O(n^2)$ ευθύγραμμα τμήματα και η πολυπλοκότητα που κυριαρχεί σε κάθε εκτέλεση είναι αυτή του βήματος 3a.

Ο χώρος που χρησιμοποιείται είναι $O(n^2)$ και χρησιμοποιείται στα βήματα 2 (για την αποθήκευση του πλήθους των σημείων που βρίσκονται στα αριστερά και στα δεξιά των $O(n^2)$ ευθυγράμμων τμημάτων) και 3a (για την αποθήκευση των $O(n^2)$ σημείων τομής). Αν υπολογίσουμε την διάταξη A(I) του συνόλου I των ευθυγράμμων τμημάτων σε $O(n^4)$ χρόνο και χώρο ως ενδιάμεσο στάδιο προεπεξεργασίας, τότε το βήμα 3a δεν είναι πλέον απαραίτητο. Στην περίπτωση αυτή, το βήμα της προεπεξεργασίας κυριαρχεί στην πολυπλοκότητα χώρου και χρόνου του αλγορίθμου.

Αντί να εκτελέσουμε το βήμα 3 του αλγορίθμου Simp-Med, μπορούμε να εφαρμόσουμε μία τοπολογική σάρωση η οποία απαιτεί $O(n^2)$ χρόνο και $O(n)$ χώρο για n ευθείες (όπως αναφέρεται στο [EG89]). Επομένως αντί να επεξεργαζόμαστε κάθε ευθύγραμμο τμήμα σειριακά, επεξεργαζόμαστε όλα τα $O(n^2)$ τμήματα ταυτόχρονα. Κάθε φορά που η καμπύλη σάρωσης συναντάει ένα σημείο τομής πάνω σε κάποιο ευθύγραμμο τμήμα, επεξεργαζόμαστε το σημείο με τον ίδιο τρόπο όπως και το βήμα 3d του αλγορίθμου Simp-Med. Από την εφαρμογή των παραπάνω τροποποιήσεων του αλγορίθμου Simp-Med προκύπτει η απόδειξη της ακόλουθης πρότασης:

Πρόταση 4.3. *Το simplicial μεσοστοιχείο ενός συνόλου n σημείων είναι δυνατόν να υπολογιστεί σε $O(n^4)$ χρόνο και $O(n^2)$ χώρο.*

Λαμβάνοντας τώρα υπόψη μας το πλεονέκτημα που μας δίνει η δομή των ευθύγραμμων τμημάτων σε αυτό το πρόβλημα, μπορούμε να μειώσουμε την πολυπλοκότητα χρόνου σε κάποια βήματα του παραπάνω αλγορίθμου, παρόλο που η συνολική πολυπλοκότητα βελτιώνεται μόνο κατά ένα σταθερό παράγοντα. Για παράδειγμα στο βήμα 2 του αλγορίθμου μπορούμε να υπολογίσουμε το βάθος ημιχώρου για κάθε σημείο του συνόλου των δεδομένων, αλλά να αποθηκεύουμε τα σημεία που καταμετρήθηκαν για κάθε ημιεπίπεδο το οποίο επεξεργαστήκαμε. Σε αυτό το σημείο κρίνεται σκόπιμο να υπενθυμίσουμε πως το βάθος ημιχώρου ενός σημείου p ως προς ένα δοθέν σύνολο n σημείων είναι το ελάχιστο πλήθος σημείων του συνόλου αυτού τα οποία περιέχονται σε οποιοδήποτε ανοιχτό ημιεπίπεδο που καθορίζεται από μία ευθεία που διέρχεται από το σημείο p. Το βάθος ημιχώρου μπορεί να υπολογιστεί σε O(n log n) χρόνο με έναν απλό αλγόριθμο που περιέγραψαν οι Rousseeuw και Ruts [RR96]. Ο αλγόριθμος αυτός παίρνει O(n) χρόνο μετά την ταξινόμηση των σημείων ακτινικά γύρω από το p. Είναι δυνατόν να ταξινομήσουμε όλα τα σημεία γύρω από κάθε σημείο του συνόλου των δεδομένων σε $O(n^2)$ χρόνο [Ede87] και επομένως η διαδικασία εύρεσης του βάθους ημιχώρου να απαιτεί μόνο O(n) χρόνο για κάθε σημείο του συνόλου των δεδομένων. Αυτό συνεπάγεται ότι η πολυπλοκότητα χρόνου του βήματος 2 μειώνεται σε $O(n^2)$ χρόνο. Η ίδια τεχνική μπορεί να χρησιμοποιηθεί και για να βελτιωθεί η πολυπλοκότητα χρόνου του υπολογισμού του τριγωνικού βάθους όλων των σημείων του συνόλου των δεδομένων που κάνουμε στο βήμα 1.

Τέλος μπορούμε να αναφερθούμε στο πως μπορούμε να εξαλείψουμε την αρχική υπόθεση των σημείων σε γενική θέση. Προφανώς τα βήματα 1 και 2 δεν επηρεάζονται από την παραπάνω υπόθεση, αφού και ο αλγόριθμος υπολογισμού της τιμής του τριγωνικού βάθους των Rousseeuw και Ruts [RR96] που μελετήσαμε στην ενότητα 3.2 λειτουργεί και για σύνολα σημείων δεδομένων που δεν βρίσκονται σε γενική θέση, αλλά και το βήμα 2 με οποιονδήποτε τρόπο υλοποιηθεί δεν λαμβάνει υπόψη του τον παραπάνω περιορισμό. Κατά την διάρκεια της επεξεργασίας του βήματος 3a για κάποιο ευθύγραμμο τμήμα λ, εάν συναντήσουμε ένα σημείο τομής το οποίο είναι ταυτόχρονα και σημείο του συνόλου των δεδομένων, μπορούμε απλά να σταματήσουμε την επεξεργασία του ευθύγραμμου τμήματος λ και να συνεχίσουμε την επεξεργασία του επόμενου ευθύγραμμου τμήματος. Αυτό μπορούμε να το κάνουμε επειδή το σημείο τομής θα σχηματίζει προφανώς ευθύγραμμα τμήματα με

καθένα από τα δύο σημεία-άκρα του λ και έτσι όλα τα σημεία του ευθύγραμμου τμήματος λ θα υποστούν επεξεργασία από τον αλγόριθμο οπωσδήποτε. Το παραπάνω γεγονός μας αφήνει με μία μόνο περίπτωση ακόμα ανοιχτή για να αντιμετωπίσουμε: Η προέκταση κάποιου τέμνοντος ευθύγραμμου τμήματος t περνά από c σημεία του δοθέντος συνόλου I. Το γεγονός αυτό μπορούμε να το αντιμετωπίσουμε εύκολα αφού γνωρίζουμε το ακριβές πλήθος των σημείων του συνόλου I που βρίσκονται αυστηρά σε κάποια πλευρά του t (γιατί έχει υπολογιστεί στο βήμα 2). Συνεπάγεται λοιπόν ότι θα έχουμε περισσότερα από ένα ευθύγραμμα τμήματα συνευθειακά με το t. Εάν αυτά είναι m στον αριθμό, τότε ισχυριζόμαστε πως πάνω σε αυτό το σημείο τομής είμαστε μέσα σε $\frac{m(c-2)}{2}$ τρίγωνα, εκτός από αυτά που υπολογίστηκαν από τον αλγόριθμο Simp-Med. Ο αριθμός αυτός προκύπτει αν ακολουθήσουμε τον παρακάτω συλλογισμό:

Με τα άκρα-σημεία ενός τέμνοντος ευθύγραμμου τμήματος και ένα από τα υπόλοιπα c σημεία δεδομένων, μπορούμε να σχηματίσουμε ένα τρίγωνο που να περιέχει το σημείο τομής αυτό. Με την παραπάνω διαδικασία και τα δύο συγκεκριμένα άκρα-σημεία μπορούμε να σχηματίσουμε c-2 τέτοια τρίγωνα. Ας υποθέσουμε ότι το q είναι το τρίτο σημείο για ένα από αυτά τα τρίγωνα. Τότε το q και ένα από τα δύο άκρα-σημεία θα σχηματίσουν ένα ακόμα τέμνον ευθύγραμμο τμήμα. Έτσι το ίδιο τρίγωνο θα μετρηθεί μία ακόμα φορά κατά την επεξεργασία του άλλου τμήματος. Επομένως έχουμε c-2 τρίγωνα να καταμετρούνται για καθένα από τα m ευθύγραμμα τμήματα και επιπλέον πολλαπλασιάζουμε με τον παράγοντα ½ για να αποφύγουμε την διπλή καταμέτρηση των τριγώνων αυτών.

ΚΕΦΑΛΑΙΟ 5

ΜΕΛΕΤΗ ΤΡΙΓΩΝΙΚΟΥ ΕΚΤΙΜΗΤΗ ΓΙΑ ΣΥΝΟΛΑ ΣΗΜΕΙΩΝ ΣΕ ΚΥΡΤΗ ΘΕΣΗ ΣΤΟ ΕΠΙΠΕΔΟ.

5.1. Εισαγωγή.

Στην μελέτη του τριγωνικού εκτιμητή ασχοληθήκαμε με σύνολα σημείων στο επίπεδο τα οποία βρίσκονται σε κυρτή θέση (είναι κορυφές του κυρτού περιβλήματος). Υλοποιήσαμε τους αλγορίθμους που αναφέρθηκαν στα προηγούμενα κεφάλαια (κεφάλαια 3 και 4) και επίσης υλοποιήσαμε ένα πρόγραμμα που κατασκευάζει σύνολα σημείων σε κυρτή θέση. Κατασκευάσαμε αρκετά σύνολα δεδομένων για πλήθος κόμβων από 4 μέχρι 20 κόμβους. Αναλυτικά τα στοιχεία των συνόλων δεδομένων φαίνονται στο παρακάτω πίνακα :

Πλήθος Κόμβων	Πλήθος Συνόλων Δεδομένων	Μέγεθος Εισόδου / Εξόδου
4	1	89 bytes / 1.3 KB
6	20	135 bytes / 5 KB

Πλήθος Κόμβων	Πλήθος Συνόλων Δεδομένων	Μέγεθος Εισόδου / Εξόδου
8	840	176 bytes / 16 KB
10	504	221 bytes / 45 KB
12	990	267 bytes / 102 KB
14	156	312 bytes / 196 KB
16	210	355 bytes / 345 KB
18	272	400 bytes / 584 KB
20	342	444 bytes / 927 KB

Πίνακας 5.1. Στοιχεία συνόλων δεδομένων.

Από τις ενδεικτικές τιμές μεγέθους της εισόδου και της εξόδου του αλγορίθμου με τις οποίες εργαστήκαμε φαίνεται ότι ενώ η είσοδος αυξάνει γραμμικά με την αύξηση των δεδομένων σημείων, η έξοδος αυξάνει πολυωνυμικά (γεγονός που συμφωνεί με την τάξη πλήθους $O(n^4)$ των υποψηφίων σημείων του αλγορίθμου). Έτσι στην προσπάθειά μας για να δούμε αν αυτή η τάξη μεγέθους ισχύει και για σημεία σε κυρτή θέση, αρχικά κάναμε μία πιθανοτική μελέτη για να μας βοηθήσει στην εξαγωγή συμπερασμάτων και να μας δείξει κάποιες από τις κατευθύνσεις στις οποίες πρέπει να κινηθούμε (τα συμπεράσματα της μελέτης αυτής παραθέτονται στην ενότητα 5.3). Επεξεργαστήκαμε τα σύνολα δεδομένων μας με το πρόγραμμα που υλοποιεί τους αλγορίθμους των δύο προηγουμένων κεφαλαίων και αφού ελέγξαμε τα αποτελέσματα συνάγαμε κάποια πολύ ενδιαφέροντα συμπεράσματα (που αναλύονται εκτεταμένα στην ενότητα 5.4). Ο ορισμός του προβλήματος και των δομών που χρησιμοποιούμε στην θεωρητική του μελέτη δίνονται στην επόμενη ενότητα 5.2.

5.2. Ορισμός του Προβλήματος και Σχετικών Δομών.

Το πρόβλημα που μελετάμε στο κεφάλαιο αυτό και αποτελεί υποπερίπτωση του γενικού προβλήματος που είδαμε στο πρώτο κεφάλαιο ορίζεται ως εξής:

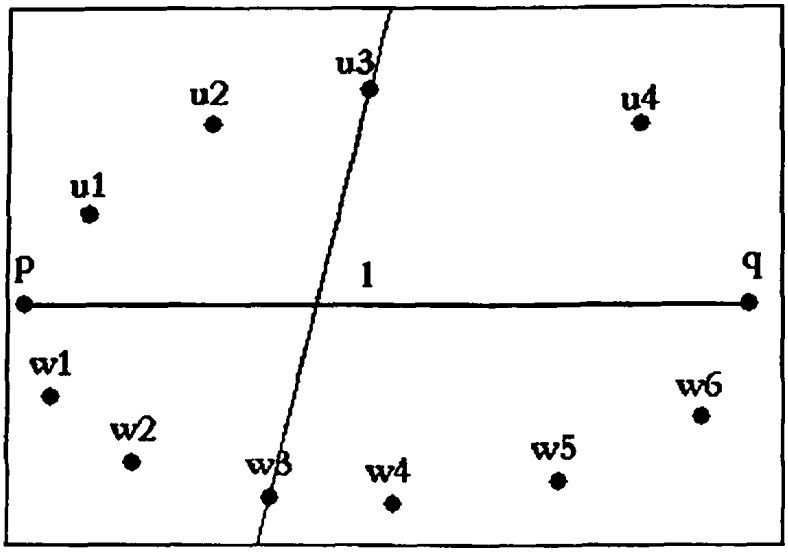
Ορισμός 5.1. *Έστω $S = \{X_1, ..., X_n\}$ ένα σύνολο από n δοθέντα σημεία τα οποία βρίσκονται σε κυρτή θέση στο επίπεδο. Να βρεθεί ένα σημείο p (ή σύνολο*

σημείων) που να περιγράφει (ή να αντιπροσωπεύει) με τον καλύτερο δυνατό τρόπο το σύνολο σημείων S χρησιμοποιώντας τον τριγωνικό εκτιμητή.

Δηλαδή θέλουμε να μελετήσουμε τον τριγωνικό εκτιμητή σε σύνολα δεδομένων που βρίσκονται σε κυρτή θέση στο επίπεδο. Στόχος μας είναι να απαντήσουμε αν το αντιπροσωπευτικό σημείο p βρίσκεται σε κάποια ιδιαίτερη θέση ή πάνω σε ευθεία που έχει κάποια συγκεκριμένη ιδιότητα, και / ή να δούμε αν μπορούμε να περιορίσουμε το πλήθος των υποψήφιων προς εξέταση σημείων που λαμβάνει υπόψη του ο αλγόριθμος Simp-Med και είναι της τάξης του $O(n^4)$ (πλήθος σημείων που είναι πολύ μεγάλο και οδηγεί σε πολύ «ακριβό» υπολογισμό).

Στην προσπάθεια μας αυτή θα μας βοηθήσει μια δομή δεδομένων, ο πίνακας των σημείων τομής T μίας ευθείας l, την οποία θα ορίσουμε παρακάτω. Ο πίνακας αυτός αποτελεί ένα πολύ χρήσιμο εργαλείο που θα μας βοηθήσει να καταλάβουμε καλύτερα τις πειραματικές διαπιστώσεις.

Ορισμός 5.2. *Ο πίνακας των σημείων τομής $T_{p,q}$ ενός ευθύγραμμου τμήματος l που ορίζεται από 2 σημεία p και q του δοθέντος συνόλου σημείων I (σε κυρτή θέση) είναι ένας πίνακας διάστασης $a \times b$, όπου a και b είναι τα πλήθη των σημείων του συνόλου I στα ημιεπίπεδα A και B αντίστοιχα (τα ημιεπίπεδα A και B είναι τα 2 ανοικτά ημιεπίπεδα που ορίζει ο φορέας του ευθύγραμμου τμήματος l και το A περιέχει το πολύ τόσα σημεία όσα και το B). Θεωρούμε ότι τα σημεία που βρίσκονται στο ημιεπίπεδο A είναι διατεταγμένα ως εξής: $p, u_1, \ldots, u_a, q$. Αντίστοιχα τα σημεία που βρίσκονται στο ημιεπίπεδο B είναι διατεταγμένα ως εξής: $p, w_1, \ldots, w_b, q$. Τότε στην θέση (i,j) του πίνακα $T_{p,q}$ βρίσκεται η καταχώρηση [+c\|-d], με c το πλήθος των σημείων του συνόλου I που βρίσκονται στο ανοικτό ημιεπίπεδο C, το οποίο ορίζεται από την ευθεία που περνάει από τα σημεία u_i και w_j και περιλαμβάνει το σημείο q, ενώ d είναι το πλήθος των σημείων του συνόλου I στο ανοικτό ημιεπίπεδο D, το οποίο ορίζεται από την ίδια ευθεία και περιλαμβάνει το σημείο p.*



Σχήμα 5.1. Κυρτό σύνολο 12 σημείων και διάσχιση ευθείας l.

Ένα παράδειγμα αυτού του πίνακα για το ευθύγραμμο τμήμα με άκρα p και q και για το σύνολο των 12 σημείων του σχήματος 5.1 φαίνεται παρακάτω:

	W_1	W_2	W_3	W_4	W_5	W_6
U_1	+9 | -1	+8 | -2	+7 | -3	+6 | -4	+5 | -5	+4 | -6
U_2	+8 | -2	+7 | -3	+6 | -4	+5 | -5	+4 | -6	+3 | -7
U_3	+7 | -3	+6 | -4	+5 | -5	+4 | -6	+3 | -7	+2 | -8
U_4	+6 | -4	+5 | -5	+4 | -6	+3 | -7	+2 | -8	+1 | -9

Πίνακας 5.2. Ο πίνακας $T_{p,q}$ του ευθύγραμμου τμήματος l του σχήματος 5.1.

Στον παραπάνω πίνακα φαίνεται η «συνεισφορά» κάθε ευθύγραμμου τμήματος που τέμνει το ευθύγραμμο τμήμα l του παραδείγματος του σχήματος 5.1. κατά την εφαρμογή του αλγορίθμου Simp-Med στην διάσχιση από το p έως το q. Ο αριθμός με το θετικό πρόσημο ισούται με το πλήθος των επιπλέον τριγώνων στα οποία εισερχόμαστε όταν φτάνουμε στο σημείο τομής που αντιστοιχεί στα συγκεκριμένα σημεία του πίνακα $T_{p,q}$ και θα αναφέρεται ως εξής ως **τιμή εισόδου** στο σημείο τομής. Όμοια, ο αριθμός με το αρνητικό πρόσημο ισούται με το πλήθος των τριγώνων από τα οποία εξερχόμαστε όταν αφήνουμε πίσω μας το σημείο τομής που αντιστοιχεί στα συγκεκριμένα σημεία του πίνακα $T_{p,q}$ και θα αναφέρεται στο εξής ως **τιμή εξόδου** στο σημείο τομής. Το άθροισμα των τιμών εισόδου και εξόδου σε κάποια θέση του πίνακα $T_{p,q}$ μας δίνει την συνολική «συνεισφορά» του αντίστοιχου σημείου τομής κατά την διάσχιση του ευθύγραμμου τμήματος l από το p στο q.

Ας πάρουμε για παράδειγμα το σημείο τομής r του ευθύγραμμου τμήματος l και του ευθύγραμμου τμήματος u_iw_i. Όταν κατά την διάσχιση του ευθύγραμμου τμήματος l φτάσουμε στο r, από την τιμή εισόδου του πίνακα $T_{p,q}$ για το συγκεκριμένο σημείο βλέπουμε ότι εισερχόμαστε σε 9 επιπλέον τρίγωνα. Ανάλογα όταν κατά την διάσχιση του l αφήνουμε πίσω μας το r, εξερχόμαστε αντίστοιχα, όπως παρατηρούμε και από την αντίστοιχη τιμή εξόδου του πίνακα $T_{p,q}$, από ένα τρίγωνο. Επομένως όταν έχουμε προσπεράσει το σημείο τομής r κατά την διάσχιση του l, έχουμε «μπει» σε 9 νέα τρίγωνα, ενώ έχουμε «βγει» από 1 τρίγωνο μέσα στο οποίο βρισκόμασταν μέχρι και την έξοδο από το σημείο r. Συνολικά, αμέσως αφού αφήσουμε πίσω μας το r, βρισκόμαστε μέσα σε 9-1=8 τρίγωνα επιπλέον από αυτά στα οποία ήμασταν ακριβώς πριν συναντήσουμε το r. Δηλαδή η τιμή εισόδου στο σημείο τομής r είναι 9 τρίγωνα, η τιμή εξόδου από το σημείο r είναι 1 τρίγωνο και η συνολική συνεισφορά του σημείου τομής r στην τιμή βάθους είναι 8 τρίγωνα.

Στο σημείο αυτό θα αναφέρουμε κάποιες ενδιαφέρουσες και χρήσιμες ιδιότητες του πίνακα $T_{p,q}$. Παρακάτω θα χρησιμοποιήσουμε τον συμβολισμό $T_{p,q}(i,j)[+]$ για την τιμή εισόδου του σημείου τομής των ευθύγραμμων τμημάτων pq και u_iw_j που αντιστοιχεί στην θέση (i,j) του πίνακα $T_{p,q}$. Ακόμη θα χρησιμοποιήσουμε τον συμβολισμό $T_{p,q}(i,j)[-]$ για την τιμή εξόδου του σημείου τομής των ευθύγραμμων τμημάτων pq και u_iw_j που αντιστοιχεί στην θέση (i,j) του πίνακα $T_{p,q}$. Τέλος θα χρησιμοποιήσουμε τον συμβολισμό $T_{p,q}(i,j)[=]$ για την τιμή συνολικής συνεισφοράς του σημείου τομής των ευθύγραμμων τμημάτων pq και u_iw_j που αντιστοιχεί στην θέση (i,j) του πίνακα $T_{p,q}$.

Ιδιότητα 5.1:

- *Η γραμμή i του πίνακα $T_{p,q}$ περιέχει τις τιμές εισόδου-εξόδου όλων των ευθυγράμμων τμημάτων που τέμνουν το pq και έχουν για ένα τους άκρο την κορυφή u_i.*
- *Η τιμή εισόδου του στοιχείου $T_{p,q}(i,j)$ είναι μικρότερη κατά 1 από την τιμή εισόδου του στοιχείου $T_{p,q}(i,j-1)$ της ίδιας γραμμής και της αμέσως προηγούμενης στήλης. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[+] = T_{p,q}(i,j-1)[+] - 1.}$
- *Η τιμή εξόδου του στοιχείου $T_{p,q}(i,j)$ είναι μεγαλύτερη κατά 1 από τη τιμή εισόδου του στοιχείου $T_{p,q}(i,j-1)$ της ίδιας γραμμής και της αμέσως προηγούμενης στήλης. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[-] = T_{p,q}(i,j-1)[-] + 1.}$

- *Η συνολική συνεισφορά του στοιχείου $T_{p,q}(i,j)$ είναι μικρότερη κατά 2 από τη τιμή εισόδου του στοιχείου $T_{p,q}(i,j-1)$ της ίδιας γραμμής και της αμέσως προηγούμενης στήλης. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[=] = T_{p,q}(i,j-1)[=] - 2.}$

Απόδειξη: Την ιδιότητα η γραμμή i του πίνακα $T_{p,q}$ να περιέχει τα σημεία τομής των ευθυγράμμων τμημάτων pq και αυτών που έχουν για ένα άκρο την κορυφή u_i την έχουμε από κατασκευής του πίνακα $T_{p,q}$ σύμφωνα με τον ορισμό 5.2. Ακόμα, η τιμή εισόδου του σημείου τομής του ευθύγραμμου τμήματος u_iw_j σε σχέση με αυτή του τμήματος u_iw_{j-1} είναι μικρότερη κατά ένα, αφού το πλήθος των σημείων που σχηματίζουν τρίγωνα στα οποία «εισερχόμαστε» όταν συναντούμε το u_iw_j μειώνεται κατά ένα σε σχέση με το πλήθος των σημείων όταν συναντούσαμε το u_iw_{j-1} (γιατί έχει αφαιρεθεί το σημείο w_j). Αντίστοιχα η τιμή εξόδου του σημείου τομής του u_iw_j αυξάνεται κατά ένα, αφού στο πλήθος των σημείων που σχηματίζουν τρίγωνα από τα οποία «εξερχόμαστε» όταν εξερχόμαστε από το u_iw_j, σε σχέση με το πλήθος των σημείων όταν εξερχόμαστε από το u_iw_{j-1}, προστίθεται το σημείο w_{j-1}. Τέλος, αφού σε σχέση με το u_iw_{j-1} η τιμή εισόδου για το u_iw_j μειώνεται κατά ένα, ενώ η τιμή εξόδου αυξάνεται κατά ένα, είναι προφανές ότι η συνολική συνεισφορά του u_iw_j θα μειώνεται κατά 2 σε σχέση με αυτή του u_iw_{j-1}. ⊡

Ιδιότητα 5.2:

- *Η στήλη j του πίνακα $T_{p,q}(i,j)$ περιέχει τις τιμές εισόδου-εξόδου όλων των ευθυγράμμων τμημάτων που τέμνουν το pq και έχουν για ένα τους άκρο την κορυφή w_j.*
- *Η τιμή εισόδου του στοιχείου $T_{p,q}(i,j)$ είναι μικρότερη κατά 1 από την τιμή εισόδου του στοιχείου $T_{p,q}(i-1,j)$ της ίδιας στήλης και της αμέσως προηγούμενης γραμμής. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[+] = T_{p,q}(i-1,j)[+] - 1.}$
- *Η τιμή εξόδου του στοιχείου $T_{p,q}(i,j)$ είναι μεγαλύτερη κατά 1 από τη τιμή εισόδου του στοιχείου $T_{p,q}(i,j-1)$ της ίδιας στήλης και της αμέσως προηγούμενης γραμμής. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[-] = T_{p,q}(i-1,j)[-] + 1.}$
- *Η συνολική συνεισφορά του στοιχείου $T_{p,q}(i,j)$ είναι μικρότερη κατά 2 από τη τιμή εισόδου του στοιχείου $T_{p,q}(i,j-1)$ της ίδιας στήλης και της αμέσως προηγούμενης γραμμής. Δηλαδή ισχύει :* $\mathbf{T_{p,q}(i,j)[=] = T_{p,q}(i-1,j)[=] - 2.}$

Απόδειξη : Η απόδειξη της ιδιότητας 5.2. είναι αντίστοιχη της απόδειξης της ιδιότητας 5.1. ▣

Ιδιότητα 5.3: *Το πλήθος των γραμμών κάθε πίνακα $T_{p,q}$ είναι μικρότερο ή ίσο από το πλήθος των στηλών του ίδιου πίνακα.*

Απόδειξη : Προκύπτει άμεσα από τον ορισμό 5.2. ▣

Ορισμός 5.3. *Στην περίπτωση που το πλήθος των γραμμών είναι ίσο με το πλήθος των στηλών στον πίνακα $T_{p,q}$, τότε λέμε ότι το ευθύγραμμο τμήμα pq λέγεται **διχοτόμος**.*

Ιδιότητα 5.4: *Η συνεισφορά των διχοτόμων σε κάθε πίνακα σημείων τομής $T_{p,q}$ είναι ίση με το μηδέν.*

Απόδειξη : Δεδομένου ότι καθένα από τα 2 ημιεπίπεδα που ορίζει ο φορέας της διχοτόμου περιέχει το ίδιο πλήθος σημείων από το δοθέν σύνολο I, συνεπάγεται ότι το πλήθος των τριγώνων, τα οποία προσθέτουμε όταν συναντάμε το σημείο τομής της διχοτόμου, είναι ίσο με το πλήθος των τριγώνων που αφαιρούμε όταν εξερχόμαστε από το σημείο αυτό. Οπότε η συνολική προσφορά της διχοτόμου είναι ίση με το μηδέν, γεγονός που φαίνεται και στον παράδειγμα του πίνακα 5.2. ▣

5.3. Πιθανοτική Μελέτη του Προβλήματος.

Σε αυτήν ενότητα θα εξετάσουμε το πρόβλημα που περιγράφεται στον ορισμό 5.1 κάτω από το πρίσμα των πιθανοτήτων. Θα προσπαθήσουμε να δούμε που είναι πιθανότερο να βρίσκεται το μέγιστο τριγωνικό βάθος και θα εξάγουμε τα πρώτα συμπεράσματά μας. Η μελέτη που αναφέρεται σε αυτήν την ενότητα δεν παίρνει υπόψη της όλους τους περιορισμούς του συγκεκριμένου προβλήματος, αλλά αναλύει μία έκδοση του προβλήματος με πιο χαλαρούς περιορισμούς. Αυτό συμβαίνει γιατί μια πλήρης πιθανοτική μελέτη του προβλήματος με όλους του τους περιορισμούς είναι εξαιρετικά πολύπλοκη, αλλά και δεν αποτελεί τον κύριο σκοπό αυτής της εργασίας.

Στην πιθανοτική μελέτη που θα ακολουθήσει θα χρησιμοποιήσουμε κάποιους συμβολισμούς που εξηγούνται στον παρακάτω πίνακα :

Συμβολισμός	Ορισμός
Τμήμα[a]	Ευθύγραμμο τμήμα με a σημεία του συνόλου I στο ημιεπίπεδο με τα λιγότερα σημεία και (n-2-a) σημεία του I στο έτερο ημιεπίπεδο. Προφανώς ισχύει a≤(n-2-a).
Διχοτόμος	Ευθύγραμμο τμήμα με ίσο πλήθος σημείων του συνόλου I και στα δύο ημιεπίπεδα που αυτό ορίζει. Αυτό ισχύει για a=(n-2)/2. Δηλαδή Διχοτόμος = Τμήμα[(n-2)/2].
Τμήμα[c|d]	Ευθύγραμμο τμήμα που έχει c σημεία του συνόλου I στο ένα ημιεπίπεδο και d σημεία του I στο άλλο ημιεπίπεδο.
Πλήθος[a][c|d]	Το πλήθος των τμημάτων Τμήμα[c|d] που τέμνουν ένα Τμήμα[a].

Πίνακας 5.3. Συμβολισμοί και ορισμοί της πιθανοτικής μελέτης.

Σε αυτό το σημείο θα αναφέρουμε κάποιες παρατηρήσεις για το πρόβλημα :

- Το πλήθος των σημείων τομής σε ένα Τμήμα[a] είναι ίσο με : $\{a*(n-2-a)\}$
- Το πλήθος των σημείων τομής μεγιστοποιείται για a=(n-2)/2. Δηλαδή το πλήθος των σημείων τομής μεγιστοποιείται πάνω στις Διχοτόμους.
- Πάνω σε ένα Τμήμα[a] έχουμε :

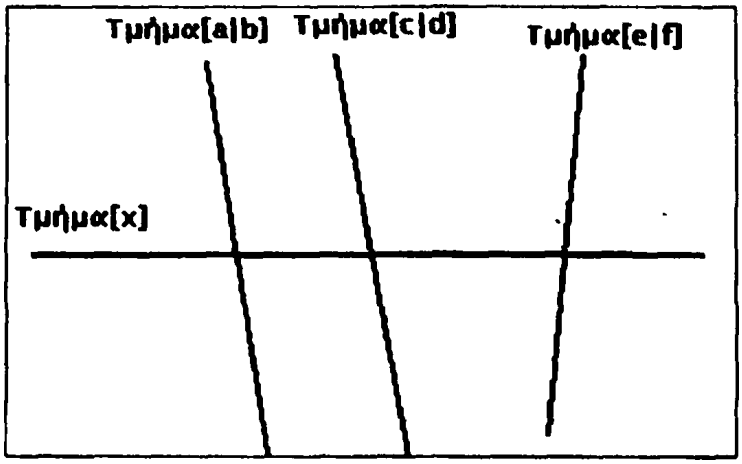
$$\text{Πλήθος}[a][c\,|\,d]=\begin{cases} a\text{ , εάν } c=d \\ \text{αλλιώς} \begin{cases} 2*c\text{ , εάν } c<=a \\ 2*a\text{ , εάν } c>a \end{cases} \end{cases}$$

Στον παραπάνω τύπο θεωρούμε πως c≤d. Ανάλογα με την φορά διάσχισης του Τμήματος[a] τα μισά Τμήματα[c|d] τα συναντάμε ως [+c|-d] και τα άλλα μισά ως [+d|-c].

- Έστω ότι το μέγιστο τριγωνικό βάθος βρίσκεται σε πάνω σε τομή ενός Τμήματος[x] και ενός Τμήματος[c|d]. Ακόμη έστω ότι το επόμενο σημείο τομής κατά την φορά διάσχισής του Τμήματος[x] είναι αυτό μεταξύ του Τμήματος[x] και του Τμήματος[e|f], ενώ το αμέσως προηγούμενο σημείο τομής είναι αυτό του Τμήματος[x] και του Τμήματος[a|b] (όπως φαίνεται και

στο σχήμα 5.2). Προφανώς για να έχω το μέγιστο τριγωνικό βάθος στο σημείο τομής του Τμήματος[x] και του Τμήματος[c|d] θα πρέπει να ισχύουν οι παρακάτω σχέσεις :

1) e < d , αλλιώς το μέγιστο δεν θα βρισκόταν πάνω στο Τμήμα[c|d].
2) f > c , επειδή e+f=c+d=n-2 και την σχέση 1).
3) b < c , αλλιώς το μέγιστο δεν θα βρισκόταν πάνω στο Τμήμα[c|d].
4) a > d , επειδή a+b=c+d=n-2 και την σχέση 3).



Σχήμα 5.2. Το σχήμα της πιθανοτικής μελέτης.

Από τους παραπάνω περιορισμούς μπορούμε να δούμε ποιο Τμήμα[c|d] είναι το πιο πιθανό για να έχει το μέγιστο τριγωνικό βάθος πάνω σε τομή του Τμήματος[x] που διασχίζει ο αλγόριθμος. Αυτό φαίνεται στους παρακάτω τύπους στους οποίους παρατηρούμε το πλήθος των ευθύγραμμων τμημάτων που μπορούν να είναι πριν και μετά το Τμήμα[c|d] για να μπορεί να έχει αυτό το μέγιστο τριγωνικό βάθος:

Συνθήκη	Πριν το Τμήμα[c|d]	Μετά το Τμήμα[c|d]
c < (n-2)/2	$\left(\sum_{i=1}^{c-1} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)$	$\left(\sum_{i=c}^{n-2} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)-1$
c = (n-2)/2	$\left(\sum_{i=1}^{c-1} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)+x-1$	$\left(\sum_{i=c+1}^{n-2} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)+x-1$

c > (n-2)/2	$\left(\sum_{i=1}^{c} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)-1$	$\left(\sum_{i=c+1}^{n-2} \frac{\Pi\lambda\acute{\eta}\theta o\varsigma[x][i \mid j]}{2}\right)$
Οι παραπάνω σχέσεις ισχύουν για 1 # x # (n-2)/2. Ακόμη, στα παραπάνω αθροίσματα ισχύει j = (n-2-i).		

Πίνακας 5.4. Πίνακας συνόλων που πληρούν τους χαλαρούς περιορισμούς.

Από τους παραπάνω τύπους για να υπολογίσω την πιθανότητα να βρίσκεται το μέγιστο τριγωνικό βάθος στο Τμήμα[c|d] θα πρέπει να υπολογίσω το πλήθος των ευθυγράμμων τμημάτων Τμήμα [a|b] που μπορούν να βρίσκονται πριν το Τμήμα[c|d], το πλήθος των ευθυγράμμων τμημάτων Τμήμα [e|f] που μπορούν να βρίσκονται μετά το Τμήμα[c|d], να πολλαπλασιάσω τους δύο αυτούς αριθμούς και να διαιρέσω με το συνολικό πλήθος όλων των δυνατών περιπτώσεων. Από τους τύπους του πίνακα 5.4 φαίνεται εύκολα ότι η πιθανότητα μεγιστοποιείται για c=(n-2)/2, δηλαδή για Διχοτόμους. Οδηγούμαστε στο συμπέρασμα αυτό αν παρατηρήσουμε ότι για συγκεκριμένο Τμήμα[x] το άθροισμα του πλήθους των ευθειών πριν (έστω α ευθείες) και μετά (έστω β ευθείες) κάθε τμήματος είναι σταθερό. Δηλαδή (α+β)=(σταθερό). Ο υπολογισμός της πιθανότητας στηρίζεται στο γινόμενο α*β, το οποίο μεγιστοποιείται όταν α=β. Αυτό γίνεται όπως φαίνεται από τους τύπους του πίνακα 5.4 για c=(n-2)/2, δηλαδή για Διχοτόμους.

5.3. Πειραματικά Συμπεράσματα και Σχολιασμός.

Σε αυτήν την ενότητα θα αναφερθούμε στα συμπεράσματα που εξαγάγαμε από την επεξεργασία των πειραματικών δεδομένων και στην συνέχεια θα προσπαθήσουμε να τα εξηγήσουμε με την βοήθεια του πίνακα των σημείων τομής Τ. Αλλά σε αυτό το σημείο είναι σκόπιμο να δούμε πιο προσεκτικά κάποιες ιδιότητες του πίνακα των σημείων τομής Τ και να δώσουμε κάποιους επιπλέον ορισμούς εννοιών που πρόκειται να χρησιμοποιήσουμε.

	W_1	W_2	W_3	W_4	W_5
U_1	+9 | -1	+8 | -2	+7 | -3	+6 | -4	+5 | -5
U_2	+8 | -2	+7 | -3	+6 | -4	+5 | -5	+4 | -6
U_3	+7 | -3	+6 | -4	+5 | -5	+4 | -6	+3 | -7
U_4	+6 | -4	+5 | -5	+4 | -6	+3 | -7	+2 | -8
U_5	+5 | -5	+4 | -6	+3 | -7	+2 | -8	+1 | -9

Πίνακας 5.5. Ο πίνακας Τ μίας Διχοτόμου του παραδείγματος της εικόνας 5.1.

Περιοχή Θετικής Βεβαιότητας u_3w_3

Περιοχές Αβεβαιότητας u_3w_3

Περιοχή Αρνητικής Βεβαιότητας u_3w_3

Από τις ιδιότητες 5.1. και 5.2. που είδαμε στην ενότητα 5.2. είναι εύκολο να δούμε πως όταν κατά την διάσχιση ενός Τμήματος[a] βρισκόμαστε πάνω σε σημείο τομής με το ευθύγραμμο τμήμα u_iw_j, τότε έχουμε περάσει σίγουρα από τα σημεία τομής των ευθυγράμμων τμημάτων u_kw_l, με k= 1 , ... , i και l= 1 , ... , j χωρίς το σημείο τομής του u_iw_j, και έχουμε δεχτεί την συνολική τους συνεισφορά (με συμβολισμό T(k,l)[=]) στην τρέχουσα τιμή βάθους. Ακόμα σύμφωνα με τις ίδιες ιδιότητες είναι προφανές πως όταν κατά την διάσχιση ενός Τμήματος[a] βρισκόμαστε πάνω σε σημείο τομής με το ευθύγραμμο τμήμα u_iw_j, τότε σίγουρα δεν έχουμε περάσει από τα σημεία τομής των ευθυγράμμων τμημάτων u_kw_l, με k= i , ... , a και l= j , ... , (n-2-a) χωρίς το σημείο τομής του u_iw_j, και δεν έχουμε προσθέσει την συνολική τους συνεισφορά στην τρέχουσα τιμή βάθους. Τις περιοχές αυτές του πίνακα Τ ενός Τμήματος[a] θα τις αναφέρουμε και ως **περιοχές βεβαιότητας** για κάποιο σημείο τομής των ευθυγράμμων τμημάτων u_iw_j και Τμήμα[a], αφού γνωρίζουμε με βεβαιότητα όταν βρισκόμαστε σε αυτό το σημείο τομής, είτε ότι έχουμε σίγουρα περάσει από κάποια σημεία τομής (και θα αναφέρεται ως **περιοχή θετικής βεβαιότητας** του πίνακα Τ ως προς το σημείο αυτό), είτε ότι δεν έχουμε περάσει από κάποια άλλα σημεία τομής (και θα αναφέρεται ως **περιοχή αρνητικής βεβαιότητας** του πίνακα Τ ως προς το σημείο αυτό). Οι δύο άλλοι υποπίνακες του πίνακα Τ θα αναφέρονται ως **περιοχές αβεβαιότητας** του πίνακα Τ. Οι δύο αυτοί υποπίνακες του πίνακα Τ ως προς το στοιχείο u_iw_j λέγονται έτσι, γιατί δεν είμαστε βέβαιοι αν έχουν συνεισφέρει στην τιμή τριγωνικού βάθους του σημείου τομής του ευθύγραμμου τμήματος u_iw_j.

Στο παράδειγμα του πίνακα 5.5. βλέπουμε τις περιοχές βεβαιότητας στον πίνακα μίας Διχοτόμου για το σημείο τομής του ευθύγραμμου τμήματος u_3w_3 (που είναι και αυτό μία Διχοτόμος), οι οποίες είναι σκιασμένες. Τα δύο λευκά τμήματα του πίνακα 5.5. αποτελούν τις περιοχές αβεβαιότητας του σημείο τομής του u_3w_3. Τέλος, το ίδιο το σημείο u_3w_3 ανήκει μισό στην θετική περιοχή βεβαιότητας (γιατί έχουμε προσθέσει την τιμή εισόδου στο τρέχον τριγωνικό βάθος, εφόσον έχουμε συναντήσει το σημείο αυτό) και μισό στην αρνητική περιοχή βεβαιότητας (γιατί δεν έχουμε αφαιρέσει την τιμή εξόδου από το τρέχον τριγωνικό βάθος, εφόσον δεν έχουμε φύγει από το σημείο αυτό).

Εξετάζοντας με προσοχή τα πειραματικά δεδομένα για σύνολα σημείων με τέσσερα έως είκοσι σημεία, τα οποία αναφέρονται στην ενότητα 5.1, παρατηρήσαμε πως οι μέγιστες τιμές τριγωνικού βάθους βρίσκονται πάντα πάνω σε διχοτόμους. Δεν μπορέσαμε να αποδείξουμε ότι αυτό ισχύει πάντα, αλλά αξίζει να σημειώσουμε ότι:

- Το γεγονός αυτό μπορεί να εξηγηθεί από τον πίνακα των σημείων τομής, αφού για τις διχοτόμους μεγιστοποιείται το πλήθος των σημείων τομής, επομένως και το πλήθος των στοιχείων του πίνακα Τ. Το γεγονός αυτό σε συνάρτηση με την δομή του πίνακα μας οδηγεί στην παρατήρηση πως μεγιστοποιείται και το μέγιστο δυνατό θετικό άθροισμα. Δηλαδή στις διχοτόμους το μέγιστο δυνατό τριγωνικό βάθος είναι μεγαλύτερο από το αντίστοιχο μέγιστο των άλλων ευθειών.
- Ακόμη στον πίνακα σημείων τομής κάθε ευθύγραμμου τμήματος τα στοιχεία που αντιστοιχούν σε σημεία τομής διχοτόμων είναι αυτά που μεγιστοποιούν το άθροισμα των συνεισφορών της περιοχής θετικής βεβαιότητας και ελαχιστοποιούν το άθροισμα των συνεισφορών της περιοχής αρνητικής βεβαιότητας ως προς όλα τα υπόλοιπα στοιχεία που βρίσκονται στην ίδια γραμμή ή την ίδια στήλη με αυτά. Δηλαδή έχουν την μεγαλύτερη σίγουρη θετική συνεισφορά στην τιμή τριγωνικού βάθους τους, ενώ ταυτόχρονα έχουν και την μεγαλύτερη απόλυτη τιμή αρνητικής συνεισφοράς που σίγουρα δεν έχει αφαιρεθεί από την τιμή τριγωνικού βάθους τους σε σχέση με τα υπόλοιπα στοιχεία που βρίσκονται στην ίδια γραμμή ή στήλη του πίνακα Τ.

Ειδικότερα, τα πειραματικά δεδομένα κατέδειξαν ότι οι μέγιστες τιμές τριγωνικού βάθους βρίσκονται πάνω σε τομές διχοτόμων. Αυτό το γεγονός μπορεί να

εξηγηθεί με την βοήθεια του πίνακα σημείων τομής Τ μίας διχοτόμου (όπως αυτόν του πίνακα 5.5) αν παρατηρήσουμε ότι τα στοιχεία που αντιστοιχούν σε σημεία τομής διχοτόμων μεγιστοποιούν το άθροισμα των συνεισφορών της περιοχής θετικής βεβαιότητας και ελαχιστοποιούν το άθροισμα των συνεισφορών της περιοχής αρνητικής βεβαιότητας ως προς όλα τα υπόλοιπα στοιχεία που βρίσκονται στην ίδια γραμμή ή την ίδια στήλη με αυτά. Στον πίνακα 5.5, τα στοιχεία που αντιστοιχούν σε διχοτόμους έχουν έντονες τις τιμές εισόδου και εξόδου, οι οποίες έχουν άθροισμα μηδέν γεγονός που συμφωνεί με την ιδιότητα 5.4. Ακόμη τα στοιχεία του πίνακα Τ που αντιστοιχούν σε τομές διχοτόμων έχουν την ιδιότητα ότι το άθροισμα των συνολικών συνεισφορών των περιοχών αβεβαιότητάς τους είναι ίσο με το μηδέν, ενώ από την δομή του πίνακα και τις ιδιότητες 5.1 και 5.2 είναι αδύνατο κάποια υποακολουθία σημείων τομής των περιοχών αβεβαιότητας να έχει άθροισμα συνεισφορών μικρότερο του μηδενός.

Σε περίπτωση που το δοθέν σύνολο διδιάστατων σημείων έχει μέγεθος ίσο με 4*m+4, όπου m θετικός ακέραιος αριθμός, για κάθε διχοτόμο υπάρχει μοναδική άλλη διχοτόμος με την ιδιότητα να διαχωρίζουν το υπόλοιπο σύνολο δεδομένων σημείων μεγέθους 4*m σε τέσσερα ισομεγέθη σύνολα των m σημείων στο επίπεδο. Την διχοτόμο αυτή θα την ονομάζουμε **κάθετη διχοτόμο** ως προς την αρχική μας διχοτόμο. Για τα πειραματικά δεδομένα συνόλων αυτού του μεγέθους (δηλαδή για 8, 12, 16 και 20 κόμβους) παρατηρήσαμε πως αν πάνω σε μία διχοτόμο υπάρχουν σημεία τομής που έχουν μέγιστο τριγωνικό βάθος, ένα από αυτά θα είναι σίγουρα το σημείο τομής r αυτής της διχοτόμου με την κάθετη διχοτόμο της. Στον πίνακα των σημείων τομής Τ, οι τιμές εισόδου και εξόδου του σημείου τομής r βρίσκονται στο κέντρο του πίνακα Τ, δηλαδή, στην θέση $T(m+1,m+1)$. Το στοιχείο αυτό του πίνακα που αντιστοιχεί στο σημείο τομής της διχοτόμου με την κάθετή της διχοτόμο, μεγιστοποιεί το άθροισμα των συνολικών συνεισφορών των στοιχείων της περιοχής θετικής βεβαιότητας, ενώ ταυτόχρονα μεγιστοποιεί και την απόλυτη τιμή του αθροίσματος των συνεισφορών των στοιχείων της περιοχής αρνητικής βεβαιότητας σε σχέση με όλα τα αντίστοιχα αθροίσματα των περιοχών βεβαιότητας των υπόλοιπων στοιχείων του πίνακα Τ.

Η απόδειξη των παραπάνω παρατηρήσεων συνεπάγεται σημαντικά οφέλη στην πολυπλοκότητα του αλγορίθμου υπολογισμού του τριγωνικού εκτιμητή για σύνολα διδιάστατων δειγμάτων σε κυρτή θέση. Συγκεκριμένα:

a) Αν αποδειχτεί ότι το μέγιστο τριγωνικό βάθος βρίσκεται πάνω σε διχοτόμο, τότε δεν είναι αναγκαίο να ελέγξουμε και τα $O(n^4)$ σημεία τομής, αλλά μόνο τα $O(n^3)$ σημεία τομής των διχοτόμων με όλα τα άλλα ευθύγραμμα τμήματα. Δηλαδή μειώνουμε την πολυπλοκότητα του αλγορίθμου κατά μία τάξη μεγέθους.

b) Αν αποδειχτεί ότι το μέγιστο τριγωνικό βάθος βρίσκεται πάνω σε σημείο τομής διχοτόμων, τότε είναι απαραίτητος ο έλεγχος μόνο των $O(n^2)$ σημείων τομής διχοτόμων. Δηλαδή καταφέρνουμε να μειώσουμε την πολυπλοκότητα του αλγορίθμου κατά δύο τάξεις μεγέθους.

c) Τέλος, για δείγματα μεγέθους 4*m+4 (με m θετικό ακέραιο αριθμό) εάν αποδειχτεί ότι ένα σημείο με μέγιστο τριγωνικό βάθος βρίσκεται πάνω σε σημείο τομής διχοτόμου με την αντίστοιχη κάθετη της διχοτόμο και ζητάμε ένα απλώς σημείο (και όχι όλα τα σημεία) με μέγιστο τριγωνικό βάθος, τότε ελέγχουμε μόνο τα $O(n)$ σημεία τομής των διχοτόμων με τις αντίστοιχες κάθετες διχοτόμους τους. Στην περίπτωση αυτή μειώνουμε το πλήθος των σημείων που πρέπει να εξετάσουμε κατά τρεις τάξεις μεγέθους.

	W_1	W_2	W_3
U_1	+5 | -1	+4 | -2	+3 | -3
U_2	+4 | -2	+3 | -3	+2 | -4
U_3	+3 | -3	+2 | -4	+1 | -5

Πίνακας 5.6. Ο πίνακας Τ μίας Διχοτόμου ενός συνόλου 8 σημείων.

Αν και δεν ήτανε δυνατή η απόδειξη των παραπάνω προτάσεων και παρατηρήσεων, ο πίνακας των σημείων τομής Τ μας παρέχει την απόδειξη των παραπάνω ισχυρισμών για μικρό αριθμό σημείων (μέχρι 8 σημεία). Είναι εύκολο να δούμε ότι για 8 σημεία, το στοιχείο που αντιστοιχεί στην κάθετη διχοτόμο έχει στην περιοχή θετική βεβαιότητας του όλες τις θετικές συνεισφορές, ενώ στην περιοχή αρνητικής βεβαιότητας του όλες τις αρνητικές συνεισφορές (όπως φαίνεται και στον πίνακα 5.6). Παράλληλα, η συνολική συνεισφορά κάθε στοιχείου στις περιοχές αβεβαιότητας είναι ίση με το μηδέν. Δηλαδή, στην τιμή τριγωνικού βάθους της κάθετης διχοτόμου έχουν προστεθεί όλες οι θετικές συνεισφορές, ενώ δεν έχει προστεθεί καμία αρνητική συνεισφορά. Αν συνυπολογίσουμε και το γεγονός ότι οι πίνακες των διχοτόμων έχουν άθροισμα θετικών συνεισφορών τουλάχιστον ίσο ή και

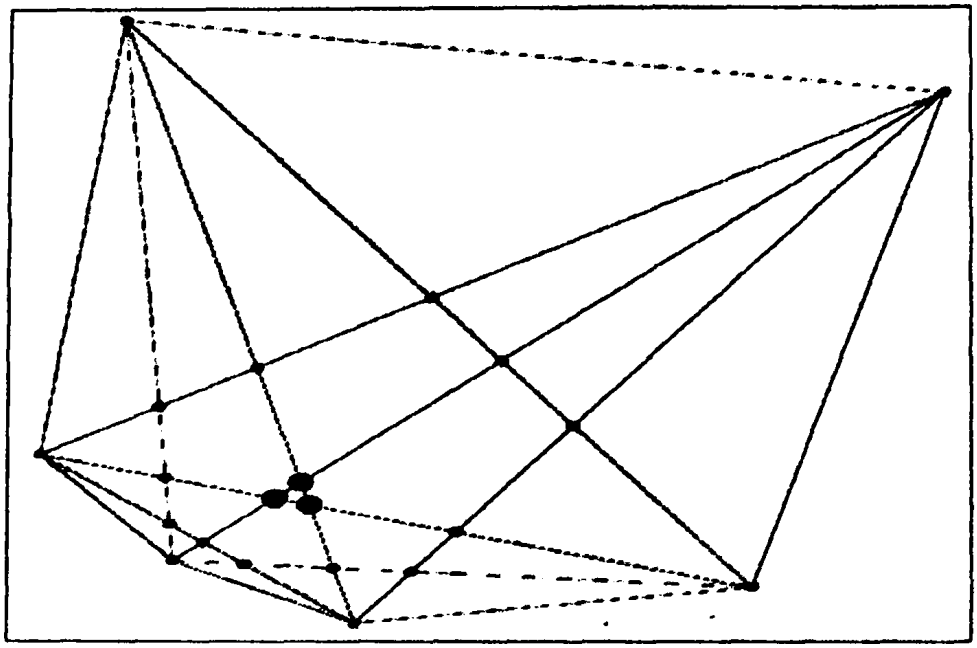
μεγαλύτερο από τους αντίστοιχους πίνακες των υπολοίπων ευθειών, καταλήγουμε στην απόδειξη του ισχυρισμού μας ότι κάποιο από τα σημεία με μέγιστο τριγωνικό βάθος για ένα σύνολο 8 σημείων θα βρίσκεται σίγουρα σε σημείο τομής διχοτόμων, και μάλιστα διχοτόμου με την αντίστοιχη κάθετή της διχοτόμο.

Η παραπάνω μελέτη και οι παρατηρήσεις μας έγιναν για σύνολα με άρτιο πλήθος σημείων, όπου ήταν πιο προφανής η εξαγωγή συμπερασμάτων και ευχερέστερη η μαθηματική τους διατύπωση. Όμως εργαστήκαμε και με αρκετά σύνολα πειραματικών δεδομένων με περιττό αριθμό σημείων, όπου οι παρατηρήσεις μας ήτανε ανάλογες με αυτές για άρτια σύνολα σημείων. Οι σημαντικότερες διαφοροποιήσεις μεταξύ περιττών και άρτιων συνόλων σημείων δίνονται παρακάτω:

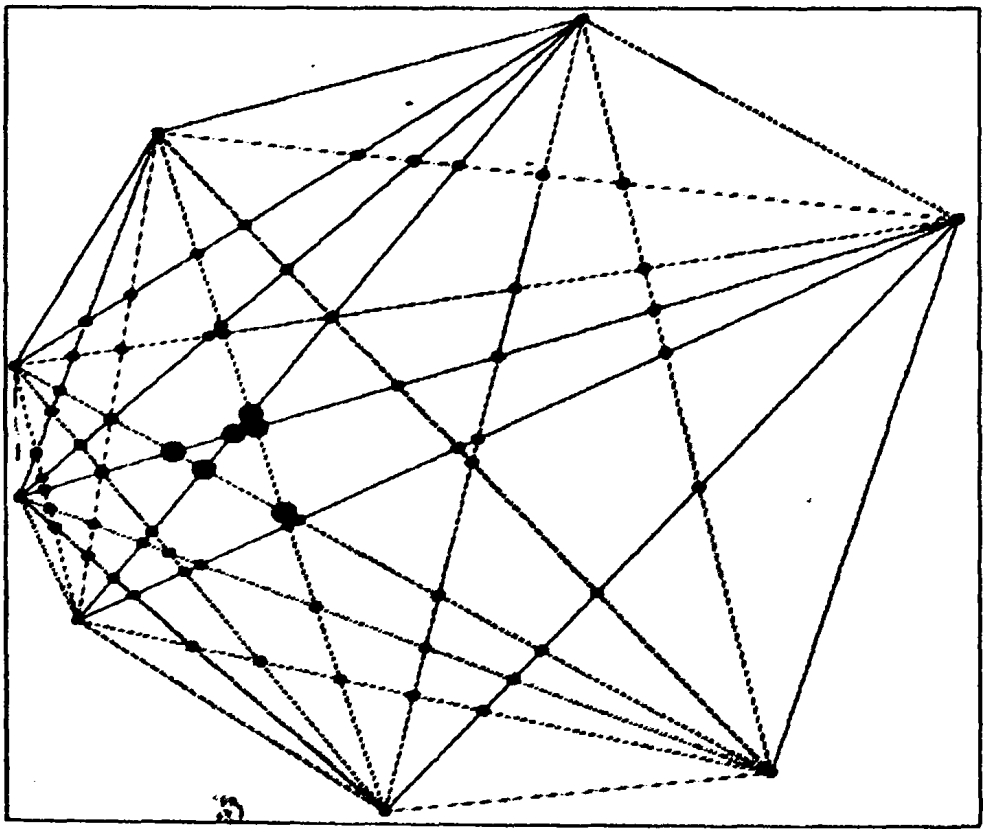
- Στα περιττά σύνολα σημείων I, η διχοτόμος ορίζεται ως το ευθύγραμμο τμήμα με πλήθος σημείων του συνόλου I το οποίο διαφέρει το πολύ κατά ένα και στα δύο ημιεπίπεδα που αυτό ορίζει.
- Κάθε σημείο του περιττού συνόλου σημείων I δεν αποτελεί το άκρο μίας διχοτόμου, αλλά δυο. Δηλαδή οι διχοτόμοι είναι n και όχι n/2, όπως στα άρτια σύνολα σημείων.
- Στα περιττά σύνολα σημείων I, η κάθετη διχοτόμος μίας διχοτόμου ορίζεται ως το ευθύγραμμο τμήμα που τέμνει την διχοτόμο και το πλήθος των σημείων του συνόλου I διαφέρει το πολύ κατά ένα στα τεταρτημόρια που ορίζονται στο επίπεδο από τα δύο αυτά ευθύγραμμα τμήματα.
- Προφανώς για κάθε διχοτόμο δεν υπάρχει μοναδική κάθετη διχοτόμος.
- Στον πίνακα των σημείων τομής η μία κάθετη διχοτόμος μεγιστοποιεί την συνεισφορά της θετική περιοχής βεβαιότητας, ενώ η άλλη κάθετη διχοτόμος μεγιστοποιεί την απόλυτη τιμή της συνεισφοράς της αρνητικής περιοχής βεβαιότητας.

Σε αυτό το σημείο μπορούμε να αναφέρουμε ότι επεξεργαστήκαμε και λίγα σύνολα σημείων σε κυρτή θέση στο επίπεδο με περισσότερα σημεία (5 σύνολα με 30 σημεία, 5 σύνολα με 40 σημεία και 5 σύνολα με 50 σημεία), αλλά αυτά ήτανε ελάχιστα σε σχέση με αυτά που αναφέρουμε στον πίνακα 5.1 και παρόλο που επιβεβαίωναν τις παρατηρήσεις μας δεν μπορούμε να βγάλουμε ασφαλή συμπεράσματα.

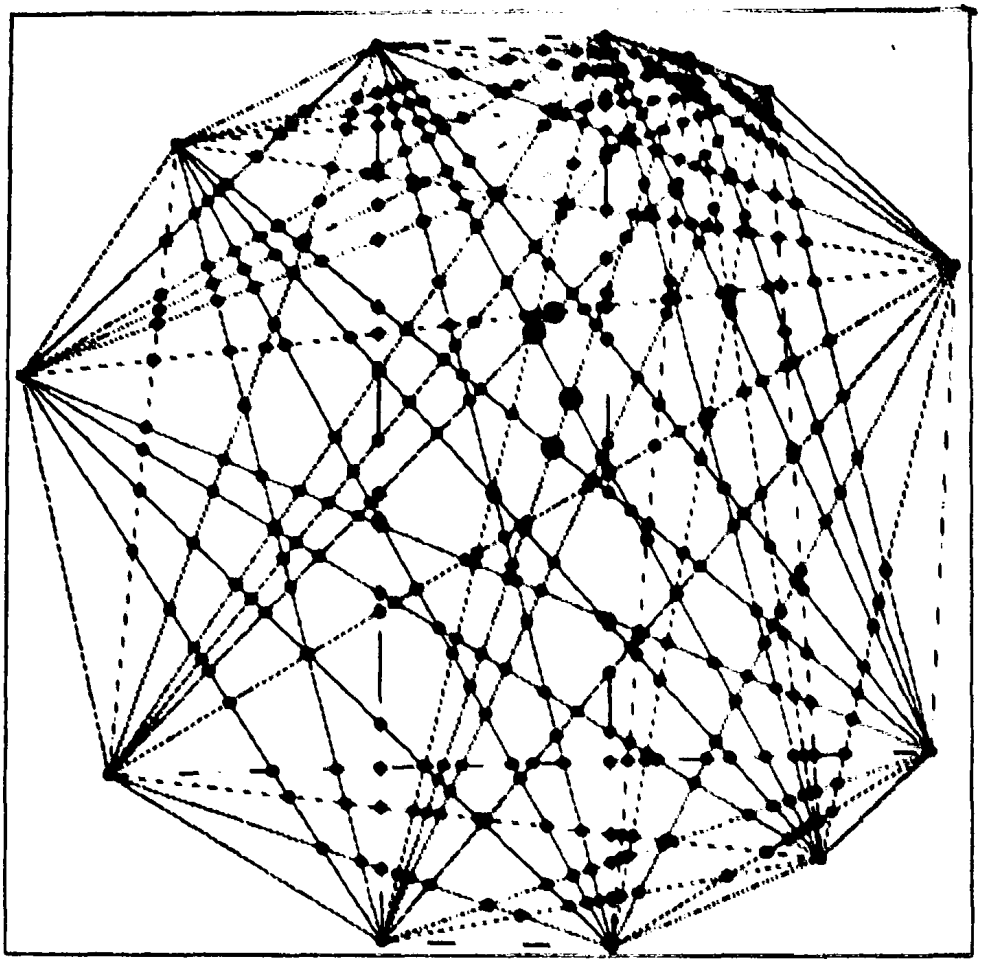
Στο τέλος θα παραθέσουμε τα σχήματα μερικών υλοποιήσεων για άρτιο αριθμό σημείων(6, 8 και 12).



Σχήμα 5.3. Σύνολο 6 σημείων σε κυρτή θέση και σημεία με μέγιστο τριγωνικό βάθος.



Σχήμα 5.4. Σύνολο 8 σημείων σε κυρτή θέση και σημεία με μέγιστο τριγωνικό βάθος.



Σχήμα 5.5. Σύνολο 12 σημείων σε κυρτή θέση και σημεία με μέγιστο τριγωνικό βάθος.

ΚΕΦΑΛΑΙΟ 6

ΕΠΙΛΟΓΟΣ

6.1. Εν Κατακλείδι.

Στην εργασία αυτή αρχικά ερευνήσαμε και μελετήσαμε την δουλειά που έχει γίνει στον τομέα των πολυδιάστατων εκτιμητών μεσοστοιχείου. Με την μελέτη της υπάρχουσας βιβλιογραφίας παρατηρήσαμε το πλαίσιο (το οποίο αναφέρεται συνοπτικά στα κεφάλαια 1 και 2) πάνω στο οποίο μπορούσαμε να στηρίξουμε τα δικά μας συμπεράσματα. Επόμενο βήμα ήταν η ειδικότερη και πιο διεξοδική μελέτη του τριγωνικού βάθους και του τριγωνικού εκτιμητή. Εξετάσαμε τους υπάρχοντες αλγορίθμους και υλοποιήσαμε τους καλύτερους μέχρι σήμερα (οι οποίοι εξετάζονται αναλυτικά στα κεφάλαια 3 και 4). Μετά εξετάσαμε το πρόβλημα του αντιπροσωπευτικού σημείου για σύνολα σημείων στο επίπεδο τα οποία βρίσκονται σε κυρτή θέση. Με την βοήθεια των υλοποιήσεων των αλγορίθμων κάναμε παρατηρήσεις πάνω σε έναν μεγάλο αριθμό πειραματικών συνόλων δεδομένων και θέσαμε το θεωρητικό υπόβαθρο για περαιτέρω κατανόηση και μελέτη του προβλήματος (που αναφέρονται στο κεφάλαιο 5).

6.2. Μελλοντική Εργασία.

Ως μελλοντική εργασία σαφώς απομένει η δύσκολη δουλειά της απόδειξης των παρατηρήσεων χρησιμοποιώντας το υπόβαθρο που αναπτύξαμε στο κεφάλαιο 5 αυτής της εργασίας. Σίγουρα θα βοηθούσαν αποδείξεις λημμάτων για την ύπαρξη της βέλτιστης διαδρομής ως προς τις συνεισφορές των σημείων τομής πάνω στον πίνακα σημείων τομής Τ μίας διχοτόμου ή λημμάτων για την ύπαρξη ακολουθίας σημείων τομής πάνω σε μία τουλάχιστον διχοτόμο που να υλοποιούν το (μέγιστο) άθροισμα όλων των θετικών συνεισφορών για όλα τα κυρτά σύνολα δεδομένων. Τέλος, η απόδειξη της ορθότητας αυτών των παρατηρήσεων θα επιτρέψει πιθανώς την εξαγωγή αντιστοίχων προτάσεων για το γενικό πρόβλημα και άρα την καλύτερη μελέτη του.

ΠΑΡΑΡΤΗΜΑ Α

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕΤΡΗΣΗΣ ΕΥΣΤΑΘΕΙΑΣ ΕΚΤΙΜΗΤΩΝ

Μία συνάρτηση βάθους **D(.;.)** για να θεωρείται ευσταθής και ταυτόχρονα να αποτελεί ένα εργαλείο για την αποτελεσματική ταξινόμηση (ordering) από το κέντρο προς το περίβλημα σημείων στον R^d είναι απαραίτητο να ικανοποιεί όσο το δυνατόν περισσότερα από τα παρακάτω επιθυμητά χαρακτηριστικά (τα οποία αναφέρονται για πρώτη φορά από την Liu [Liu90] στην μελέτη της σχετικά με την συνάρτηση του τριγωνικού βάθους). Θα αναφέρουμε την κλάση όλων των κατανομών στον R^d ως **F** και για οποιοδήποτε τυχαίο διάνυσμα **X** την αντίστοιχη συνάρτηση βάθους ως $\mathbf{F_x}$ (συμβολισμός κατά Zuo και Serfling [ZS00]).

- **P1: Αμεταβλητότητα σε γραμμικούς μετασχηματισμούς.**

  $\mathbf{D(Ax+b\ ;\ F_{Ax+b}) = D(x\ ;\ F_x)}$ για οποιοδήποτε τυχαίο διάνυσμα X στον R^d, για κάθε $d \times d$ nonsingular πίνακα A και κάθε d-διάστατο διάνυσμα b.

- **P2: Μεγιστοποίηση στο Κέντρο.**

  Για κάθε $f \in F$ η οποία έχει κέντρο το σημείο θ (π.χ. το σημείο συμμετρίας σχετικά με κάποια έννοια συμμετρίας) θα πρέπει να ισχύει $\mathbf{D(\theta;f) = \sup_x D(x\ ;\ f)}$, για κάθε $x \in R^d$.

- **P3: Μονοτονικότητα ως προς το Βαθύτερο Σημείο.**

  Για κάθε $f \in F$ η οποία έχει ως βαθύτερο σημείο το θ (π.χ. το σημείο με το μέγιστο βάθος) θα πρέπει να ισχύει **D(x ; f) ≤ D(θ + a*(x-θ) ; f)**, για κάθε a∈[0,1].

- **P4: Μηδενισμός στο άπειρο.**

  Για κάθε f∈**F** ισχύει **D(x ; f)→0** καθώς το $\|x\| \to \infty$.

- **P5: Σημείο Κατάρρευσης.**

  Αποτελεί ίσως το κυριότερο μέτρο αξιολόγησης της ευστάθειας ενός εκτιμητή. Το σημείο κατάρρευσης (breakdown point) που αντιπροσωπεύει το σύνολο δεδομένων μας αποτελεί το ποσοστό των σημείων του συνόλου που πρέπει να μετακινηθούν στο άπειρο ώστε και ο εκτιμητής να πράξει ανάλογα [Sma90]. Υπάρχουνε όμως και δύο ορισμοί-εκδοχές του παραπάνω ορισμού που αναφέρονται από τους Donoho και Huber [DH83] και οι οποίοι έχουνε επικρατήσει ως οι πιο ποσοτικές εκτιμήσεις της συνολικής ευστάθειας ενός εκτιμητή και είναι το σημείο κατάρρευσης λόγω προσθήκης (addition breakdown point ή ABP) και το σημείο κατάρρευσης λόγω αντικατάστασης (replacement breakdown point ή RBP). Έστω $X = \{X_1, ..., X_n\}$ ένα σύνολο n σημείων στον R^d. Τότε το σημείο κατάρρευσης λόγω προσθήκης ενός εκτιμητή T σε συνάρτηση με το σύνολο σημείων X ορίζεται ως:

$$ABP(X,T) = \min\left\{\frac{m}{m+n} : \sup_{Y^m}\left\|T\left(X \cup Y^m\right) - T(X)\right\| = \infty\right\}$$

όπου το Y^m αντιστοιχεί σε ένα σύνολο m σημείων τυχαίων τιμών και το $X \cup Y^m$ αντιστοιχεί στο σύνολο που προκύπτει από την ένωση των συνόλων X και Y^m. Αντίστοιχα το σημείο κατάρρευσης λόγω αντικατάστασης ενός εκτιμητή T σε συνάρτηση με το σύνολο σημείων X ορίζεται ως:

$$RBP(X,T) = \min\left\{\frac{m}{n} : \sup_{X^m}\left\|T(X^m) - T(X)\right\| = \infty\right\}$$

όπου το X^m αντιστοιχεί σε ένα σύνολο n σημείων που προκύπτει από το X με την αντικατάσταση m σημείων του με τυχαίες τιμές σημείων. Με απλά λόγια, το σημείο κατάρρευσης λόγω προσθήκης είναι το ελάχιστο κλάσμα προσθήκης και το σημείο κατάρρευσης λόγω αντικατάστασης είναι το ελάχιστο κλάσμα αντικατάστασης αντίστοιχα για τα οποία ο εκτιμητής δεν φράσσεται (δηλαδή, τείνει στο άπειρο).

ΠΑΡΑΡΤΗΜΑ Β

ΧΡΗΣΙΜΟΙ ΟΡΙΣΜΟΙ ΚΑΙ ΘΕΩΡΗΜΑΤΑ

- **Θεμελιώδες θεώρημα του Caratheodory:** Κάθε σημείο του κυρτού περιβλήματος ενός συνόλου σημείων S στον R^d βρίσκεται μέσα στον κυρτό συνδυασμό d+1 ή λιγότερων σημείων του S.

- **Θεώρημα του Helly:** Εάν F είναι μία οικογένεια από περισσότερα των d φραγμένων κυρτών συνόλων στον R^d και εάν κάθε H_d (όπου H_d είναι ο αριθμός του Helly) μέλη της F έχουν ένα τουλάχιστον κοινό σημείο , τότε όλα τα μέλη της F έχουν τουλάχιστον ένα κοινό σημείο.

- **Αριθμός του Helly:** Δεδομένου ενός Ευκλείδειου χώρου διάστασης d, ο αριθμός του Helly είναι $H_d = d+1$.

- **Θεώρημα του Radon:** Κάθε σύνολο d+2 σημείων στον R^d μπορεί πάντα να διαχωριστεί σε δύο υποσύνολα V_1 και V_2, τέτοια ώστε τα κυρτά περιβλήματα των δύο υποσυνόλων να τέμνονται.

- **Regression:** Μία μέθοδος για «ταίριασμα» μίας καμπύλης (όχι απαραίτητα μίας ευθείας γραμμής) σε ένα σύνολο από σημεία χρησιμοποιώντας ένα κριτήριο ταιριάσματος. Ο πιο κοινός τύπος regression είναι ο γραμμικός.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[ABET00] N.Amenta, M.Bern, D.Eppstein, and S.Teng. Regression depth and center points, *Discrete and Computational Geometry,* 123(3):305-323,2000.

[ACG+01] G.Aloupis, C.Cortes, F.Gomez, M.Soss, and G.Toussaint. Lower bounds for computing statistical depth, *Technical Report SOCS-01.1,* School of Computer Science, McGill University, February 2001.

[Alo01] G.Aloupis. On computing geometric estimators of location, *M.Sc. thesis,* McGill University, 2001.

[ALST01] G.Aloupis, S.Langerman, M.Soss, and G.Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines, *Submitted for publication,* 2001.

[AST01] G.Aloupis, M.Soss, and G.Toussaint. On the computation of the bivariate median and a Fermat-Torricelli problem, *Technical Report SOCS-01.2,* School of Computer Science, McGill University, February 2001.

[Avi82] D.Avis. On the complexity of finding the convex hull of a set of points, *Discrete Applied Math*, 4:81-86, 1982.

[Bal95] I.Balaban. An optimal algorithm for finding segment intersections, in *Proc. of the 11th Annual ACM Symposium of Computational Geometry*, pages 211-219, 1995.

[Bar76] V.Barnett. The ordering of multivariate data, *Journal of the Royal Statistical Society ser.A*, 139:318-355, 1976.

[Bar82] I.Barany. A generalization of Caratheodory's theorem, *Discrete Math*, 40:141-150, 1982.

[Bar00] R.Barbara. The Fermat-Torricelli points of n lines, *Mathematical Gazette*, 84:24-29, 2000.

[Bas91] G.Bassett. Equivariant monotonic, 50% breakdown estimators, *The American Statistician*, 45(2):135-137, 1991.

[BF84] E.Boros and Z.Furedi. The maximal number of covers by the triangles of a given vertex set on the plane, *Geometriae Dedicata*, 17:69-77, 1984.

[Cha85] B.Chazelle. On the convex layers of a planar set, *IEEE Transactions on Information Theory*, IT-31(4):509-517, 1985.

[CLR90] T.Cormen, C.Leiserson, and R.Rivest. Introduction to Algorithms, *MIT Press*, 1990.

[CO98] A.Cheng and M.Ouyang. On algorithms for simplicial depth, *Technical Report dcs-tr-368*, Department of Computer Science, Rutgers University, 1998.

[CSY87] R.Cole, M.Shamir, and K.Yap. On k-hulls and related problems, *SIAM J. Comput.*, 16(1):61-77, 1987.

[DG92] D.Donoho and M.Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *Annals of Statistics*, 20:1803-1827, 1992.

[DH83] D.Donoho and P.Huber. The notion of breakdown point, in P.Bickel, K.Doksum, and J.Hodges (editors), *A Festschrift for Erich L.Lehmann*, pages 157-184, Belmont, California, 1983, Wadsworth International Group.

[Don82] D.Donoho. Breakdown properties of multivariate location estimators, *Ph.D. thesis*, Harvard University, 1982.

[Ede87] H.Edelsbrunner. Algorithms in Combinatorial Geometry, *Springer-Verlag*, Berlin, 1987.

[Eel30] W.Eells. A mistaken conception of the center of the population, *Journal of the American Statistical Association*, 25:33-40, 1930.

[Gal33] L.Galvani. Sulla determinazione del centro di gravita e del centro mediano di una populazione, *Metron*, 11(1):17-47, 1933.

[Gow74] J.Gower. The mediancentre, *Applied Statistics*, 23(3):466-470, 1974.

[Gre81] P.J.Green. Peeling Bivariate Data, in V.Barnett (editor), *Interpreting Multivariate Data*, New York, Wiley, 1981.

[Gri27] F.Griffin. Abstract: Points of minimum travel for a distributed population, *Bulletin of the American Mathematical Society*, 33:516, 1927.

[GS98] C.Gross and T.Strempel. On generalizations of conics and on a generalization of the Fermat-Torricelli problem, *American Mathematical Monthly*, 105(8):732-743, 1998.

[GSW92] J.Gill , W.Steiger, and A.Wigderson. Geometric medians, *Discrete Mathematics*, 108:37-51, 1992.

[Ham68] R.Hampel. Contributions to the theory of robust estimation, *Ph.D. thesis*, University of California, Berkley, 1968.

[Har69] F.Harary. Graph Theory, *Addison-Wesley*, Reading, Massachusetts, 1969.

[Hot29] H.Hotelling. Stability in competition, *Economic Journal*, 39:41-57, 1929.

[Hub72] P.Huber. Robust statistics: A review, *The Annals of Mathematical Statistics*, 43(3):1041-1067, 1972.

[Jar73] R.Jarvis. On the identification of the convex hull of a finite set of points in the plane, *Information Processing Letters*, 2:18-21, 1973.

[JM94] S.Jadhav and A.Mukhopadhyay. Computing a centerpoint of a finite planar set of points in linear time, *Discrete and Computational Geometry*, 12:291-312, 1994.

[KM89] S.Khuller and J.Mitchel. On a triangle counting problem, *Information Processing Letters*, 33:319-321, 1989.

[Lan01] S.Langerman. Algorithms and Data Structures in Computational Geometry, *Ph.D. thesis*, Department of Computer Science, Rutgers University, New Brunswick, 2001.

[Liu90] R.Liu. On the notion of data depth based upon random simplices, *The Annals of Statistics*, 18:405-414, 1990.

[Liu95] R.Liu. Control charts for multivariate processes, *Journal of the American Statistical Association*, 90(432):1380-1387, 1995.

[Lop92] H.Lopuhaa. Highly efficient estimator of the multivariate location, *The Annals of Statistics*, 20(1):398-413, 1992.

[LPS99] R.Liu, J.Parelius, and K.Singh. Multivariate analysis of data depth: descriptive statistics and inference, *Annals of Statistics*, 27(2):783-858, 1999.

[LS00] S.Langerman and W.Steiger. The complexity of the hyperplane depth in the plane, in *Japan Conference on Discrete and Computational Geometry*, November 2000.

[Mat91] J.Matousek. Computing the center of planar point sets, *In J.Goodman, R.Pollack, and W.Steiger (editors), Computational Geometry: Papers from the DIMACS special year*, volume 6, pages 221-230, *American Mathematical Society*, 1991.

[Moo41] A.Mood. On the joint distribution of the medians in samples from a multivariate population, *The Annals of Mathematical Statistics*, 12:268-278, 1941.

[MRR+01] K.Miller, S.Ramaswami, P.Rousseeuw, T.Sellares, D.Souvaine, I.Streinu, and A.Struyf. Fast implementation of depth contours using topological sweep, in *Proc. 12th Symposium on Discrete Algorithms (SODA)*, Washington D.C., 2001.

[NON92] A.Niinimaa, H.Oja, and J.Nyblom. Algorithm AS 277: The Oja bivariate median, *Applied Statistics*, 41:611-617,1992.

[NOT90] A.Niinimaa, H.Oja, and M.Tableman. The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version, *Statistics and Probability Letters*, 10:325-328, 1990.

[Oja83] H.Oja. Descriptive statistics for multivariate distributions, *Statistics and Probability Letters*, 1:327-332, 1983.

[ON89] H.Oja and J.Nyblom. Bivariate sign tests, *Journal of the American Statistical Association*, 84(405):249-259, 1989.

[O'R95] J.O'Rourke. Computational Geometry in C, *Cambridge University Press*, 1995.

[OL81] M.Overmars and J.van Leeuwen. Maintenance of configurations in the plane, *Journal of Comput. And Syst. Sci.*, 23:166-204, 1981.

[PS85] F.Preparata and M.Shamos. Computational Geometry, An Introduction, *Springer-Verlag*, 1985.

[RH99a] P.Rousseeuw and M.Hubert. Depth in an arrangement of hyperplanes, *Discrete Computational Geometry*, 22:167-176, 1999.

[RH99b] P.Rousseeuw and M.Hubert. Regression Depth, *Journal of the American Statistical Association*, 94:388-402, 1999.

[RL91] P.Rousseeuw and H.Lopuhaa. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, *The Annals of Statistics*, 119(1):229-248, 1991.

[Rou85] P.Rousseeuw. Multivariate estimation with high breakdown point, in W.Grossman, G.Pflug, I.Vincze, and W.Wertz (editors), *Mathematical Statistics and Applications*, Volume B, pages 283-297, Dordrecht, Reidel Publishing Company, 1985.

[RR96] P.Rousseeuw and I.Ruts. Bivariate location depth, *Applied Statistics*, 45:516-526, 1996.

[RR98] P.Rousseeuw and I.Ruts. Constructing the bivariate Tukey median, *Statistica Sinica*, 8:828-839, 1998.

[RR99] P.Rousseeuw and I.Ruts. The depth function of a population distribution, *Metrica*, 49(3):213-244, 1999.

[Sca33] D.Scates. Locating the median of the population in United States, *Metron*, 11(1):49-65, 1933.

[Sha76] M.Shamos. Geometry and statistics: Problems at the interface, in J.Traub (editor), *Recent Results and New Directions in Algorithms and Complexity*, pages 251-280, Academic Press, 1976.

[Sma90] C.Small. A survey of multidimensional medians, *International Statistical Review*, 58:263-277, 1990.

[Sta81] W.Stahel. Robust Estimation: Infinitesimal optimality and covariance matrix estimators, *Ph.D. thesis*, ETH, Zurich, 1981.

[Tit78] D.Titterington. Estimation of correlation coefficients by ellipsoidal trimming, *Applied Statistics*, 27:227-234, 1978.

[TP79] G.T.Toussaint and R.S.Poulsen. Some new algorithms and software implementation methods for pattern recognition research, *in Proc. of the IEEE Computer Society Conference on Computer Software*, pages 55-63, 1979.

[Tuk75] J.Tukey. Mathematics and the picturing of data, *in Proceedings of the International Congress of Mathematicians*, pages 523-531, Vancouver, 1975.

[vKMR99] M.van Kreveld, J.Mitchell, P.Rousseeuw, M.Sharir, J.Snoeyink, and B.Speckmann. Efficient algorithms for maximum regression depth, *in Proceedings of the 15th Symposium on Computational Geometry*, ACM, pages 31-40, 1999.

[Web09] A.Weber. Uber den Standort der Industrien, Tubingen, *English translation by C.Friedrich (1929), Alfred Weber's Theory of Location of Industries*, University of Chicago Press, 1909.

[Yao81] A.Yao. A lower bound to finding convex hulls, *Journal of the ACM*, 28(4):780-787, 1981.

[ZS00] Y.Zuo and R.Serfling. Structural properties and convergence results for contours of sample statistical depth functions, *Final version to appear in Annals of Statistics*, 2000.