ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Πολυδιάστατα Μοντέλα Περιγραφής Σφαιρικών και Ελλειπτικών Δεδομένων

Παντελεήμων Μαθιουδάκης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Αθήνα
Φεβρουάριος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μάρκος Κούτρας (Επιβλέπων)
- Αναπληρωτής Καθηγητής Δημήτριος Αντζουλάκος
- Αναπληρωτής Καθηγητής Γεώργιος Τζαβελάς

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS
School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN **APPLIED STATISTICS**

Multivariate models for Spherical and Elliptical Data

Panteleimon Mathioudakis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Athens
February 2023

Στους γονείς μου και στον αδερφό μου.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κύριο Μάρκο Κούτρα που ως επιβλέπων της Διπλωματικής Διατριβής, η βοήθειά του αποδείχθηκε καταλυτική για την ολοκλήρωσή της.
Κλείνοντας να ευχαριστήσω και την οικογένειά μου που στέκεται πάντα βοήθημα και υποστηρικτής στις προσπάθειές μου.

ΠΕΡΙΛΗΨΗ

Στη Στατιστική είναι γνωστό πως χρησιμοποιούνται κατά κόρον πιθανοθεωρητικά μοντέλα και Κατανομές με σκοπό την ερμηνεία του πληθυσμού μελέτης. Αναντίρρητα, όσο αυξάνεται η πολυπλοκότητα των δεδομένων που λαμβάνονται, τόσο επιτακτική είναι η ανάγκη ανάπτυξης προχωρημένων μεθόδων ανάλυσής τους. Ειδικά στην Θεωρία Κατανομών έχει υπάρξει πρόοδος που μπορεί να καλύψει τον απαιτητικό χώρο της Ανάλυσης Δεδομένων, τόσο για Μονοδιάστατες μεταβλητές όσο και για Πολυδιάστατες. Κύριος στόχος της Διπλωματικής είναι η παρουσίαση, σε θεωρητικό αλλά και πρακτικό επίπεδο με πραγματικά δεδομένα, της πολύ χρήσιμης οικογένειας των Σφαιρικών και Ελλειπτικών Κατανομών.

Abstract

Probabilistic models and Distributions are used quite often in Statistics to describe the available data and carry out Analysis and Inference on them. Without any doubt as the complexity of data rises, the need for advanced analytical methods arises as well. Hopefully, large advances in Distribution theory guarantee the existence of state of the art models and Distributions that can cope with this ever-growing complexity in unidimensional and multidimensional data. The main goal of this Dissertation is to present the theory of the family of Spherical and Elliptical Distributions and highlight its use in practical problems.

Keywords:
Spherical distributions, Elliptical distributions, Multivariate models, Clustering algorithms, Discriminant Analysis

Κατάλογος Περιεχομένων

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Αλγορίθμων Python

Κεφάλαιο 1

Εισαγωγικά Στοιχεία για Πολυδιάστατες Τυχαίες Μεταβλητές

1.1 Εισαγωγή

Στη Στατιστική η τυχαία μεταβλητή (τ.μ.) που περιγράφει κάποιο χαρακτηριστικό αποτελεί το βασικό στοιχείο ανάλυσης. Αναντίρρητα, κάθε τυχαία μεταβλητή, συγκεκριμένα στην παραμετρική Στατιστική, διέπεται απο κάποιους κανόνες που αποτυπώνουν την συμπεριφορά της. Στοιχεία που αφορούν την θέση της όπως ο μέσος όρος ή η διάμεσος αλλά και μέτρα διασποράς που συμπεριλαμβάνουν διάφορα στατιστικά μέτρα όπως η διακύμανση (το άπλωμα των δεδομένων περί του μέσου όρου) και η κύρτωση (αφορά την ύπαρξη ακραίων παρατηρήσεων) δίνουν συγκεκριμένο αποτύπωμα και μας βοηθούν να κατανοήσουμε καλύτερα τη συμπεριφορά της. Όλη αυτή η πληροφορία εμπεριέχεται στην κατανομή της μεταβλητής.

Στην Πολυμεταβλητή Στατιστική Ανάλυση, η κύρια βαρύτητα συγκεντρώνεται στη μελέτη πολλών τυχαίων μεταβλητών $X_1, X_2, \ldots, X_p$ ταυτοχρόνα. Συγκεκριμένα το ενδιαφέρον έγκειται στη μελέτη της απο κοινού κατανομής των δεδομένων. Η μαθηματική έκφραση αυτής είναι η απο κοινού συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) για συνεχείς μεταβλητές, ενώ για διακριτά δεδομένα η απο κοινού συνάρτηση πιθανότητας (σ.π.). Για να υπολογίσουμε την πιθανότητα σε ένα υποσύνολο του πεδίου ορισμού της θα υπολογίσουμε το ολοκλήρωμα της απο κοινού σ.π.π. (εάν πρόκειται για συνεχή μεταβλητή) ή το άθροισμα της σ.π. στον αντίστοιχο χώρο που μας ενδιαφέρει.

1.2 Απο Κοινού Συνάρτηση Πιθανότητας

Έστω $X_1, X_2, \ldots X_p$ ένα σύνολο τριών τυχαίων μεταβλητών, τέτοιες ώστε $X_i = X_i(\omega)$ με $\omega \in \boldsymbol{\Omega}$. Το σύνολο $\boldsymbol{\Omega}$ αποτελείται απο όλα τα δυνατά αποτελέσματα ενός πειράματος, ενώ $\boldsymbol{D}_{X_i}$ είναι το σύνολο όλων των τιμών $X_i(\omega)$, με $X_i(\omega)$ απεικόνιση στο $\mathbb{R}$. Επιπρόσθετα, το $\boldsymbol{D_X}$ το πεδίο τιμών της απο κοινού τυχαίας μεταβλητής $\boldsymbol{X} = (X_1, X_2, \ldots X_p)^T$. Τότε ορίζεται η απο κοινού αθροιστική κατανομή πιθανότητας (CDF) η οποία γράφεται ως :

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_p \leq x_p)$$

Αν και θα επισυγκεντρωθούμε περισσότερο στις συνεχείς μεταβλητές, για $X_1, X_2, \ldots, X_p$ διακριτές τυχαίες μεταβλητές (π.χ. $X_i \in \{0, 1, 2, ..\}\ \forall i \in \{1, \ldots, p\}$), η συνάρτηση πιθανότητας $P(x_1, x_2, \ldots x_p)$ έχει κάποιες αξιοσημείωτες ιδιότητες :

- $P(x_1, x_2, \ldots x_p) \geq 0\ \forall (x_1, x_2, \ldots, x_p) \in \boldsymbol{D_X}$
- $\sum_{x_1} \sum_{x_2} \ldots \sum_{x_p} P(x_1, x_2, \ldots, x_p) = 1.$

Οι παραπάνω ιδιότητες είναι όμοιες με αυτές που ισχύουν για συνεχείς τ.μ. αρκεί να αντικαταστήσουμε τα αθροίσματα με ολοκληρώματα. Για $X_1, X_2, \ldots, X_p$ συνεχείς τ.μ. ορίζεται απο κοινού αθροιστική συνάρτηση πιθανότητας $F(x_1, x_2, \ldots, x_p) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_p \leq x_p)$, η οποία αποτελεί το ολοκλήρωμα μιας συνάρτησης πυκνότητας πιθανότητας $f(x_1, x_2, \ldots, x_p)$. Γενικά ενδιαφερόμαστε για κατανομές απολύτως συνεχείς για τις οποίες έχουμε :

$$\frac{\partial^p F(x_1, x_2, \ldots x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p} = f(x_1, x_2, \ldots, x_p). \tag{1.1}$$

Συγκεκριμένα, η ιδιότητα (1.1) περιορίζει το σύνολο των κατανομών διαισθητικά σε αυτές που ως προς το σχήμα τους δεν έχουν ασυνέχειες, ενώ παράλληλα είναι διαφορίσιμες σχεδόν παντού.

Στην μονοδιάστατη περίπτωση για μια μεταβλητή X, η Αθροιστική Συνάρτηση πιθανότητας ωρίζεται ως $F(x) = \int_{-\infty}^{x} f(u)\, du$ που ερμηνεύεται ως η πιθανότητα η μεταβλητή X να πάρει το πολύ την τιμή x. Στην παρακάτω ενότητα θα οριστούν οι πιθανότητες στην περίπτωση των πολλών διαστάσεων (κάθε διασταση αποτελεί μία μεταβλητή).

1.3 Σχέση Αθροιστικής Συνάρτησης Κατανομής και Συνάρτησης Πιθανότητας

Στην περίπτωση των πολλών διαστάσεων p, έχοντας την πυκνότητα $f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ ή απο κοινού πιθανότητα για διακριτά δεδομένα $P(x_1, \ldots, x_p)$, οι αθροιστικές συναρτήσεις πιθανότητας δίνονται απο τους τύπους :

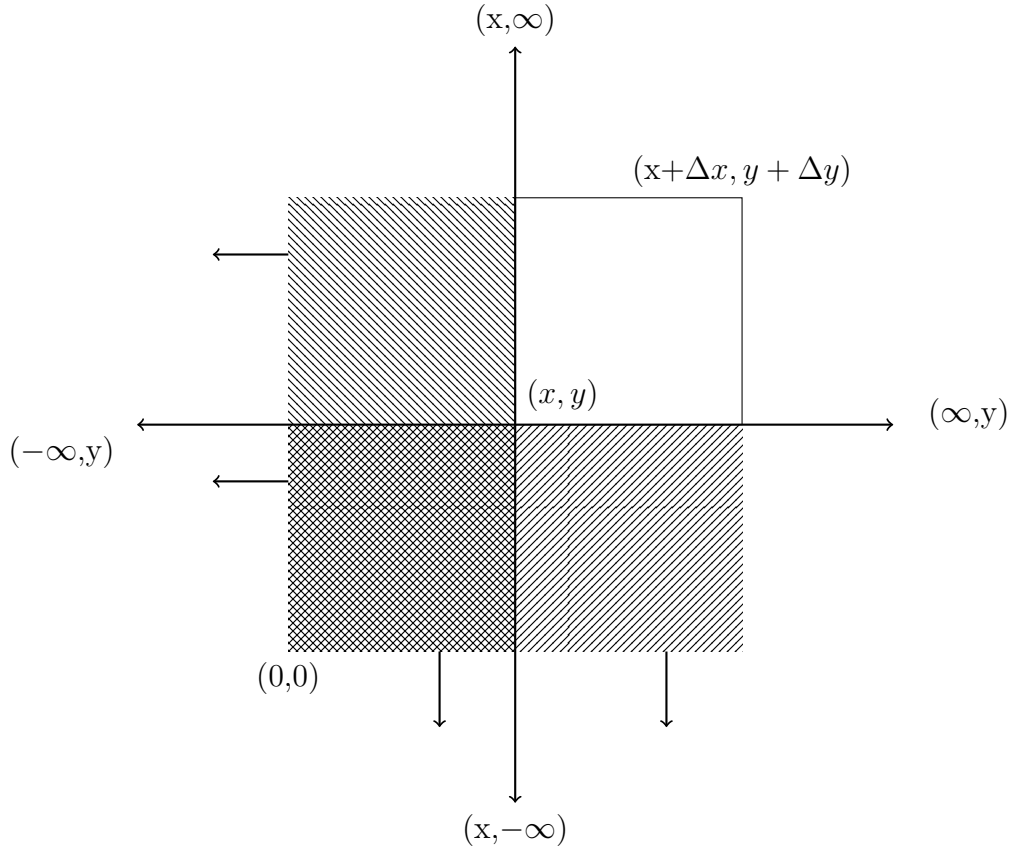
$$F(x_1, x_2, \ldots x_p) = \begin{cases} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_p} f(u_1, u_2, \ldots, u_p)\, du_1\, du_2 \cdots du_p & \text{εάν } X_i \text{ Συνεχής τ.μ.} \\ \sum_{u_1=0}^{x_1} \sum_{u_2=0}^{x_2} \ldots \sum_{u_p=0}^{x_p} P(u_1, u_2, \ldots, u_p) & \text{εάν } X_i \text{ Διακριτή τ.μ.} \end{cases}$$

Επιπλέον έχοντας την πυκνότητα $f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ ή απο κοινού πιθανότητα για διακριτά δεδομένα $P(x_1, \ldots, x_p)$, οι πιθανότητες σε διάφορα υποσύνολα δίνονται απο τους τύπους :

$$P(\boldsymbol{X} \in A) = \int \int \underset{\boldsymbol{X} \in A}{\ldots} \int f(x_1, \ldots, x_p)\, dx_1 \cdots dx_p \qquad \text{εάν } \boldsymbol{X} \in \mathbb{R}^p, \boldsymbol{A} \subseteq \mathbb{R}^p$$

$$P(\boldsymbol{X} \in A) = \sum \sum \underset{\boldsymbol{X} \in A}{\ldots} \sum P(x_1, \ldots, x_p) \qquad \text{εάν } \boldsymbol{X} \text{ διάνυσμα διακριτών τ.μ.}$$

Στη συνέχεια θα παρουσιαστεί ένα παράδειγμα στις δύο διαστάσεις μαζί με γράφημα για την καλύτερη κατανόηση του ορισμού της πιθανότητας σε υποσύνολα. Πολλές φορές ενδιαφερόμαστε να βρούμε την πιθανότητα δύο μεταβλητές X και Y να ανήκουν εντός ορισμένων διαστημάτων $[x, x+\Delta x]$, $[y, y+\Delta y]$ αντίστοιχα, με $\Delta x > 0$, $\Delta y > 0$. Συγκεκριμένα ζητείται η $P(x \leq X \leq x+\Delta x,\, y \leq Y \leq y+\Delta y) = F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y) - F(x, y+\Delta y) + F(x,y) = \int_y^{y+\Delta y} \int_x^{x+\Delta x} f(u,v)\, du\, dv$.



Η ζητούμενη πιθανότητα $P(x \leq X \leq x+\Delta x,\, y \leq Y \leq y+\Delta y)$ που είναι το (X,Y) να ανήκει εντός του λευκού παραλληλόγραμμου, είναι το $F(x+\Delta x, y+\Delta y)$ (η προέκταση του λευκού παραλληλογράμμου όπως δείχνουν τα μικρά βελάκια) μείον η περιοχή που ορίζεται ως προέκταση του γκρί παραλληλογράμου, μείον η αντίστοιχη προέκταση του άλλου γκρί, σύν την προέκταση του σκούρου παραλληλογράμου. Παρατηρείστε πως αυτές οι περιοχές που αφαιρούνται, αποτελούν περιοχές με αντίστοιχες πιθανότητες $F(x+\Delta x, y)$ και $F(x, y+\Delta y)$. Ωστόσο επειδή η πιθανότητα $F(x,y)$ του σκούρου επιπέδου έχει αφαιρεθεί δύο φορές, γι αυτό προστίθεται μία φορα για να μείνει εν τέλει μόνο το σύνολο $[x, x+\Delta x] \times [y, y+\Delta y]$, του οποίου η πιθανότητα είναι $P(x \leq X \leq x+\Delta x,\, y \leq Y \leq y+\Delta y) = F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y) - F(x, y+\Delta y) + F(x,y) = \int_x^{x+\Delta x} \int_y^{y+\Delta y} f(u,v)\, du dv$.

1.4 Σημαντικές Κατανομές και Μέτρα

Ένα άλλο είδος κατανομής που μπορεί να μας ενδιαφέρει είναι η περιθώρια κατανομή. Αν έχουμε τις τυχαίες μεταβλητές $X_1, X_2 \ldots X_p$, η περιθώρια συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της X_1 δίνεται ως :

$$f(x_1) = \int_{x_2} \int_{x_3} \cdots \int_{x_p} f(x_1, x_2, \ldots, x_p)\, dx_2 \cdots dx_p.$$

Δηλαδή ολοκληρώνουμε τη συνάρτηση πυκνότητας ως προς όλες τις υπόλοιπες μεταβλητές.

Επιπρόσθετα, άλλη μια χρήσιμη κατανομή για την Στατιστική είναι η δεσμευμένη, όταν δηλαδή γνωρίζουμε τις τιμές καποιων μεταβλητών (που σημαίνει πως πλέον είναι σταθερές). Ορίζεται ως εξής :

$$f(x_1|x_2,\ldots,x_p) = \frac{f(x_1,x_2,\ldots,x_p)}{f(x_2,x_3,\ldots,x_p)}.$$

Εν συνεχεία, ο μέσος όρος αποτελεί σημαντικό χαρακτηριστικό στην θεωρία κατανομών, και ορίζεται όπως παρακάτω :

$$\mathbb{E}(X_1) = \int x_1 f(x_1)\, dx_1$$

Τελικώς, η έννοια της συνδιακύμανσης είναι αναπόσπαστο κομμάτι στην ανάλυση των δεδομένων, ειδικά όταν διατίθενται πολλές μεταβλητές, διότι χαρακτηρίζεται ως μέτρο που φανερώνει τις σχέσεις μεταξύ των μεταβλητών, δηλαδή κατα πόσο συσχετίζονται μεταξύ τους. Ο τύπος της :

$$\begin{aligned}\mathrm{Cov}(X_1,X_2) = \mathbb{E}[(X_1-\mathbb{E}(X_1))\cdot(X_2-\mathbb{E}(X_2))] = \\ \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2).\end{aligned}$$

Βασικές ιδιότητες της συνδιακύμανσης :

- $\mathrm{Cov}(X_1,X_2) = 0$, εάν X_1, X_2 Ανεξάρτητες (το αντίστροφο δεν ισχύει πάντα),
- $\mathrm{Cov}(X_1+X_2,X_3) = \mathrm{Cov}(X_1,X_3) + \mathrm{Cov}(X_2,X_3)$,
- $\mathrm{Cov}(\alpha X_1+\beta,\gamma X_2+\delta) = \alpha\gamma\mathrm{Cov}(X_1,X_2)$.

Βέβαια, εφόσον οι μεταβλητές που συνήθως διατίθονται είναι πολλές, μπορεί η πληροφορία του μέσου και των διακυμάνσεων-συνδιακυμάνσεων να συμπτυχθεί με μορφή πινάκων χρησιμοποιώντας γραμμική άλγεβρα. Έστω ένα διάνυσμα τ.μ. $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$. Τότε ο μέσος όρος και ο πίνακας διακύμανσης-συνδιακύμανσης ορίζεται με τον εξής τρόπο :

$$\mathbb{E}(\boldsymbol{X}) = \begin{pmatrix}\mathbb{E}(X_1)\\ \vdots \\ \mathbb{E}(X_p)\end{pmatrix},$$

$$\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix}\mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \ldots & \mathrm{Cov}(X_1,X_p)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2,X_p)\\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_p,X_1) & \mathrm{Cov}(X_p,X_2) & \ldots & \mathrm{Var}(X_p).\end{pmatrix}$$

Οι παραπάνω μορφές πινάκων μας δίνουν όλη την δυνατή πληροφορία για το μέσο όρο των μεταβλητών αλλά και για την διακύμανση και την συνδιακύμανσή τους (κατα πόσο δηλαδή συσχετίζονται). Παρατηρείστε πως η συνδιακύμανση της μεταβλητής με τον εαυτό της είναι στην ουσία η διακύμανση.

Στην περίπτωση που έχουμε μία καινούρια μεταβλητή $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{X}$, με $\boldsymbol{C} = (c_1, c_2, \ldots, c_p)$ σταθερές συνιστώσες του διανύσματος, τότε $\mathbb{E}(\boldsymbol{Y}) = \mathbb{E}(\boldsymbol{C}\cdot\boldsymbol{X}) = \boldsymbol{C}\cdot\mathbb{E}(\boldsymbol{X})$ και $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{C}\cdot\mathrm{Cov}(\boldsymbol{X})\cdot\boldsymbol{C}^T$

Κεφάλαιο 2

Η Οικογένεια των Σφαιρικών και των Ελλειπτικών Κατανομών

2.1 Εισαγωγή

Ως γνωστόν, απο την θεωρία κατανομών γνωρίζουμε αρκετές χρήσιμες πυκνότητες πιθανοτήτων οι οποίες χρησιμοποιούνται ευρέως για στατιστική συμπερασματολογία όπως ελέγχους υποθέσεων αλλά και για την ερμηνεία του πληθυσμού που αναλύεται. Τέτοιες κατανομές αφορούν τις πολλοί γνωστές όπως η Κανονική, η Student-t και λοιπές. Παρατηρώντας τα χαρακτηριστικά και την μορφή τους μπορούμε να συμπεραίνουμε πως έχουν αρκετές κοινές ιδιότητες όπως η συμμετρία (πχ μονοκόρυφες συμμετρικές πυκνότητες πιθανοτήτων). Στην πραγματικότητα αυτές οι κατανομές δύναται να ενοποιηθούν υπο μία γενική οικογένεια που ορίζεται ως Σφαιρικές και Ελλειπτικές κατανομές. Με άλλα λόγια αποτελούν συναρτήσεις πυκνοτήτων που αν αναπαραχθούν γραφικώς θα μοιάζουν με κυκλικούς δίσκους στις δύο διαστάσεις ή σφαίρες στις τρείς (σφαιρικές κατανομές) ή με ελλείψεις όταν πρόκειται για ελλειπτικές κατανομές. Προφανώς μπορούν να γενικευτούν και σε p διαστάσεις στον ευκλείδιο χώρο $\mathbb{R}^p$. Αρχικά μελετήθηκαν κυρίως απο τον Steerneman and van Perlo-ten Kleij (2005), τον Kelker (1970), τους Cambanis et al. (1981), επιπλέον τους Cacoulos and Koutras (1984), Koutras (1986), Koutras (1987) αλλά και τους Fang et al. (1990). Η μαθηματική τους μορφή χαρακτηρίζεται απο την εξής σ.π.π. :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \eta_p |\Sigma|^{-\frac{1}{2}} \cdot g((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})),$$

όπου η_p μια σταθερά (μπορεί να γραφεί και c_p), $g : \mathbb{R}^+ \to \mathbb{R}^+$ και $\boldsymbol{\mu}$ ο μέσος και $\boldsymbol{\Sigma}$ θετικά ορισμένος πίνακας (scale matrix). Η συνάρτηση g είναι πολύ σημαντική (θα χρησιμοποιηθεί και στις παρακάτω Ενότητες) καθώς είναι υπεύθυνη για την δημιουργία της κατανομής μιας και οι υπόλοιποι παράγοντες αποτελούν σταθερές. Στις παρακάτω ενότητες θα διερευνηθεί αρχικά η έννοια της χαρακτηριστικής συνάρτησης που είναι σημαντική για την ερμηνεία αυτών των κατανομών, ενώ στη συνέχεια θα αναλυθούν κυρίως θεωρητικά, ξεκινώντας απο τις σφαιρικές.

2.2 Χαρακτηριστική Συνάρτηση

Η χαρακτηριστική συνάρτηση αποτελεί αναπόσπαστο κομμάτι των συμμετρικών κατανομών. Μπορεί να οριστεί για οποιαδήποτε πραγματική τυχαία μεταβλητή. Αναντίρρητα είναι χρήσιμη κυρίως διότι ορίζει άν μία μεταβλητή κατανέμεται σφαιρικά ή ελλειπτικά, όπως θα δούμε

παρακάτω. Το όνομά της δεν είναι τυχαίο αφού χαρακτηρίζει πλήρως τη συνάρτηση κατανομής και τις ιδιότητές της. Μπορούμε π.χ. να εξάγουμε τις ροπές (άν υπάρχουν) απο αυτήν. Επιπλέον λειτουργεί και ως τρόπος ταυτοποίησης αν δύο κατανομές είναι ακριβώς ίδιες, διότι θα έχουν ίδιες χαρακτηριστικές συναρτήσεις. Συνήθως αυτό γίνεται όταν είναι δύσκολο να αποδειχθεί άμεσα η ισότητα $F_X(x) = F_Y(y)$. Η χαρακτηριστική συνάρτηση ορίζεται ως εξής :

$$\begin{aligned}\phi_X(t) &= \mathbb{E}(e^{itX}) \\ &= \mathbb{E}(\cos(tX) + i\sin(tX)) \\ &= \mathbb{E}(\cos(tX)) + i\mathbb{E}(\sin(tX)),\end{aligned}$$

αλλιώς μπορεί να γραφεί στην πιό τυπική μορφή για Χ συνεχή τυχαία μεταβλητή (τ.μ.) ως εξής:

$$\phi_X(t) = \int e^{itx} dF_X.$$

Στην περίπτωση συνεχούς τυχαίας μεταβλητής για την οποία υπάρχει η πυκνότητα πιθανότητας $f_X(x)$, τότε γράφεται ως :

$$\phi_X(t) = \int e^{itx} \cdot f_X(x)dx.$$

2.2.1 Βασικές Ιδιότητες

Οι σημαντικότερες ιδιότητες της χαρακτηριστικής συνάρτησης καταγράφονται παρακάτω

- $|\phi_X(t)| \leq 1$ και $\phi_X(0) = 1$.
- Αν $Y = aX + b$ τότε $\phi_Y(t) = e^{ibt}\phi_X(at)$.
- Αν X, Y είναι ανεξάρτητες τυχαίες μεταβλητές και $Z = X+Y$ τότε $\phi_Z(t) = \phi_X(t)\cdot\phi_Y(t)$.
- Εάν η ροπογεννήτρια $M_X(s) < \infty$ με $s \in [-\epsilon, \epsilon], \epsilon > 0$, τότε $\phi_X(t) = M_X(it)$, $\forall t \in \mathbb{R}$.

Ένα ισχυρό πλεονέκτημα της χαρακτηριστικής συνάρτησης έναντι της ροπογεννήτριας είναι οτι η πρώτη υπάρχει για κάθε πραγματική τυχαία μεταβλητή.

Τέλος, άλλη μια σημαντική ιδιότητα είναι η ένα-προς-ένα αντιστοιχία μεταξύ της χαρακτηριστικής συνάρτησης και της συνάρτησης κατανομής. Πιο αναλυτικά, μπορούμε να βρούμε την μοναδική χαρακτηριστική συνάρτηση μιας τυχαίας μεταβλητής X, την $\phi_X(t)$, γνωρίζοντας την κατανομή $F_X(x)$. Ισχύει βέβαια και το αντίστροφο, καθώς απο μία γνωστή $\phi_X(t)$ υπάρχει τρόπος εύρεσης της μοναδικής $F_X(x)$. Αυτό επιβεβαιώνει και την άρρηκτη σύνδεση μεταξύ των δύο συναρτήσεων, $\phi_X(t) \leftrightarrow F_X(x)$. Το επόμενο θεώρημα δίνει ένα τύπο με τον οποίο συνδέεται η πυκνότητα πιθανότητας με την χαρακτηριστική συνάρτηση.

Θεώρημα 2.2.1 (*Θεώρημα Αντιστροφής*)

Έστω X μια συνεχής τυχαία μεταβλητή, η οποία έχει πυκνότητα πιθανότητας $f_X(x)$ και χαρακτηριστική συνάρτηση $\phi_X(t) = \int e^{itx} f_X(x)dx$. Τότε η πυκνότητα πιθανότητας μπορεί να βρεθεί απο την χαρακτηριστική συνάρτηση ως :

$$f_X(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} e^{-itx} \phi_X(t) dt.$$

Σε οποιοδήποτε σημείο η $f_X(x)$ είναι διαφορίσιμη.

2.3 Σφαιρικότητα

Σε αυτή την ενότητα θα επισυγκεντρωθούμε στις κατανομές που διατηρούν σ.π.π. του τύπου $f_{\boldsymbol{X}}(\boldsymbol{x}) = c_p \cdot g(\boldsymbol{x}^T\boldsymbol{x})$ με $\mathbb{E}(\boldsymbol{X}) = 0$ και c_p μια σταθερά. Δηλαδή αναφερόμαστε σε κατανομές που επηρεάζονται αποκλειστικά απο την ευκλείδια νόρμα στο τετράγωνο $||\boldsymbol{x}||_2^2 = \boldsymbol{x}^T\boldsymbol{x}$ ή αλλιώς απο το τετραγωνικό μήκος του διανύσματος $\boldsymbol{x}$ γράφοντάς το ως $r^2 = \boldsymbol{x}^T\boldsymbol{x}$ (άλλες φορές ορίζουμε το $r = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ που είναι το μήκος του $\boldsymbol{x}$). Στη συνέχεια παρουσιάζεται μία ιδιότητα τέτοιων μεταβλητών.

Ορισμός 2.3.1 (*Σφαιρική Κατανομή*)

Έστω $\boldsymbol{X}$ ένα τυχαίο διάνυσμα p-διαστάσεων με $\mathbb{E}(\boldsymbol{X}) = 0$. Τότε το $\boldsymbol{X}$ είναι σφαιρικά κατανεμημένο, ή απλά σφαιρικό, αν και μόνο αν $\boldsymbol{X} \overset{d}{=} \boldsymbol{O}\boldsymbol{X}$ για κάθε ορθογώνιο πίνακα $\boldsymbol{O}_{p\times p}$, ($\boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I_p}$).

Δηλαδή αποτελούν μεταβλητές που δεν επηρεάζονται απο περιστροφές. Στη βιβλιογραφία εμφανίζονται με διάφορα ονόματα όπως ακτινικές (Kelker (1970)) ή ισοτροπικές (Bingham and Kiesel (2005)). Ο τρόπος με τον οποίο θα ταυτοποιείται μία μεταβλητή ως σφαιρική είναι ο εξής. Εάν υπάρχει συνάρτηση $\psi : [0, \infty) \to \mathbb{R}$ τέτοια ώστε :

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = \psi(||\boldsymbol{t}||_2^2),$$

η $\boldsymbol{X}$ είναι σφαιρική και θα χρησιμοποιείται ο συμβολισμός $\boldsymbol{X} \sim \mathcal{S}_p(g)$. Με $||\boldsymbol{t}||_2$ συμβολίζουμε την ευκλείδια νόρμα, δηλαδή $||\boldsymbol{t}||_2 = (\sum_p t_i^2)^{\frac{1}{2}}$ με $\boldsymbol{t} = (t_1, \cdots, t_p)^T$ και με $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ τη χαρακτηριστική συνάρτηση της τ.μ. $\boldsymbol{X}$. Μία καλύτερη απεικόνισή της θα ήταν χρησιμοποιώντας πολικές συντεταγμένες. Με αυτό τον τρόπο θα μπορέσουμε να διασπάσουμε το τυχαίο διάνυσμα σε ένα παράγοντα που αφορά το μήκος του διανύσματος (διότι όπως ειπώθηκε η τ.μ. επηρεάζεται απο το $||x||_2$) και σε ένα άλλο που αφορά την κατεύθυνσή του (γωνία).

Θεώρημα 2.3.1 (*Πολική Αναπαράσταση Σφαιρικής τυχαίας μεταβλητής*)

Έστω $\boldsymbol{X} \sim \mathcal{S}_p(g)$ με $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{0}$ και σ.π.π. $f_{\boldsymbol{X}}(\boldsymbol{x}) = c_p \cdot g(r^2)$ με $r^2 = \boldsymbol{x}^T\boldsymbol{x}$, c_p μία σταθερά και $g : \mathbb{R}^+ \to \mathbb{R}^+$. Επιπλέον ορίζουμε πολικές συντεταγμένες ως εξής :

$$\begin{aligned}
X_1 &= R \sin\Theta_1 \sin\Theta_2 \cdots \sin\Theta_{p-2} \sin\Theta_{p-1} \\
X_2 &= R \sin\Theta_1 \sin\Theta_2 \cdots \sin\Theta_{p-2} \cos\Theta_{p-1} \\
X_3 &= R \sin\Theta_1 \sin\Theta_2 \cdots \cos\Theta_{p-2} \\
&\vdots
\end{aligned}$$

$$\vdots$$
$$X_{p-1} = R\sin\Theta_1\sin\Theta_2$$
$$X_p = R\cos\Theta_1$$

με $R \in [0,+\infty)$, $\Theta_i \in [0,\pi]$, $i = \{1,\dots,p-2\}$, $\Theta_{p-1} \in (0,2\pi]$. *Τότε η* $\boldsymbol{X}$ *είναι Σφαιρική τ.μ. αν και μόνο αν τα* $R,\Theta_1,\dots,\Theta_{p-1}$ *είναι ανεξάρτητα και τα* $R = \sqrt{\boldsymbol{X}^T\boldsymbol{X}}$, Θ_i *και* Θ_{p-1} *έχουν σ.π.π. :*

$$f_R(r) = 2\frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2})}\cdot r^{p-1}\cdot g(r^2)$$
$$f_{\Theta_i}(\theta_i) = \frac{1}{B(\frac{1}{2},\frac{p-i}{2})}\cdot \sin^{p-i-1}\theta_i$$
$$f_{\Theta_{p-1}}(\theta_{p-1}) = \frac{1}{2\pi}.$$

Επιπρόσθετα ο πίνακας διακύμανσης-συνδιακύμανσης των Σφαιρικών κατανομών δίνεται απο τον τύπο $\mathrm{Cov}(\boldsymbol{X}) = \frac{\mathbb{E}(R^2)}{p}\boldsymbol{I}_p$ (Cambanis et al. (1981)). Ουσιαστικά με την πολική αναπαράσταση αναλύθηκε το τυχαίο διάνυσμα της σφαιρικής κατανομής σε δύο μέρη. Το πρώτο αφορά την κατεύθυνση του διανύσματος (με γωνίες $\theta_1,\dots,\theta_{p-1}$ ως προς τους άξονες των συντεταγμένων) και το δεύτερο κομμάτι αφορά το μήκος του διανύσματος ($||\boldsymbol{x}||_2 = \sqrt{\boldsymbol{x}^T\boldsymbol{x}} = r$). Το μήκος αυτό καθορίζει την κατανομή των δεδομένων σχετικά με το πόσο κοντά ή μακριά θα βρίσκονται τα δεδομένα απο την αρχή των αξόνων. Στην περίπτωση που τα σημεία έχουν μέσο διαφορετικό απο το $\boldsymbol{0} = (0,\dots,0)^T_{1\times p}$ τότε μπορούμε να τα κεντρικοποιήσουμε για να έχουν μέσο την αρχή των αξόνων, δηλαδή η απόσταση στο τετράγωνο να ορίζεται ως $R^2 = (\boldsymbol{X}-\boldsymbol{\mu})^T(\boldsymbol{X}-\boldsymbol{\mu})$. Αυτά τα συμπεράσματα μας οδηγούν άμεσα στον επόμενο ορισμό.

Θεώρημα 2.3.2 (*Στοχαστική Αναπαράσταση Σφαιρικής Κατανομής*)

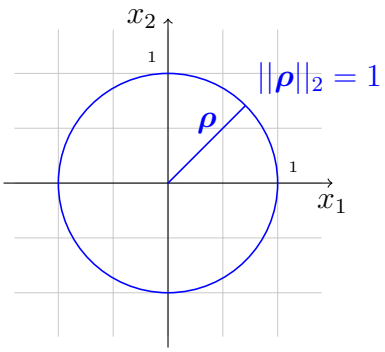
Έστω $\boldsymbol{S}$ *ομοιόμορφα κατανεμημένο στην μοναδιαία υπερσφαίρα* $\{\boldsymbol{x}\in\mathbb{R}^p : ||\boldsymbol{x}||_2 = 1\}$ p *διαστάσεων.*
Τότε $\boldsymbol{X}\sim\mathcal{S}_p(g)$ *αν και μόνο αν* $\boldsymbol{X}\overset{d}{=}R\boldsymbol{S}$ *όπου* $R\geq 0$ *με* $R\sim\sqrt{\boldsymbol{X}^T\boldsymbol{X}}$ *είναι τυχαία μεταβλητή ανεξάρτητη της* $\boldsymbol{S}$ *και* $\boldsymbol{S}\sim U(\{\boldsymbol{x}\in\mathbb{R}^p : ||\boldsymbol{x}||_2 = 1\})$ *δηλαδή σημεία* $\boldsymbol{x}$ *ομοιόμορφα κατανεμημένα* (U) *πάνω στην μοναδιαία υπερσφαίρα* ($\{\boldsymbol{x}\in\mathbb{R}^p : ||\boldsymbol{x}||_2 = 1\}$).

Η μονοδιάστατη τ.μ. R αποτελεί και το πιό σημαντικό κομμάτι απο την διάσπαση της Στοχαστικής Αναπαράστασης καθώς ο παράγοντας $\boldsymbol{S}$ είναι κοινός για κάθε σφαιρική κατανομή, δηλαδή δεν αλλάζει (δεδομένα πάνω στην υπερσφαίρα ακτίνας μήκους 1). Συγκεκριμένα το R είναι αυτό που καθορίζει την κατανομή της σφαιρικής τ.μ.. Εφόσον η κατανομή εξαρτάται απο την ποσότητα $R = \sqrt{\boldsymbol{X}^T\boldsymbol{X}} = ||\boldsymbol{X}||_2$ που αποτελεί την ευκλείδια νόρμα. Στα Σχήματα (2.1) και (2.2) φαίνεται η γεωμετρική απεικόνιση του παράγοντα $\boldsymbol{S}$ για $p=2$ καθώς και κάποιες προσομοιωμένες τιμές. Για την σ.π.π. του μήκους της ακτίνας θα επιλέξουμε την $f_R(r) = re^{-\frac{r^2}{2}}$ (δηλαδή $R\sim\sqrt{\chi^2_2}$).

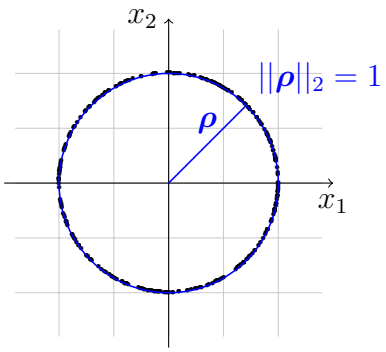
```
1 import numpy as np, matplotlib.pyplot as plt
2 from scipy.stats import chi2
```

```
3
4 # Παραγωγή Κανονικής Κατανομής μέσω Στοχαστικής Αναπαράστασης Σφαιρικών Μεταβλητών
5
6 deg = np.random.randint(0,360,300)
7 # S Κομμάτι, τυχαία σημεία επι του μοναδιαίου κύκλου
8 S = np.array([np.cos(deg),np.sin(deg)])
9 # R κομμάτι, προσομοίωση απο τετρ. ρίζα χ^2_2 κατανομής
10 chisq = np.random.chisquare(df = 2, size = 300)
11 R = np.sqrt(chisq)
12
13 #Μέσος όρος η αρχή των αξόνων
14 mu = np.repeat([0,0], 300).reshape(300,2)
15
16 # Σφαιρική Κανονική Κατανομή N(0, I_p) απο τον τύπο X = μ + RS με μ = 0
17 X = mu + np.multiply(R,S).T
18 plt.plot(X[:,0],X[:,1], "bo"); plt.show()
```

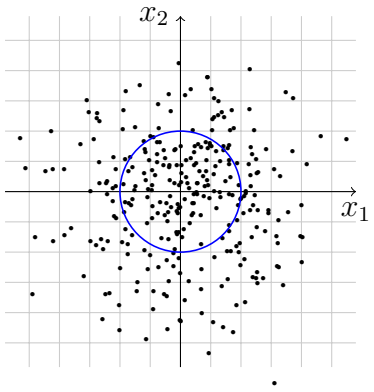
Προσομοίωση Σφαιρικών δεδομένων



Σχήμα 2.1: Ο μοναδιαίος κύκλος $\{\boldsymbol{x} \in \mathbb{R}^2 \ : \ ||\boldsymbol{x}||_2 = 1\}$



Σχήμα 2.2: Προσομοίωση τιμών στον μοναδιαίο κύκλο, μία αναπαράσταση της μεταβλητής $\boldsymbol{S}$



Σχήμα 2.3: Αποτέλεσμα γινομένου των δεδομένων στον κύκλο με τις τιμές της τ.μ. R, $\boldsymbol{X} = R\boldsymbol{S}$

2.4 Ελλειπτικότητα

Έχοντας παρουσιάσει την θεωρία των Σφαιρικών κατανομών, εύκολα μπορούμε να γενικεύσουμε το είδος των κατανομών σε Ελλειπτικού είδους. Ειδικά όσον αφορά την σ.π.π., θα αναφερόμαστε σε κατανομές με τον εξής τύπο :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \eta_p \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \cdot g((\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})) \tag{2.1}$$

Ορισμός 2.4.1 (*Ελλειπτική Κατανομή*)

Ένα τυχαίο διάνυσμα $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ p-διάστατο έχει ελλειπτική κατανομή με πίνακα $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$ και μέσο $\boldsymbol{\mu}_{p\times 1}$ εάν

$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Y}$$

όπου $\boldsymbol{Y} \sim \mathcal{S}_p(g)$, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$

Ξεκινώντας απο τον ορισμό της χαρακτηριστικής συνάρτησης, στην περίπτωση των Ελλειπτικών κατανομών θα έχουμε :

$$\begin{aligned}\phi_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = \mathbb{E}(e^{i\boldsymbol{t}^T(\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{Y})}) \\ &= e^{i\boldsymbol{t}^T\boldsymbol{\mu}} \cdot \mathbb{E}(e^{i(\boldsymbol{A}^T\boldsymbol{t})^T\boldsymbol{Y}}) = e^{i\boldsymbol{t}^T\boldsymbol{\mu}} \cdot \psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}).\end{aligned} \tag{2.2}$$

Επομένως η χαρακτηριστική συνάρτηση των ελλειπτικά συμμετρικών κατανομών μπορεί να γραφεί συναρτήσει της τετραγωνικής μορφής $\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$ που αυτό μας οδηγεί στο εξής θεώρημα.

Θεώρημα 2.4.1 (*Στοχαστική Αναπαράσταση Ελλειπτικής Κατανομής*)

Έστω $\boldsymbol{X}$ ενα p-διάστατο διάνυσμα. Το $\boldsymbol{X}$ ορίζεται ως Ελλειπτικά κατανεμημένο, ή απλά Ελλειπτικό, αν και μόνο αν υπάρχει διάνυσμα $\boldsymbol{\mu} \in \mathbb{R}^p$ και ενας θετικά ορισμένος πίνακας $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ τέτοια ώστε η χαρακτηριστική συνάρτηση του $\boldsymbol{X}-\boldsymbol{\mu}$ να δίνεται ως $\phi_{\boldsymbol{X}-\boldsymbol{\mu}}(\boldsymbol{t}) = \psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$, με μια συνάρτηση $\psi : \mathbb{R}_+ \to \mathbb{R}$. Το $\boldsymbol{X}$ θα ορίζεται πλέον ως :

$$\boldsymbol{X} \sim \mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$$

Πάλι βασισμένοι στην περίπτωση των Σφαιρικών κατανομών, για να προκύψει ο τύπος για τις Ελλειπτικές κατανομές, απλά θα χρησιμοποιήσουμε την στοχαστική αναπαράσταση των σφαιρικών μεταβλητών και θα κάνουμε μία τροποποίηση σε αυτά. Όπως ειπώθηκε, για την δημιουργία της έλλειψης αρκεί να μετασχηματίσουμε την αρχική σφαίρα με κάποιον πίνακα $\boldsymbol{A}$ που αποτελεί παράγοντα απο την διάσπαση Cholesky του πίνακα $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$, και να προσθέσουμε έναν διανυσματικό μέσο που θα αλλάξει απλά την θέση της κατανομής. Συμπερασματικά, ο τύπος ορίζεται ως :

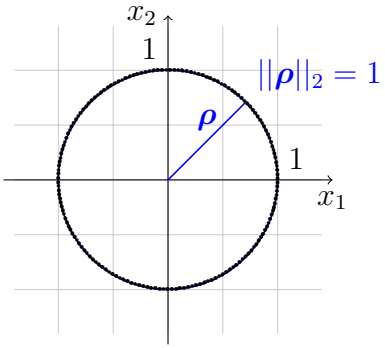
$$\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + R\boldsymbol{A}\boldsymbol{S}$$

Όπου τα $R, \boldsymbol{S}$ ορίζονται όπως στην Σφαιρική περίπτωση. Επιπλέον το $\mathrm{Cov}(\boldsymbol{X}) = \frac{\mathbb{E}(R^2)}{p}\boldsymbol{\Sigma}$. Ως εκ τούτου, η σφαίρα αλλάζει μορφή βάσει του πίνακα $\boldsymbol{A}$, και μετατοπίζεται απο το σημείο μηδέν

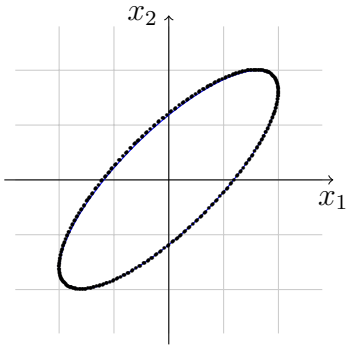
(αρχή αξόνων) στον διανυσματικό μέσο $\boldsymbol{\mu}$. Στη συνέχεια θα παρουσιαστεί μία προσομοίωση ενος δισδιάστατου ελλειπτικού Κανονικής Κατανομής χρησιμοποιώντας την Ελλειπτική θεωρία. Συνολικά παράγονται είκοσι δεδομένα ενώ τα στάδια δημιουργίας των Ελλειπτικών παρατηρήσεων αποδίδονται στα παρακάτω γραφήματα.

```
import numpy as np
import matplotlib.pyplot as plt
import scipy as sc
from scipy.stats import f
from scipy.stats import chi2

#Παραγωγή Κανονικής Κατανομής μέσω Στοχαστικής Αναπαράστασης Σφαιρικών Μεταβλητών
size = 300
deg = np.random.randint(0,360,size)
#S Κομμάτι, τυχαία σημεία επι του μοναδιαίου κύκλου
S = np.array([np.cos(deg),np.sin(deg)])
#R κομμάτι, προσομοίωση απο τετρ. ρίζα χ^2_2 κατανομής
chisq = np.random.chisquare(df = 2, size = size)
R = np.sqrt(chisq)

#Μέσος όρος η αρχή των αξόνων
mu = np.repeat([0,0], size).reshape(size,2)

#Το A = Σ^(1/2)
Sigma = np.array([[1,0.8],[0.8,1]])
A = np.linalg.cholesky(Sigma)

#Σφαιρική Κανονική Κατανομή N(0,I_d) απο τον τύπο X = RAS με μ = 0
X = mu + np.multiply(R,A@S).T
plt.plot(X[:,0],X[:,1], "bo"); plt.show()
```

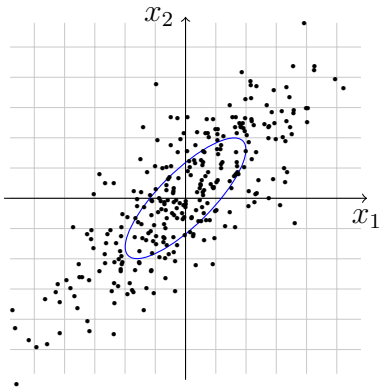
Προσομοίωση Ελλειπτικών δεδομένων απο δισδιάστατη Κανονική Κατανομή



Σχήμα 2.4: $\boldsymbol{X} = \boldsymbol{S}$. Διανύσματα προσομοιωμένα στον μοναδιαίο κύκλο αναπαριστώντας την μεταβλητή $\boldsymbol{S}$



Σχήμα 2.5: $\boldsymbol{X} = \boldsymbol{AS}$. ο $\boldsymbol{A}$ αλλάζει την κυκλική διάταξη των σημείων και τα μετατρέπει σε έλλειψη



Σχήμα 2.6: $\boldsymbol{X} = R\boldsymbol{A}\boldsymbol{S}$. Ο παράγοντας R αλλάζει τα μήκη των διανυσμάτων (σημεία) που βρίσκονται στην έλλειψη δημιουργώντας ένα νέφος σημείων

με $R = \sqrt{\chi^2_2}$ για παραγωγή τιμών απο $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ με $\boldsymbol{\mu} = (0,0)^T$ και $\boldsymbol{A}$ απο διάσπαση Cholesky $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$ με :

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0.520 & 0.854 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.52 \\ 0.52 & 1 \end{pmatrix}.$$

Το $\mathrm{Cov}(X, Y) = 0.52$ υποδηλώνει θετική συσχέτιση των X, Y γι' αυτό και παρατηρείται το ελλειπτικό με θετική κλίση. Στα Σχήματα (2.4), (2.5), (2.6) παρουσιάζεται η διαδικασία παραγωγής ελλειπτικών δισδιάστατων δεδομένων βήμα-βήμα. Στο Σχήμα (2.4) προσομοιώνουμε δεδομένα πάνω στον μοναδιαίο κυκλο, έχοντας έτσι τον πρώτο παράγοντα $\boldsymbol{S}$ της Στοχαστικής Αναπαράστασης. Στη συνέχεια τα δεδομένα αυτά πολλαπλασιάζονται με τον πίνακα $\boldsymbol{A}$ απο την διάσπαση Cholesky, δίνοντας έτσι την ελλειπτική μορφή των σημείων, όπως φαίνεται στο Σχήμα (2.5). Εν τέλει στη σχέση μπαίνει και ο παράγοντας R που αφορά την μή αρνητική μονοδιάστατη τ.μ. που θα δώσει διαφορετικά μήκη στα διανύσματα (δεδομένα), απο Σχήμα (2.6). Συγκεκριμένα αυτό που κάνει είναι να αλλάζει τα μήκη των διανυσμάτων, διαχέοντάς τα γύρω απο την έλλειψη δημιουργώντας έτσι την ελλειπτική κατανομή.

2.5 Σημαντικές Ιδιότητες των Ελλειπτικών Κατανομών

Μία ιδιότητα που θα φανεί πολύ χρήσιμη κυρίως στην Στατιστική είναι η εξής. Εάν το $\boldsymbol{X} \sim \mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ τότε η μετασχηματισμένη μεταβλητή $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{S}_p(g)$. Δηλαδή γενικά όταν μία κατανομή είναι Ελλειπτική τότε μπορούμε να την μετατρέψουμε μέσω μετασχηματισμού σε Σφαιρική. Π.χ. στην περίπτωση που ένας πληθυσμός ακολουθεί πολυδιάστατη Κανονική Κατανομή με συσχετισμένες μεταβλητές, χρησιμοποιώντας τον παραπάνω μετασχηματισμό μπορούμε να παράγουμε νέα μεταβλητή $\boldsymbol{Y}$ που θα αποτελείται πλέον απο ασυσχέτιστες συνιστώσες $Y_1, \ldots, Y_p$ διατηρώντας την Κανονικότητα των δεδομένων. Ουσιαστικά η Σφαιρικότητα σημαίνει πως οι συνιστώσες της πολυδιάστατης τυχαίας μεταβλητής είναι ασυσχέτιστες. Όσον αφορά τους γραμμικούς μετασχηματισμούς παρουσιάζονται κάποιες σημαντικές ιδιότητες. Για $\boldsymbol{X} \sim \mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ και $\boldsymbol{a}_{k\times 1}$, $\boldsymbol{B}_{k\times p}$ ισχύει πως $\boldsymbol{B}\boldsymbol{X} + \boldsymbol{a} = \mathcal{EC}_p(\boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^T, g)$. Επιπλέον για $\boldsymbol{\alpha} \in \mathbb{R}^p$ τότε $\boldsymbol{\alpha}^T\boldsymbol{X} \sim \mathcal{EC}_{p=1}(\boldsymbol{\alpha}^T\boldsymbol{\mu}, \boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha}, g)$. Για τις περιθώριες κατανομές διακρίνονται τα συγκεκριμένα χαρακτηριστικά. Στην τυπική περίπτωση $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ δηλαδή η $\boldsymbol{X}$ είναι Ελλειπτική κατανομή $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \sim \mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ όπου $\boldsymbol{X}_1 \sim$

$\mathcal{EC}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, g)$ και $\boldsymbol{X}_2 \sim \mathcal{EC}_{p-k}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}, g)$, δηλαδή και οι περιθώριες κατανομές των Ελλειπτικών μεταβλητών είναι Ελλειπτικές. Επιπλέον μπορεί να δειχθεί πως ακόμη και οι δεσμευμένες κατανομές των Ελλειπτικών κατανομών είναι και αυτές Ελλειπτικές.

Εν κατακλείδι, όπως παρουσιάστηκε προηγουμένως φάνηκε πως το σημαντικό κομμάτι είναι να βρεθεί η Κατανομή της R^2 η οποία δίνεται απο την τετραγωνική μορφή $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Γνωρίζοντας την κατανομή του R^2 μπορούμε εύκολα να προσομοιώσουμε ελλειπτικά δεδομένα απο αυτές, όπως έγινε παραπάνω. Συγκεκριμένα η μεταβλητή $\boldsymbol{X}$ κατασκευάζεται ως $\boldsymbol{X} = \sqrt{R^2}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{S}$ με $\boldsymbol{S}$ προσομοιωμένες τιμές πάνω στην μοναδιαία υπερσφαίρα και $\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{A}$ με $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$.

Κεφάλαιο 3

Οι Κυριότερες Σφαιρικές και Ελλειπτικές Κατανομές

Αυτό το Κεφάλαιο παρουσιάζει κάποιες σημαντικές πολυδιάστατες σφαιρικές και ελλειπτικές κατανομές, τις ιδιότητές τους καθώς και τον ρόλο τους στην Πολυμεταβλητή Στατιστική Ανάλυση.

3.1 Πολυδιάστατη Κανονική Κατανομή

Στη στατιστική ένας απο τους πιο σημαντικούς τομείς της αποτελεί η συμπερασματολογία, απο την οποία εξάγονται οι έλεγχοι των υποθέσεων, οδηγώντας έτσι στην λήψη αποφάσεων. Οι κατανομές αποτελούν την βάση των στατιστικών διαδικασιών διότι χωρίς αυτές δεν δύναται να κατανοήσουμε σε βάθος τα δεδομένα. Αναντίρρητα, απο τις πιό κλασικές και δημοφιλείς κατανομές αποτελεί η Κανονική. Για να γίνει αντιληπτή η χρησιμότητα αυτής της κατανομής, αρκεί να αναφέρουμε πως οι περισσότεροι έλεγχοι γίνονται βάσει της Κανονικής κατανομής (συνήθως με την χρήση του Κεντρικού Οριακού Θεωρήματος). Η πιό γνωστή μορφή της Κανονικής κατανομής είναι η μονοδιάστατη περίπτωση στην οποία ο μέσος και η διακύμανση αποτελούν αριθμούς. Επιπρόσθετα, προφανώς υπάρχουν και γενικεύσεις της σε περισσότερες διαστάσεις όπως η δισδιάστατη που πλέον μελετάται η απο κοινού συμπεριφορά δύο μεταβλητών σε μορφή διανύσματος στον ευκλείδιο χώρο $\mathbb{R}^2$. Εν τέλει, η πολυδιάστατη μορφή της αποτελεί την κατανομή, όπως ορίζει η πυκνότητά της, του διανύσματος τυχαίων μεταβλητών (κάθε συνιστώσα ως μεταβλητή) $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$. Άλλο ένα σημαντικό χαρακτηριστικό αυτών των κατανομών είναι οι περιθώριές τους. Στην περίπτωση της πολυδιάστατης Κανονικής, οι περιθώριές της είναι μονοδιάστατες Κανονικές η κάθε μία. Ωστόσο το αντίστροφο δεν ισχύει πάντα. Αναλυτικά, όταν οι περιθώριες είναι Κανονικές (δηλαδή κάθε μεταβλητή απο μόνη της χωρίς να λογαριαστεί η απο κοινού συμπεριφορά της με τις άλλες) τότε δεν είναι σίγουρο ότι και η απο κοινού συνάρτηση πυκνότητας πιθανότητας είναι η Κανονική. Επιστρέφοντας στο παρόν θέμα, η συγκεκριμένη κατανομή μπορεί να οριστεί μαθηματικώς όπως παρακάτω. Η πυκνότητα πιθανότητας της πολυδιάστατης Κανονικής κατανομής με μέσο $\boldsymbol{\mu}_{p\times 1}$ και πίνακα διακύμανσης-συνδιακύμανσης $\boldsymbol{\Sigma}_{p\times p}$ δίνεται απο τον τύπο :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{3.1}$$

Συνοπτικά μπορούμε να γράφουμε ότι τότε το $\boldsymbol{X}_{p\times 1} \sim N_p(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Sigma}_{p\times p})$. Προφανώς στην περίπτωση μίας μεταβλητής ο μέσος όρος και η διακύμανση αποτελούν αριθμούς και η σ.π.π. της $X \sim N(\mu, \sigma^2)$ παίρνει την μορφή :

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}.$$

Επιπλέον, σε δισδιάστατα δεδομένα, δηλαδή έχοντας δύο μεταβλητές X_1, X_2, το διάνυσμα $\boldsymbol{X} = (X_1, X_2)^T$, έχει μέσο όρο και διακύμανση αντιστοίχως :

$$\mathbb{E}(\boldsymbol{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{pmatrix} \qquad \begin{aligned} \mathrm{Cov}(\boldsymbol{X}) &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1,X_2} \\ \sigma_{X_2,X_1} & \sigma^2_{X_2} \end{pmatrix}. \end{aligned}$$

Η συνάρτηση πυκνότητας πιθανότητας έχει τον εξής τύπο :

$$f_{X_1,X_2}(x_1, x_2|\boldsymbol{\mu}_{2\times 1}, \boldsymbol{\Sigma}_{2\times 2}) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_{X_1})^2}{\sigma^2_{X_1}} - 2\rho\frac{x_1-\mu_{X_1}}{\sigma_{X_1}}\frac{x_2-\mu_{X_2}}{\sigma_{X_2}} + \frac{(x_2-\mu_{X_2})^2}{\sigma^2_{X_2}}\right]}.$$

Το $\rho = \frac{\sigma_{X_1,X_2}}{\sigma_{X_1}\sigma_{X_2}}$ αποτελεί τον συντελεστή συσχέτισης των δύο μεταβλητών. Ουσιαστικά ο παραπάνω τύπος είναι ανοιγμένη μορφή της γενικής συνάρτησης (3.1) (για δισδιάστατα δεδομένα). Επιστρέφοντας στον τύπο της $p-$διάστατης Κανονικής κατανομής (3.1), παρατηρείται πως η πυκνότητα της $\boldsymbol{X}$ εξαρτάται αποκλειστικά απο την ποσότητα $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ (τετραγωνική απόσταση Mahalanobis). Αυτό μας δίνει την δυνατότητα να αφαιρέσουμε την ύπαρξη συσχέτισης απο τα δεδομένα και να παραχθούν νέες μετασχηματισμένες μεταβλητές οι οποίες πλέον θα είναι ασυσχέτιστες.

Όσον αφορά την Στοχαστική Αναπαράσταση της πολυδιάστατης (σφαιρικής) Κανονικής κατανομής ($\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$), δίνεται ως $\boldsymbol{X} = R\boldsymbol{S}$, με $\boldsymbol{S}$ τυχαία μεταβλητή στην μοναδιαία υπερσφαίρα (πιο κατανοητά επάνω στον κύκλο για δισδιάστατες μεταβλητές) και R μή αρνητική τυχαία μεταβλητή ανεξάρτητη της $\boldsymbol{S}$ (βλέπε Θεώρημα 2.3.1). Ουσιαστικά το R όπως θα δούμε καθορίζει την κατανομή των δεδομένων ενώ το $\boldsymbol{S}$ αφορά την σφαιρική δομή τους. Ένα εύλογο ερώτημα είναι πως θα παράγουμε σφαιρικά δεδομένα απο μία πολυδιάστατη Κανονική Κατανομή. Το ζητούμενο συνεπώς είναι πώς θα καθοριστεί το R. Αυτό θα γίνει ακολουθώντας την εξής παρακάτω διαδικασία για να βρούμε την κατανομή του R.
Έστω $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Εφόσον το $\boldsymbol{X}$ γνωρίζουμε πως είναι σφαιρικό (απο την χαρακτηριστική του συνάρτηση που είναι συνάρτηση του $||\boldsymbol{t}||_2$), τότε το $\boldsymbol{X} = R\boldsymbol{S}$. Άρα ισχύει η ισότητα

$$\boldsymbol{X}^T\boldsymbol{X} = (R\boldsymbol{S})^T R\boldsymbol{S} = R\boldsymbol{S}^T\boldsymbol{S}R.$$

Επειδή όμως οι συνιστώσες $\boldsymbol{S} = (S_1, S_2, \ldots, S_p)^T$ είναι ασυσχέτιστες με $||\boldsymbol{S}||_2 = 1$, ισχύει :

$$R\boldsymbol{S}^T\boldsymbol{S}R = R^2.$$

Η $\boldsymbol{X}^T\boldsymbol{X} = \chi^2_p$ είναι δηλαδή χ^2 κατανομή (διότι $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$) άρα :

$$R^2 = \chi^2_p,$$

καταλήγοντας πως το $R = \sqrt{\chi^2_p}$. Στην περίπτωση της πολυδιάστατης ελλειπτικής Κανονικής κατανομής, η μόνη διαφορά απο την περίπτωση της σφαιρικής είναι πως θα αντικαταστήσουμε το $\boldsymbol{S}$ με $\boldsymbol{AS}$ στον τύπο της Στοχαστικής αναπαράστασης, παίρνοντας έτσι την μορφή $\boldsymbol{X} = R\boldsymbol{AS}$ με $R^2 \sim \chi^2_p$ (ή αλλιώς $R \sim \sqrt{\chi^2_p}$).

Θεώρημα 3.1.1 (*Μετασχηματισμός Κανονικής Κατανομής*)

Έστω $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *όπου* $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$. *Τότε*

$$\boldsymbol{Y} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$$

όπου τα $Y_i \; \forall\, i \in \{1, \ldots, p\}$ *είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την μονοδιάστατη Κανονική κατανομή* $N(0, 1)$.

Αυτό αποδεικνύεται αρκετά σημαντική ιδιότητα καθώς μπορούμε να μετασχηματίσουμε τις αρχικές μεταβλητές $X_1, X_2, \ldots, X_p$ σε $Y_1, Y_2, \ldots, Y_p$, όπου πλέον είναι ανεξάρτητες, δηλαδή $Y_i \sim N(0, 1)$. Υπάρχει ένας πολύ σημαντικός λόγος που είναι επιθυμητός ένας τέτοιος μετασχηματισμός. Όπως προϋπόθηκε, ένα απο τα βασικά στοιχεία της Στατιστικής είναι η συμπερασματολογία η οποία βασίζεται αρκετά στις πιθανότητες καθώς και στα Διαστήματα Εμπιστοσύνης. Στην περίπτωση μίας διάστασης τα πράγματα είναι απλά, αρκεί να βρούμε την κατανομή της μεταβλητής και απο κεί και πέρα μπορούν να εξαχθούν p_values κτλ. Ωστόσο στην πολυδιάστατη περίπτωση η μεταβλητή αποτελεί ένα πολύπλοκο διάνυσμα με συνιστώσες μεταβλητές που συνήθως συσχετίζονται μεταξύ τους. Αυτό και μόνο καθιστά τη συμπερασματολογία αρκετά δύσκολη. Συνεπώς είναι αναγκαία η ύπαρξη μετασχηματισμού ο οποίος θα εξαλείψει τις συσχετίσεις των μεταβλητών. Για την καλύτερη κατανόηση, παρακάτω παρουσιάζεται γραφικά η μετατροπή που είδαμε πριν.

$$N\left[\begin{pmatrix}\mu_1\\ \vdots \\ \mu_p\end{pmatrix}, \begin{pmatrix}\mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p)\\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_p)\\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}(X_p)\end{pmatrix}\right]$$

$$\Big\downarrow \boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$$

$$N\left[\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \begin{pmatrix}1 & 0 & \ldots & 0\\ 0 & 1 & \ldots & 0\\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1\end{pmatrix}\right].$$

Μπορεί να διαπιστώσει κανείς πως η εξίσωση $(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) = c^2$, $c \in \mathbb{R}$ ορίζει ελλειψοειδές, γι'αυτό το λόγο και οι συναρτήσεις πολυδιάστατων Κανονικών κατανομών δημιουργούν ελλείψεις. Είναι γνωστό απο το προηγούμενο θεώρημα πως το $\boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{X} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{I_p})$ δηλαδή είναι ανεξάρτητες τυχαίες μεταβλητές. Απο αυτό συμπεραίνουμε ότι :

$$
\begin{gathered}
(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = \\
[\boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{x} - \boldsymbol{\mu})]^T [\boldsymbol{\Sigma}^{-\frac{1}{2}} (\boldsymbol{x} - \boldsymbol{\mu})] = \\
\boldsymbol{y}^T \boldsymbol{y} = \\
\sum_{i=1}^{p} y_i^2
\end{gathered}
$$

εφόσον τα $Y_i \sim N(0,1)$, τότε το $\boldsymbol{y}^T \boldsymbol{y} = \sum_{i=1}^{p} y_i^2 \sim \chi_p^2$, επειδή η κατανομή χ^2 προκύπτει ως άθροισμα τετραγώνων ανεξάρτητων μονοδιάστατων Κανονικών κατανομών. Το τελευταίο αυτό συμπέρασμα είναι αρκετά σημαντικό καθώς μας δίνει την δυνατότητα κατασκευής ελλειπτικών περιοχών πρόβλεψης. Αναλυτικά έχοντας επίπεδο σημαντικότητας α θα ισχύει η ισότητα :

$$P((\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi_{p,\alpha}^2) = 1 - \alpha,$$

οπότε η ανισότητα που υπάρχει εντός της πιθανότητας για οποιοδήποτε α, μας δίνει περιοχές εμπιστοσύνης με συντελεστή $1 - \alpha$.

3.1.1 Γεωμετρική Ερμηνεία

Γενικώς όλες οι κατανομές του τύπου $\boldsymbol{f_Y}(\boldsymbol{y}) = \eta_p |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot g[(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})]$, με η_p μία σταθερά έχουν ελλειπτική γεωμετρική αναπαράσταση. Προφανώς αυτό ισχύει και για τον τύπο της Κανονικής πολυδιάστατης κατανομής, όπως αυτή έχει οριστεί στο (3.1). Ευδιάκριτη είναι η αντιστοιχία $\eta_p \to (2\pi)^{-\frac{p}{2}}$, $g(t) \to e^{-\frac{1}{2}\boldsymbol{t}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{t}}$. Συγκεκριμένα όλα τα $\boldsymbol{X}$ για τα οποία ισχύει η ισότητα που ακολουθεί :

$$\varepsilon(\boldsymbol{x}) = \boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2$$

ορίζουν μία έλλειψη στον ευκλείδιο χώρο $\mathbb{R}^p$ Βέβαια, το γιατί ορίζεται η μορφή της τυχαίας μεταβλητής $\boldsymbol{X}$ ως έλλειψη μπορεί να αποδειχθεί χρησιμοποιώντας την φασματική ανάλυση του πίνακα $\boldsymbol{\Sigma}$. Εφόσον ο πίνακας $\boldsymbol{\Sigma}$, είναι συμμετρικός θα διαγωνοποείται και μπορεί να γραφεί σε μορφή $\boldsymbol{\Sigma} = \boldsymbol{S}\,\boldsymbol{\Lambda}\,\boldsymbol{S}^{-1} = \boldsymbol{S}\,\boldsymbol{\Lambda}\,\boldsymbol{S}^T$ όπου ο $\boldsymbol{S}$ είναι ο ορθογώνιος πίνακας ιδιοδιανυσμάτων. Στη συνέχεια, λόγω αυτής της διάσπασης, μπορούμε να γράψουμε :

$$
\begin{gathered}
(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{S}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{S}^T (\boldsymbol{x} - \boldsymbol{\mu}) = \\
[\boldsymbol{S}^T (\boldsymbol{x} - \boldsymbol{\mu})]^T \boldsymbol{\Lambda}^{-1} [\boldsymbol{S}^T (\boldsymbol{x} - \boldsymbol{\mu})] \overset{\boldsymbol{y} = \boldsymbol{S}^T (\boldsymbol{x} - \boldsymbol{\mu})}{=} \boldsymbol{y}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{y} = \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}
\end{gathered}
\tag{3.2}
$$

Επομένως η ισότητα $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c^2$ μπορεί να γραφεί ισοδύναμα ως :

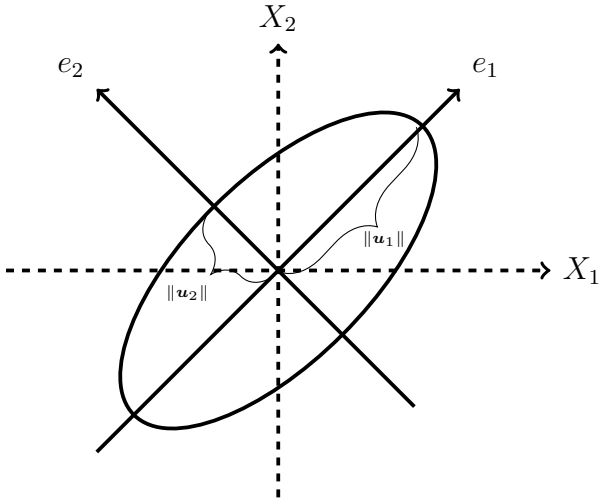
$$\frac{y_1^2}{(\sqrt{\lambda_1})^2} + \frac{y_2^2}{(\sqrt{\lambda_2})^2} + \cdots + \frac{y_p^2}{(\sqrt{\lambda_p})^2} = c^2.$$

Η τελευταία εξίσωση περιγράφει μία έλλειψη. Το c^2 είναι μια σταθερά που καθορίζει το μέγεθος του ελλειψοειδούς. Οι άξονες του ελλειπτικού δίνονται απο τα ιδιοδιανύσματα, ενώ τα μήκη του κάθε ιδιοδιανύσματος εξαρτώνται απο τις αντίστοιχες ιδιοτιμές τους. Η κύρια διεύθυνση της έλλειψης ορίζεται απο το ιδιοδιάνυσμα με την μεγαλύτερη ιδιοτιμή, ενω μικρότερες ιδιοτιμές καθορίζουν τα μικρότερα μήκη πάνω στους άξονες.

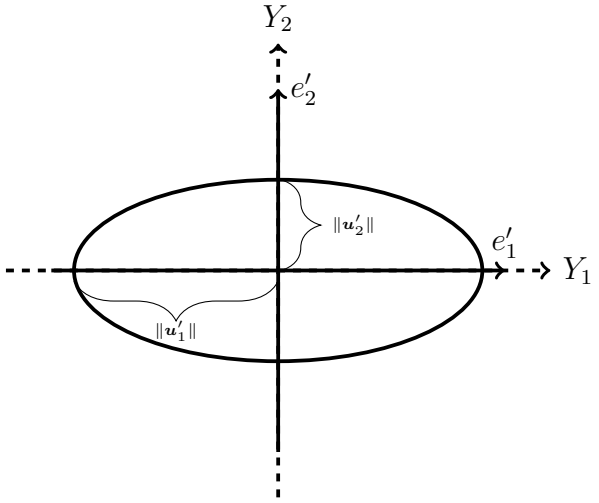
Συνδέοντάς το και με την Στατιστική, η σταθερά c^2 αποτελεί ουσιαστικά το ποσοστιαίο σημείο της $\chi^2_{p,a}$, γι αυτό το λόγο μπορούμε να δημιουργήσουμε απο κοινού περιοχές πρόβλεψης των παρατηρήσεων. Π.χ. αν θελήσουμε να έχουμε 40% κάλυψη σε δισδιάστατα δεδομένα, τότε το ποσοστιαίο πρέπει να ισούται με $\chi^2_{p=2,a=0.6}$ που είναι περίπου 1. Παρακάτω θα ακολουθήσει ένα παράδειγμα απεικόνισης δισδιάστατων ελλειπτικών.

$$\frac{y_1^2}{(\sqrt{\lambda_1})^2} + \frac{y_2^2}{(\sqrt{\lambda_2})^2} = \chi^2_{p=2,a=0.6} \simeq 1.$$

Άλλες πιο συνήθεις περιπτώσεις περιλαμβάνουν σταθερές $\chi^2_{p=2,a=0.1} \simeq 4.6057, \chi^2_{p=2,a=0.05} \simeq 5.9914, \chi^2_{p=2,a=0.01} \simeq 9.21034$.



Σχήμα 3.1: Έλλειψη περιοχής εμπιστοσύνης 1-α των μεταβλητών $\boldsymbol{X}$



Σχήμα 3.2: Έλλειψη περιοχής εμπιστοσύνης 1-α των μετασχηματισμένων μεταβλητών $\boldsymbol{Y}$

$$\|\boldsymbol{u}_1\|_2 = \sqrt{\lambda_1 \cdot \chi^2_{p=2,a}}$$
$$\|\boldsymbol{u}_2\|_2 = \sqrt{\lambda_2 \cdot \chi^2_{p=2,a}}$$

$$\|\boldsymbol{u}'_1\|_2 = \sqrt{\lambda_1 \cdot \chi^2_{p=2,a}}$$
$$\|\boldsymbol{u}'_2\|_2 = \sqrt{\lambda_2 \cdot \chi^2_{p=2,a}}$$

Ο άξονας e_1 αποτελεί την προέκταση του ιδιοδιανύσματος u_1 με το μεγαλύτερο μηκος $\|u_1\|_2$ ενώ το e_2 είναι ο άξονας του μικρότερου ιδιοδιανύσματος (u_2 με μήκος $\|u_2\|_2 < \|u_1\|_2$) απο την φασματική ανάλυση των μεταβλητών X_1, X_2. Απο την άλλη το Σχήμα (3.2) αποτελεί περιστροφή του ελλειπτικού απο το Σχήμα (3.1) όπου πλέον οι άξονες και τα διανύσματα έχουν διαφορετική διεύθυνση συμβολίζοντάς τα με $u'_i, e'_i, i = \{1, 2\}$. Αυτή η περιστροφή μας οδηγεί σε ενα νέο σύστημα μεταβλητών Y_1, Y_2 που προκύπτει απο το αρχικό X_1, X_2. Επιπρόσθετα, τα μήκη των ιδιοδιανυσμάτων είναι $\sqrt{\lambda_i \cdot c^2}$, ωστόσο απο τον τύπο (3.2) είναι γνωστό πως το c^2 αποτελεί το ποσοτιαίο της $\chi^2_{p,a}$ κατανομής, γι αυτό και εν τέλει το μήκος είναι $\sqrt{\lambda_i \cdot \chi^2_{p,a}}$. Προ-

φανώς, μεγάλα ποσοστιαία οδηγούν σε μεγαλύτερα ελλειψοειδή. Απο την άλλη, εάν παρατηρήσουμε την έλλειψη των μετασχηματισμένων μεταβλητών $\boldsymbol{Y}$, επιβεβαιώνεται πως οι νέες αυτές μεταβλητές είναι ασυσχέτιστες μεταξύ τους, αφού τα ιδιοδιανύσματα ανήκουν στους ίδιους τους άξονες των συντεταγμένων. Σχετικά με την περιστρόφή του ελλειπτικού απο το Σχήμα 3.1, μιας και αναφερόμαστε σε δύο διαστάσεις, μπορεί εύκολα να υπολογιστεί η γωνία περιστροφής. Αυτή προκύπτει απο την σχέση $\boldsymbol{X} = \boldsymbol{S} \cdot \boldsymbol{Y}$ με $\boldsymbol{Y}$ μεταβλητή που προέκυψε απο την φασματική ανάλυση του πίνακα $\boldsymbol{\Sigma}$. Ο $\boldsymbol{S}$ δημιουργεί την περιστροφή δεξιόστροφα με μορφή :

$$\boldsymbol{S} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix},$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}.$$

Η περιστροφή αριστερόστροφα επιτυγχάνεται απο την σχέση $\boldsymbol{Y} = \boldsymbol{S}^T \cdot \boldsymbol{X}$. Ένας τρόπος να βρούμε την γωνία περιστροφής είναι να επισυγκεντρωθούμε στην γωνία μεταξύ του ιδιοδιανύσματος στο 1ο τεταρτημόριο και στον άξονα x. Απο την τριγωνομετρία είναι γνωστό οτι $< \boldsymbol{x}, \boldsymbol{y} >= cos\theta \cdot \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{y}\|_2$, οπότε η γωνία ισούται :

$$cos\theta = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}. \tag{3.3}$$

Αυτά τα δεδομένα ειναι κεντρικοποιημένα, δηλαδή $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{0}, \mathbb{E}(\boldsymbol{Y}) = \boldsymbol{0}$. Η τελευταία σχέση δίνει :

$$cos\theta = \frac{\Big(x_1, x_2, \ldots, x_p\Big) \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}}{\sqrt{< \boldsymbol{x}, \boldsymbol{x} > \cdot < \boldsymbol{y}, \boldsymbol{y} >}} =$$

$$\frac{\sum_{i=1}^{p} x_i \cdot y_i}{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{p} y_i^2} = \frac{\sigma_{\boldsymbol{XY}}}{\sqrt{\sigma_{\boldsymbol{X}}^2 \sigma_{\boldsymbol{Y}}^2}} = \rho_{\boldsymbol{XY}}.$$

Συνεπώς η γωνία $\theta = \text{τοξσυν}\Big(\rho_{\boldsymbol{XY}}\Big)$.

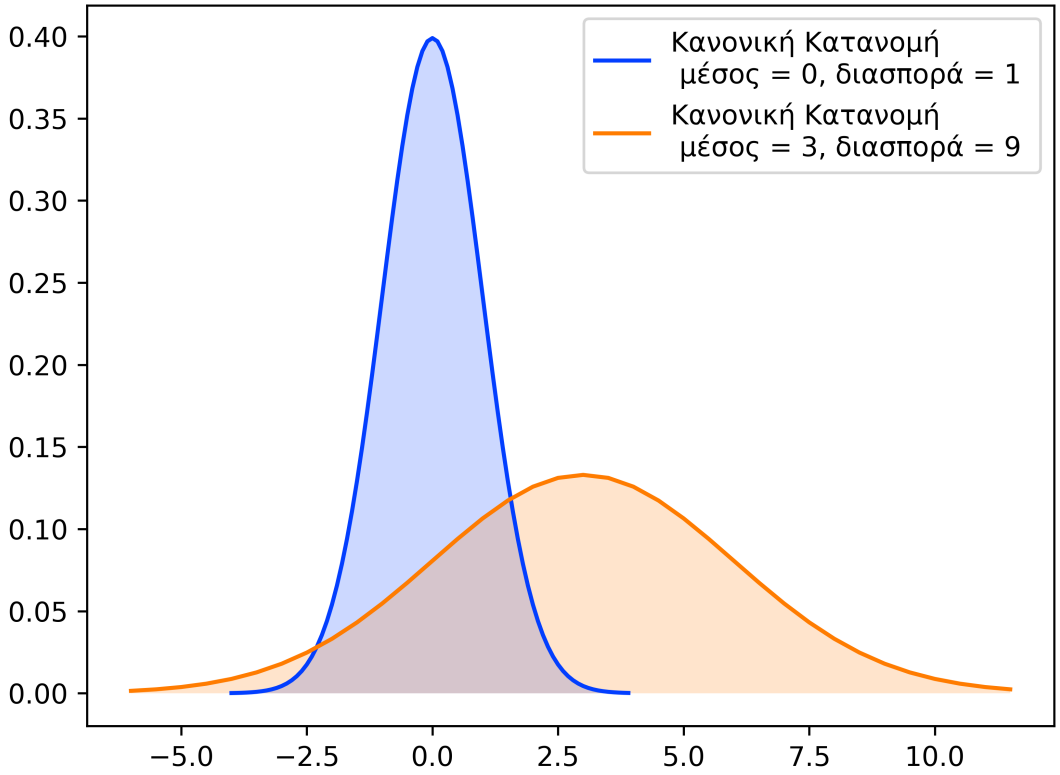
3.1.2 Γραφική Απεικόνιση

Σε αυτή την ενότητα θα παρουσιαστούν κάποιες ιδιότητες των πολυδιάστατων σφαιρικών και ελλειπτικών Κανονικών κατανομών. Θα δοθεί έμφαση στην οπτικοποίησή τους καθώς και στις παραμέτρους που τις διέπουν. Το λογισμικό που θα χρησιμοποιηθεί είναι η **PYTHON**,

Έκδοση 3.9.5. Ιδιαίτερη βαρύτητα θα δοθεί σε προσομοιωμένα δεδομένα δύο και τριών διαστάσεων για να δωθεί η δυνατότητα γεωμετρικής αποτύπωσής τους. Θα ξεκινήσουμε με την αναπαράσταση της Κανονικής κατανομής, πρώτα σε μία διάσταση και μετά σε περισσότερες.

Αλγόριθμος 3.1.1: Μονοδιάστατες Κανονικές Κατανομές

```
'''Εισαγωγή βασικών πακέτων (γραμμές 2-11)'''
import math
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
from scipy.stats import multiVariate_normal
from scipy.stats import norm
import pandas as pd
from matplotlib import cm
from matplotlib.ticker import LinearLocator
from mpl_toolkits.mplot3d import axes3d

'''Παραγωγή τιμων x1, x2 στα οποία θα στηριχθούν η πυκνότητες norm1,
 ↪  norm2 '''
x1 = np.arange(-4,4,0.1)
x2 = np.arange(-6,12,0.5)
norm1 = norm(0, 1)
norm2 = norm(3,3)
y1 = norm1.pdf(x1)
y2 = norm2.pdf(x2)
arr1 = np.vstack([x1,y1]).T
arr2 = np.vstack([x2,y2]).T

'''Γράφημα '''
df1 = pd.DataFrame(arr1, columns=['x1','y1'])
df2 = pd.DataFrame(arr2, columns=['x2','y2'])
sns.set_palette("bright")
lp1 = sns.lineplot(data= df1,x = x1,y = y1,label = ('Κανονική
 ↪  Κατανομή\n μέσος = 0, διασπορά = 1)
lp2 = sns.lineplot(data= df2, x = x2, y = y2, label ='Κανονική
 ↪  Κατανομή\n μέσος = 3, διασπορά = 9')
lp1.fill_between(x1, y1, interpolate=True, alpha =0.2)
lp2.fill_between(x2, y2, interpolate=True, alpha =0.2)
plt.show()
```



Σχήμα 3.3: Διάγραμμα Κανονικών μονοδιάστατων μεταβλητών

Ως γνωστόν, η Κανονική κατανομή χαρακτηρίζεται απο 2 παραμέτρους :

- Τον μέσο που καθορίζει την θέση της πυκνότητας. Φαίνεται πως η μπλέ πυκνότητα συγκεντρώνεται σε χαμηλότερες τιμές απο την πορτοκαλί λόγω μικρότερου μέσου.
- Την διακύμανση που επηρεάζει το εύρος των τιμών. Η μπλέ κατανομή, εξ'αιτίας της μικρότερης διασποράς, απλώνεται λιγότερο εκατέρωθεν του μέσου απ'ότι η πορτοκαλί με μεγαλύτερη διακύμανση.

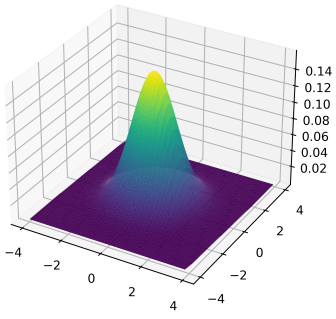
Ας προχωρήσουμε πλέον σε πιο πολύπλοκες περιπτώσεις. Στην συνέχεια θα παρουσιαστεί η δομή της δισδιάστατης Κανονικής κατανομής καθώς και οι ιδιότητές της. Όπως έχει γίνει γνωστό, η μεταβλητή αποτελεί ένα διάνυσμα τυχαίων μεταβλητών οι οποίες συνθέτουν την απο κοινού συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της Κανονικής κατανομής στις δύο διαστάσεις. Γι'αυτό το λόγο και οι παράμετροι ορίζονται στον ίδιο χώρο, αποτελώντας διανύσματα ή πίνακες. Συγκεκριμένα θα παρουσιαστούν περιπτώσεις με μέσο $(0,0)^T$ και με πίνακες διακυμάνσεων-συνδιακυμάνσεων :

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

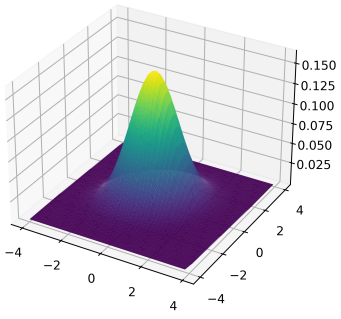
Ο κώδικας **PYTHON** παρατίθεται παρακάτω.

Αλγόριθμος 3.1.2: Δισδιάστατη Απεικόνιση Κανονικών Κατανομών

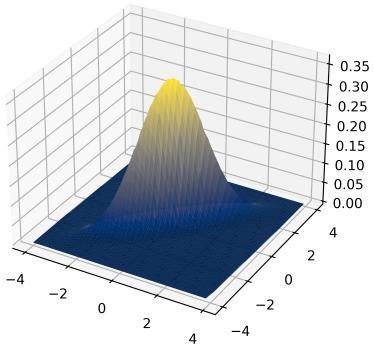
```
'''Μέσω εντολής multiVariate_normal ορίζονται οι πυκνότητες 2 διαστάσεων'''
norm_ind = multiVariate_normal([0,0],[[1,0],[0,1]])
norm_02 = multiVariate_normal([0,0],[[1,0.2],[0.2,1]])
norm_09 = multiVariate_normal([0,0],[[1,0.9],[0.9,1]])
seq1 = np.repeat(np.arange(-4,4,0.05),50)
seq2 = np.asarray(list(np.arange(-4,4,0.05))*50)
xy = np.vstack([seq1,seq2]).T
dns_i = norm_ind.pdf(xy)
dns_02 = norm_02.pdf(xy)
dns_09 = norm_09.pdf(xy)

'''Γράφημα '''
fig = plt.figure()
ax = fig.gca(projection='3d')
ax.plot_trisurf(seq1, seq2, dns_i, cmap=plt.cm.viridis, linewidth=0.2)
plt.show()

fig = plt.figure()
ax = fig.gca(projection='3d')
ax.plot_trisurf(seq1, seq2, dns_02, cmap=plt.cm.viridis, linewidth=0.2)
plt.show()

fig = plt.figure()
ax = fig.gca(projection='3d')
ax.plot_trisurf(seq1, seq2, dns_09, cmap=plt.cm.cividis, linewidth=1)
plt.show()
```



Σχήμα 3.4: Κανονική κατανομή δύο διαστάσεων με $\mathrm{Cov}(X,Y)=0$



Σχήμα 3.5: Κανονική κατανομή δύο διαστάσεων με $\mathrm{Cov}(X,Y)=0.2$



Σχήμα 3.6: Κανονική κατανομή δύο διαστάσεων με $\mathrm{Cov}(X, Y) = 0.9$

Απο τα Σχήματα (3.4), (3.5) και (3.6) των δισδιάστατων Κανονικών πυκνοτήτων διακρίνονται τα εξής :

- Η θέση της κατανομής ορίζεται απο το διάνυσμα του μέσου $\mathbb{E}(\boldsymbol{X}) = (\mathbb{E}(X_1), \mathbb{E}(X_2))$. Όλες οι κατανομές έχουν κέντρο το $\boldsymbol{O} = (0, 0)^T$. Η αλλαγή του μέσου αλλάζει μόνο την θέση της πυκνότητας.

- Το μέγεθος της σ.π.π. διαμορφώνεται απο τον πίνακα διακύμανσης-συνδιακύμανσης και συγκεκριμένα απο την διαγώνιό του. Αν επιθυμούσαμε μεγαλύτερη διασπορά ως προς τον άξονα x, τότε το στοιχείο σ_{11} έπρεπε να είναι μεγαλύτερο. Το ίδιο ισχύει και για τον άξονα y με το σ_{22}. Τα μή διαγώνια στοιχεία ορίζουν την συσχέτιση μεταξύ των μεταβλητών. Π.χ. στο Σχήμα (3.5) η συνδιακύμανση είναι $\mathrm{Cov}(X_1, X_2) = 0.2$, που σημαίνει πως υπάρχει θετική ασθενής συσχέτιση μεταξύ τους. Το ελλειψοειδές βέβαια που δημιουργείται απέχει λίγο απο την σφαιρικότητα. Ωστόσο στην περίπτωση του Σχήματος (3.6) η συσχέτιση είναι αρκετά υψηλή, γι'αυτό και η κατανομή είναι επιμηκυμένη έχοντας θετική κλίση λόγω του θετικού συντελεστή συσχέτισης σ_{12} ή σ_{21}.

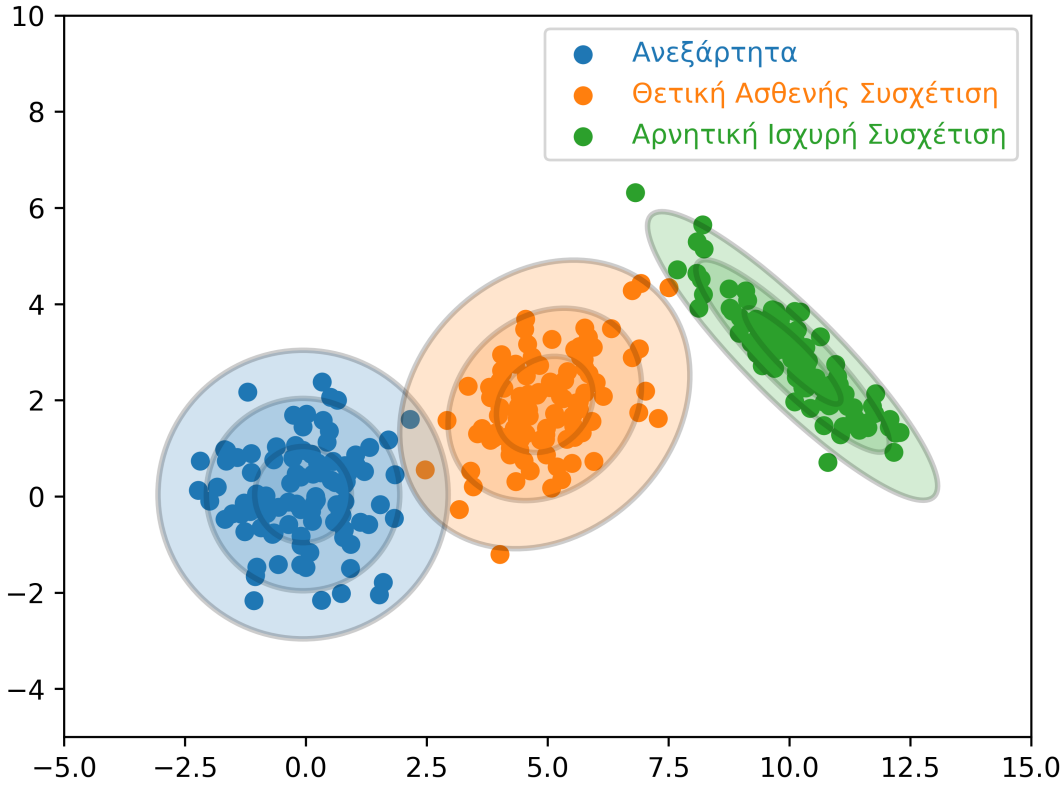
Όπως αναφέρθηκε στην υποενότητα Γεωμετρική Ερμηνεία, η Κανονική κατανομή δημιουργεί ελλειπτικές δομές όταν υπάρχει συσχέτιση στις μεταβλητές, αλλιώς τα δεδομένα κατανέμονται σφαιρικά. Αυτά τα ελλειψοειδή μας βοηθούν στη δημιουργία περιοχών πρόβλεψης αλλά και περιοχών εμπιστοσύνης (Confidence Regions) όπου πλέον η στατιστική συμπερασματολογία αφορά όλες τις μεταβλητές ταυτόχρονα και όχι μόνο την καθεμία ξεχωριστά. Για την καλύτερη κατανόηση της θεωρίας θα επισυγκεντρωθούμε σε προσομοίωση δισδιάστατων δεδομένων απο διαφόρου είδους Κανονικές κατανομές με σκοπό την γραφική παρουσίαση των ελλειπτικών περιοχών πρόβλεψης των παρατηρήσεων.

Αλγόριθμος 3.1.3: Δισδιάστατη Απεικόνιση Κανονικών Ελλειπτικών Κατανομών

```
1 import numpy as np
2 from scipy.stats import multiVariate_normal
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
```

```
cov_ind = [[1,0],[0,1]]
pmf_ind = multiVariate_normal([0,0],[[1,0],[0,1]])
x_ind = pmf_ind.rvs(size=100)
cov_2 = [[1,0.2],[0.2,1]]
pmf_2 = multiVariate_normal([5,2],[[1,0.6],[0.6,1]])
x_2 = pmf_2.rvs(size=100)
cov_9 = [[1,-0.9],[-0.9,1]]
pmf_9 = multiVariate_normal([10,3],[[1,-0.9],[-0.9,1]])
x_9 = pmf_9.rvs(size=100)

df = np.concatenate([x_ind,x_2,x_9],axis=1)
cov = [cov_ind,cov_2,cov_9]
costheta = rotation_rad = rotation_deg = []

fig, ax = plt.subplots()
ax.set_xlim(-5,15)
ax.set_ylim(-5,10)
col = ['#1f77b4', '#ff7f0e', '#2ca02c']

for i in np.arange(0,6,2) :
  lambda_, v = np.linalg.eig(cov[int(np.floor(i/2))])
  lambda_ = np.sqrt(lambda_)
  ax.scatter(df[:,i], df[:,(i+1)], label=label2d[int(np.floor(i/2))])
  for j in range(4):
      ell = Ellipse(xy=(np.mean(df[:,i]), np.mean(df[:,(i+1)])),
                    width=lambda_[0]*j*2, height=lambda_[1]*j*2,
                    angle=np.rad2deg(np.arccos(v[1, 0])), alpha=0.2,
                     ↪  linestyle ='-', edgecolor =  'black', linewidth
                     ↪  = 2, facecolor = col[int(np.floor(i/2))])
      ax.add_artist(ell)
    costheta.append(np.asarray(v[0])@[1,0])
leg = ax.legend()

for h, t in zip(leg.legendHandles, leg.get_texts()):
    t.set_color(h.get_facecolor()[0])
plt.show()

rotation_rad = np.arccos(costheta)
rotation_deg = np.rad2deg(rotation_rad)

for c in range(3):
...   print('περιστροφή του ελλειπτικού δεδομένων πίνακα {0} σε ακτίνια
 ↪  = {1:.2f} και μοίρες =
 ↪  {2:.0f}'.format(cov[c],rotation_rad[c],rotation_deg[c]))
...
περιστροφή του ελλειπτικού δεδομένων πίνακα [[1, 0], [0, 1]] σε ακτίνια
 ↪  = 0.00 και μοίρες = 0
```

```
περιστροφή του ελλειπτικού δεδομένων πίνακα [[1, 0.2], [0.2, 1]] σε ακτίνια = 0.79 και μοίρες = 45
περιστροφή του ελλειπτικού δεδομένων πίνακα [[1, -0.9], [-0.9, 1]] σε ακτίνια = 0.79 και μοίρες = 45
```



Σχήμα 3.7: Προσομοίωση δισδιάστατων ελλειπτικών δεδομένων

Τα χρωματιστά δεδομένα αποτελούν προσομοιώσεις απο Κανονικές κατανομές δύο διαστάσεων. Οι διαφορές τους βρίσκονται στην θέση τους :

$$\boldsymbol{\mu}_1 = (0,0)^T \qquad \boldsymbol{\mu}_2 = (5,2)^T \qquad \boldsymbol{\mu}_3 = (10,3)^T$$

αλλά και στον πίνακα Cov :

$$\text{Cov}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \text{Cov}_2 = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \qquad \text{Cov}_3 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$$

Διαφαίνεται και εμπειρικά πως η σφαιρικότητα των δεδομένων προκύπτει απο ασυσχέτιστες μεταβλητές καθώς ο πίνακας διακύμανσης-συνδιακύμανσης των μπλέ σημείων είναι ο ταυτοτικός πίνακας. Τα πορτοκαλί δεδομένα έχουν θετική ασθενή συσχέτιση, γι'αυτό δημιουργείται ένα ελλειπτικό με θετική κλίση. Τέλος, οι πράσινες παρατηρήσεις δημιουργούν μία αρκετά

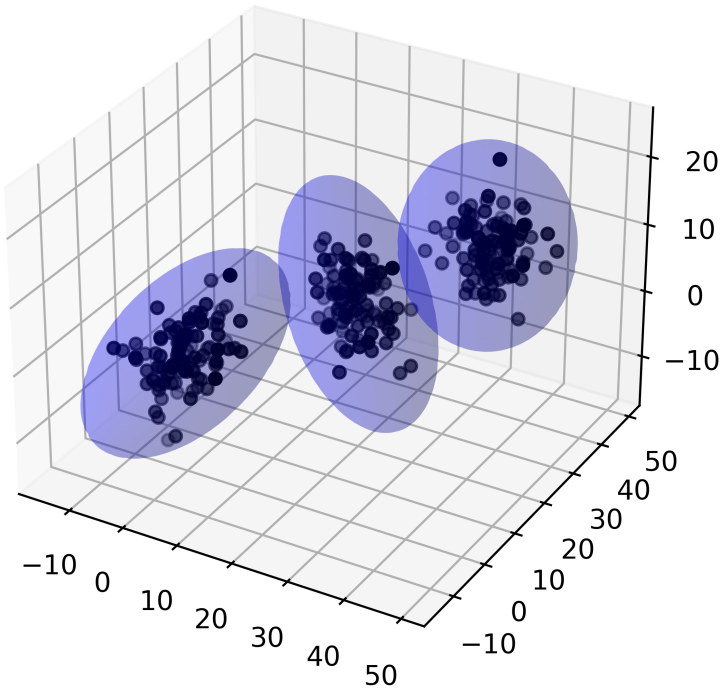
επιμηκυμένη έλλειψη με αρνητική κλίση λόγω της ισχυρής αρνητικής συσχέτισης ($\sigma_{12} = 0.9$).

Οι περιοχές πρόβλεψης των ελλειπτικών είναι οι ελλείψεις με ίδιο κέντρο (μέσο όρο) με γκρί χρώμα που περικλείουν τα δεδομένα. Το μικρότερο ελλειπτικό έχει κάλυψη 40%, το μεσαίο περίπου 87% και το μεγαλύτερο περίπου 99%, διότι αποτελούνται απο ελλειπτικά που δημιουργήθηκαν λαμβάνοντας τα πολλαπλάσια των ιδιοδιανυσμάτων τους. Συγκεκριμένα παράγονται παίρνοντας το πρώτο, δεύτερο και τρίτο πολλαπλάσιο των ιδιοτιμών τους. Τέλος, βάσει του τύπου (3.3), είναι εφικτός ο υπολογισμός τη γωνίας περιστροφής των ελλειπτικών. Απο τον αλγόριθμο (3.1.3) γραμμές 44-49, εκτυπώνεται η γωνία περιστροφής. Αναλυτικά, βρίσκεται υπολογίζοντας την γωνία του μεγαλύτερου ιδιοδιανύσματος με το διάνυσμα $(1, 0)^T$ γραμμή 34, δηλαδή τη γωνία του με τον άξονα X. Εν τέλει, θα παρουσιαστεί και η περίπτωση τρισδιάστατων δεδομένων που προκύπτουν προφανώς απο τρείς μεταβλητές καθώς και η γραφική απεικόνιση του ελλειπτικού τους.

Αλγόριθμος 3.1.4: Τρισδιάστατα Ελλειπτικά

```
A = np.array([[1.0, 0.3, 0.5],
              [0.3, 1.0, 0.2],
              [0.5, 0.2 , 1.0]])

A2 = np.array([[1.0, -0.3, -0.5],
              [-0.3, 1.0, 0.2],
              [-0.5, 0.2 , 1.0]])

A3 = np.array([[1.0, 0, 0],
              [0, 1.0, 0],
              [0, 0, 1.0]])

###_3D_Ellipse_####
cov = A
cov2 = A2
cov3 = A3
ppf = scipy.stats.chi2.ppf(0.9975,df = 3)
sr_half= ppf/2
cov, cov2, cov3 = cov*sr_half, cov2*sr_half, cov3*sr_half
plus = np.array([[sr_half,0,0],[0,sr_half,0],[0,0,sr_half]])
cov, cov2, cov3 = cov + plus, cov2+plus, cov3+plus
nx1, ny1, nz1 = multiVariate_normal([0,0,0],cov).rvs(100).T
nx2, ny2, nz2 = multiVariate_normal([20,20,4],cov2).rvs(100).T
nx3, ny3, nz3 = multiVariate_normal([35,35,10],cov3).rvs(100).T

fig = plt.figure(figsize=plt.figaspect(1))
ax = fig.add_subplot(111, projection='3d')
u = np.linspace(0, 2 * np.pi, 100)
v = np.linspace(0, np.pi, 100)

x = np.outer(np.cos(u), np.sin(v))
y = np.outer(np.sin(u), np.sin(v))
z = np.outer(np.ones_like(u), np.cos(v))
```

```
34
35
36 ellipsoid = (cov @ np.stack((x, y, z), 0).reshape(3, -1) ).reshape(3,
    ↪ *x.shape)+np.array(np.repeat([0,0,0],10000)).reshape(*ellipsoid.sha
    ↪ pe)
37 ellipsoid2 = (cov2 @ np.stack((x, y, z), 0).reshape(3, -1) ).reshape(3,
    ↪ *x2.shape)+np.array(np.repeat([20,20,4],10000)).reshape(*ellipsoid.
    ↪ shape)
38 ellipsoid3 = (cov3 @ np.stack((x, y, z), 0).reshape(3, -1) ).reshape(3,
    ↪ *x3.shape)+np.array(np.repeat([35,35,10],10000)).reshape(*ellipsoid
    ↪ .shape)
39
40 ax.plot_surface(*ellipsoid, rstride=4, cstride=4, color='b', alpha=0.2)
41 ax.scatter(nx1,ny1,nz1, c= 'k' )
42 ax.plot_surface(*ellipsoid2, rstride=4, cstride=4, color='b', alpha=0.2)
43 ax.scatter(nx2,ny2,nz2, c= 'k' )
44 ax.plot_surface(*ellipsoid3, rstride=4, cstride=4, color='b', alpha=0.2)
45 ax.scatter(nx3,ny3,nz3, c= 'k' )
46 plt.show()
```



Σχήμα 3.8: Προσομοίωση τρισδιάστατων ελλειπτικών δεδομένων

Απο το Σχήμα (3.8), ως επέκταση της δισδιάστατης περίπτωσης, μπορούμε να πάρουμε παρόμοια αποτελέσματα και στις τρείς διαστάσεις. Οι παρατηρήσεις που προκύπτουν απο ασυσχέτιστες μεταβλητές φαίνεται να δημιουργούν μία τρισδιάστατη σφαίρα (το 3ο απο αριστερά σχήμα). Αντιθέτως, τα δεδομένα που παρουσιάζουν συσχετίσεις μεταξύ των μεταβλητών τους δημιουργούν τρισδιάστατα ελλειπτικά που μοιάζουν με μπάλες του ράγκμπι. Εύκολα διακρίνεται το ελλειπτικό που έχει μία αρνητική καθοδική τάση, προκύπτει απο τον πίνακα A2 στην 5η γραμμή του κώδικα 1.4, λόγω της αρνητικής σχέσης της X με την Y και Z ($\text{Cov}_{A2}(X,Y) < 0$ $\text{Cov}_{A2}(X,Z) < 0$). Τέλος, το πρώτο ελλειπτικό απο αριστερά έχει θετική κλίση μιας και όλες οι ανα δύο συσχετίσεις των μεταβλητών είναι θετικές (βλέπε πίνακα A κώδικας 1.4 γραμμή 1).

3.2 Κατανομή Student

Η κατανομή Student-t είναι ιδιαίτερα χρήσιμη καθώς ο ρόλος της είναι σημαντικός στους ελέγχους υποθέσεων και στην κατασκευή διαστημάτων εμπιστοσύνης. Χρησιμοποιείται στα Χρηματοοικονομικά αλλα και σε πολλούς άλλους επιστημονικούς τομείς όπως σεισμολογία, μετεωρολογία, χρονολογικές σειρές κτλ. Ειδικά στον τομέα ανάλυσης κινδύνου και χαρτοφυλακίων, έχει βρεί ιδιαίτερες εφαρμογές στην μελέτη ακραίων γεγονότων. Άξιο αναφοράς είναι πως η Student αποτελεί μία κατανομή της ευρύτερης οικογένειας των κατανομών με Βαριές Ουρές οι οποίες έλαβαν το όνομα αυτό λόγω της μεγαλύτερης πιθανότητας ακραίων γεγονότων απο την Κανονική κατανομή, δοθέντος ότι η σύγκριση γίνεται υπο τον ίδιο μέσο όρο και διακύμανση. Αναλυτικότερα θα επεξηγηθεί αυτή η ιδιότητα παρακάτω. Απο ιστορική σκοπιά, αυτού του είδους οι κατανομές μελετήθηκαν πρώτα απο τον *Vilfredo Pareto* και απο τον *Paul Levy*. Όσον αφορά την Student, αυτή αναλύθηκε πρώτα απο τον Gosset (1908) ο οποίος την δημοσίευσε με το όνομα Student (Wolfgang and Leopold (2014)).

Αρχικά, για την κατανόησή της τόσο θεωρητικά όσο και πρακτικά, θα ξεκινήσουμε απο την μονοδιάστατη εκδοχή και στην συνέχεια θα επεκταθούμε σε πολυδιάστατες δομές.

3.2.1 Πυκνότητα Πιθανότητας

Ορισμός 3.2.1 (*Μονοδιάστατη Κατανομή Student-t*)

Έστω ότι η μεταβλητή X είναι Κανονικά κατανεμημένη με μέσο $\mu = 0$ και διακύμανση σ^2 και Y μία άλλη τυχαία μεταβλητή τέτοια ώστε το $\frac{\Upsilon^2}{\sigma^2}$ να ακολουθεί κατανομή χ^2 με n βαθμούς ελευθερίας. Άν οι X , Y είναι ανεξάρτητες και ορίσουμε την τυχαία μεταβλητή :

$$Z = \frac{X \cdot \sqrt{n}}{Y}$$

τότε η Z ακολουθεί κατανομή Student-t με n βαθμούς ελευθερίας, και συνάρτηση πυκνότητας πιθανότητας :

$$f_Z(z;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\cdot\Gamma(\frac{n}{2})} \cdot \left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}}$$

με $-\infty < z < \infty$ όπου $\Gamma(\alpha) = \int_0^\infty z^{a-1}\cdot e^{-z}\,dz$.

Για την κατανομή t αποδεικνύεται ότι :

1. μέσος $(\mu) = 0$.

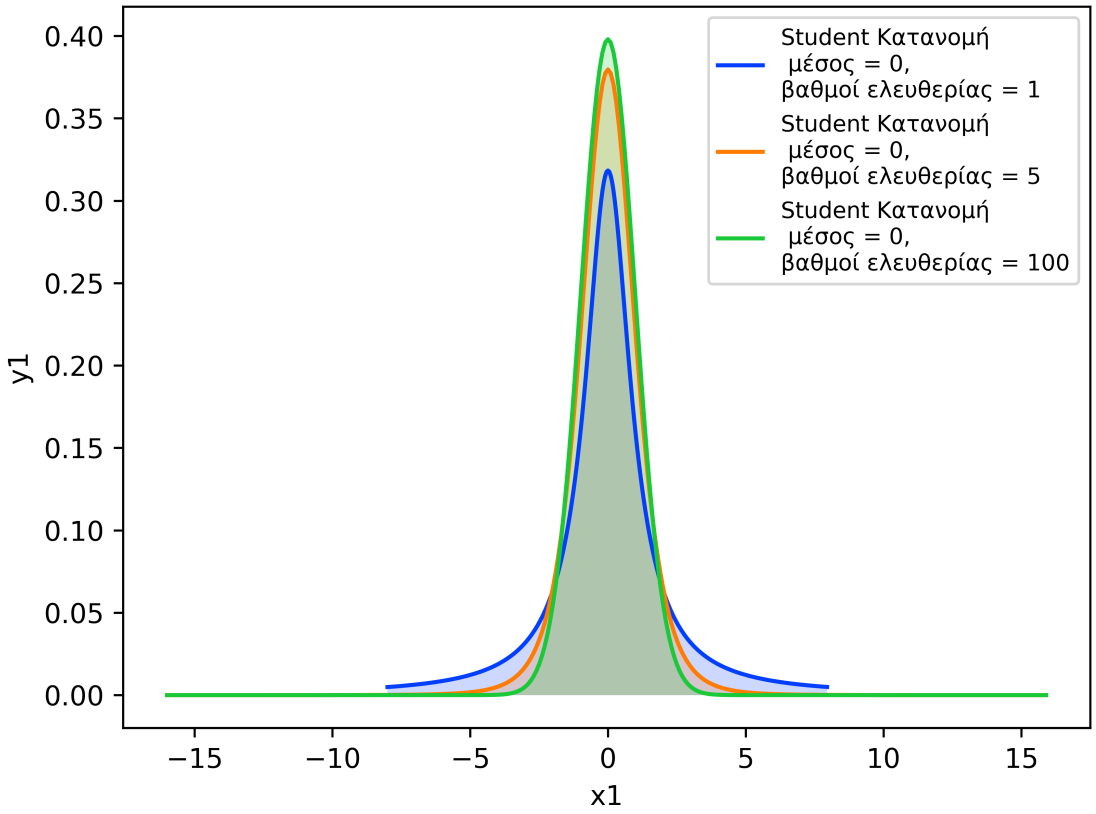
2. Διακύμανση $(\sigma^2) = \frac{n}{n-2}$.
3. Λοξότητα $= 0$.
4. Κύρτωση $= 3 + \frac{6}{n-4}$.

Η μηδενική λοξότητα σημαίνει πως η κατανομή είνα συμμετρική, ενώ η κύρτωση φαίνεται πως είναι μεγαλύτερη του 3 (για $n > 4$). Γι' αυτό το λόγο η Student αποτελεί πυκνότητα με βαριές ουρές μιας και η κύρτωσή της είναι μεγαλύτερη απο αυτή της Κανονικής κατανομής. Αυτός είναι ο λόγος που μπορεί να περιγράψει ικανοποιητικά ακραία γεγονότα, κάτι το οποίο θα ήταν σχεδόν αδύνατο με την Κανονική κατανομή αφού διαθέτει λεπτές ουρές. Παρακάτω θα απεικονιστούν γραφήματα τα οποία θα παρουσιάσουν τόσο την Student όσο και την σύγκρισή της με την Κανονική.

Αλγόριθμος 3.2.1: Student-t Κατανομές Διαφορετικών Βαθμών Ελευθερίας

```
import math
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
from scipy.stats import \multiVariate_normal
from scipy.stats import t
from scipy.stats import norm
import pandas as pd
from matplotlib import cm
from matplotlib.ticker import LinearLocator
from mpl_toolkits.mplot3d import axes3d

x1 = np.arange(-8,8,0.05)
x2 = np.arange(-16,16,0.1)
t1 = t(df=1)
t2 = t(df=5)
t3 = t(df=100)

y1 = t1.pdf(x1)
y2 = t2.pdf(x2)
y3 = t3.pdf(x2)
yn = norm.pdf(x1,0,1)
arr1 = np.vstack([x1,y1]).T
arr2 = np.vstack([x2,y2]).T
arr3 = np.vstack([x2,y3]).T
arrn = np.vstack([x1,yn]).T

df1 = pd.DataFrame(arr1, columns=['x1','y1'])
df2 = pd.DataFrame(arr2, columns=['x2','y2'])
df3 = pd.DataFrame(arr3, columns=['x3','y3'])
dfn = pd.DataFrame(arrn, columns=['xn','yn'])
sns.set_palette("bright")
```

```
lp1 = sns.lineplot(data= df1,x = 'x1',y = 'y1',label = 'Student Κατανομή \n μέσος = 0,\nβαθμοί ελευθερίας = 1')
lp2 = sns.lineplot(data= df2, x = 'x2', y = 'y2', label ='Student Κατανομή \n μέσος = 0,\nβαθμοί ελευθερίας = 5')
lp3 = sns.lineplot(data= df3, x = 'x3', y = 'y3', label ='Student Κατανομή \n μέσος = 0,\nβαθμοί ελευθερίας = 100')
'''
lpn = sns.lineplot(data=dfn, x = 'xn', y = 'yn', label ='Student Κατανομή \n μέσος = 0,\nβαθμοί ελευθερίας = 100')
'''
lp1.fill_between(x1, y1, interpolate=True, alpha =0.2)
lp2.fill_between(x2, y2, interpolate=True, alpha =0.2)
lp3.fill_between(x2, y3, interpolate=True, alpha =0.2)
'''
lpn.fill_between(x1, yn, interpolate=True, alpha =0.2)
'''
plt.legend(fontsize='small'); plt.show()
```



Σχήμα 3.9: Student-t πυκνότητες πιθανοτήτων για διάφορους βαθμούς ελευθερίας

Σχηματικά, η κατανομή Student-t είναι παρόμοια με την Κανονική κατανομή. Η βασική τους διαφορά έγκειται στο οτι η Student έχει πιο παχιές ουρές. Η μορφή της κατανομής Student επηρεάζεται απο τους βαθμούς ελευθερίας. Όσο οι βαθμοί ελευθερίας αυξάνονται, τόσο

λεπταίνουν οι ουρές τις και η σ.π.π. τείνει να γίνει η Κανονική κατανομή. Προχωρώντας στην πολυδιάστατη περίπτωση έχουμε τον εξής ορισμό :

Ορισμός 3.2.2 (*Πολυδιάστατη Κατανομή Student-t*)

Έστω $\boldsymbol{X}$, Y *ανεξάρτητες τ.μ. με* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *και* $Y \sim \chi^2_n$ *με* $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{X}\sqrt{\frac{n}{Y}}$. *Τότε η* $\boldsymbol{Z}$ *έχει κατανομή Student-t, συμβολικά ως* $\boldsymbol{Z} \sim t_p(n, 0, \boldsymbol{I}_p)$, *αν η σ.π.π. γράφεται ως :*

$$f_{\boldsymbol{Z}}(\boldsymbol{z}; n, \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}) n^{\frac{p}{2}} \pi^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot [1 + \frac{1}{n}(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})]^{-\frac{n+p}{2}}$$

με n *βαθμούς ελευθερίας, θετικά ορισμένο πίνακα* $\boldsymbol{\Sigma}$, *μέσο* $\boldsymbol{\mu}$ *και* $\mathrm{Cov}(\boldsymbol{Z}) = \frac{n}{n-2}\boldsymbol{\Sigma}$.

Σχετικά με την Στοχαστική Αναπαράσταση στην περίπτωση που έχουμε $\boldsymbol{X} \sim t_p(n, 0, \boldsymbol{I}_p)$ δηλαδή Student p-διάστατη με διανυσματικό μέσο μηδέν και πίνακα $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, n βαθμών ελευθερίας τότε $R^2 = \boldsymbol{X}^T\boldsymbol{X}$ και επειδή η Student γράφεται ως $\sqrt{W}\boldsymbol{Z}$ με $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ Κανονική τυποποιημένη κατανομή και $W = \frac{n}{V}$, $V \sim \chi^2_n$ τότε $R^2 = \boldsymbol{X}^T\boldsymbol{X} = \sqrt{W}\boldsymbol{Z}^T\boldsymbol{Z}\sqrt{W} = W\boldsymbol{Z}^T\boldsymbol{Z}$. Οπότε καταλήγουμε :

$$\frac{R^2}{p} = \frac{\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{p}}{\frac{\frac{W}{n}}{n}} = \frac{\frac{\chi^2_p}{p}}{\frac{\chi^2_n}{n}} \sim F(p, n).$$

Εφόσον η F ορίζεται ως πηλίκο ανεξάρτητων χ^2 κατανομών. Εύκολα πλέον προκύπτει η κατανομή του R, καθώς $R \sim \sqrt{p \cdot F(p, n)}$. Απο την σχέση $\boldsymbol{X} = R\boldsymbol{S}$ της στοχαστικής αναπαράστασης, μπορούμε να παράγουμε δείγμα απο την Student βάσει των $\boldsymbol{S}$ (προσομοίωση στον μοναδιαίο κύκλο) και του R (απο το $\sqrt{p \cdot F(p, n)}$). Άν θελήσουμε πολυδιάστατη ελλειπτική Student-t, τότε το μόνο που έχουμε να κάνουμε είναι να πολλαπλασιάσουμε την σφαιρική Student-t ($R\boldsymbol{S}$) με έναν πίνακα $\boldsymbol{A}$ όπου για τον πίνακα $\boldsymbol{\Sigma}$ να ισχύει $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$, παίρνοντας έτσι την μορφή $\boldsymbol{X} = \boldsymbol{A}R\boldsymbol{S} = R\boldsymbol{A}\boldsymbol{S}$.

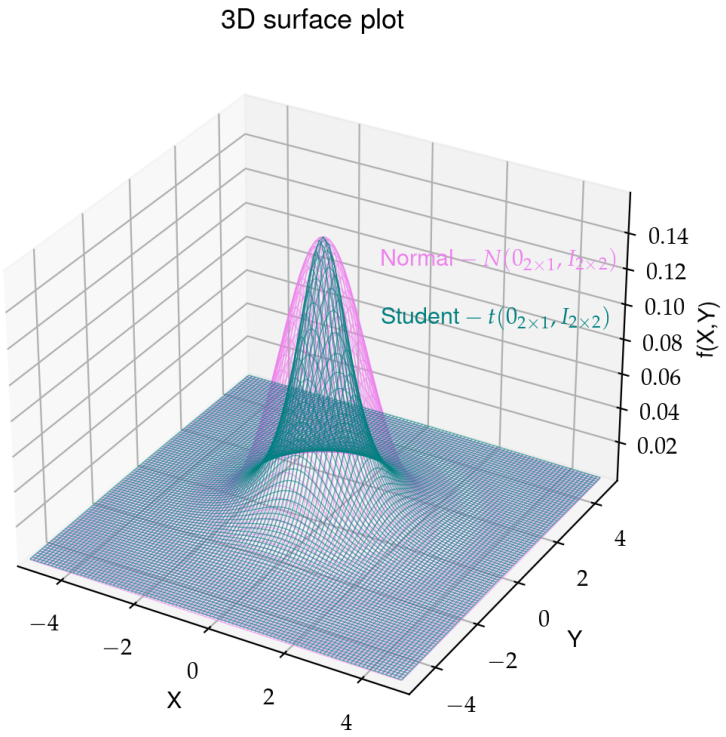
Στη συνέχεια παρουσιάζεται τρισδιάστατο γράφημα που απεικονίζει την σ.π.π της Student-t καθώς και της Κανονικής κατανομής για ίδιους μέσους και πίνακες $\boldsymbol{\Sigma}$.

3.2.2 Τρισδιάστατη Γραφική Απεικόνιση

Αλγόριθμος 3.2.2: Δισδιάστατες Student-t και Κανονική Κατανομή

```
from scipy.stats import multiVariate_t
import matplotlib
matplotlib.rcParams['text.usetex'] = True

mv_t = multiVariate_t(loc = [0,0],df = 1)
x = np.arange(-5,5,0.1)
y = np.arange(-5,5,0.1)
X,Y = np.meshgrid(x,y)
z = np.array([X.reshape(1,10000)[0],Y.reshape(1,10000)[0]]).T
Z = mv_t.pdf(z)
znorm = multiVariate_normal([0,0],[[1,0],[0,1]]).pdf(z)
Z, znorm = Z.reshape(100,100), znorm.reshape(100,100)
```

```
13
14 fig = plt.figure(figsize=(12,6))
15 ax = fig.add_subplot(111, projection='3d')
16
17 surf = ax.plot_wireframe(X, Y, Z, rstride=3, cstride=3, color ='teal',
   ↪  alpha=0.5)
18 surf_norm = ax.plot_wireframe(X, Y, znorm, rstride=3, cstride=3,
   ↪  color='violet', alpha=0.5)
19
20 ax.set_xlabel('X')
21 ax.set_xlim(-5, 5)
22 ax.set_ylabel('Y')
23 ax.set_ylim(-5, 5)
24 ax.set_zlabel('Z')
25 ax.set_zlim(np.min(Z), np.max(Z))
26 ax.set_title('3D surface plot')
27
28 eq_norm = r"$\text{Normal}-N(\bm{0}_{2\times 1},I_{2\times 2})$"
29 eq_st = r"$\text{Student}-t(0_{2\times 1}, I_{2\times 2})$"
30 ax.text(-0.7,4,0.10, eq_norm, color = 'violet')
31 ax.text(0.7,1.5,0.10, eq_st, color = 'teal')
32 plt.show()
```



Σχήμα 3.10: Δισδιάστατες πυκνότητες Student και Κανονικής (Σύγκριση)

Βλέπουμε πως η Κανονική κατανομή δύο διαστάσεων διατηρεί μεγαλύτερη πυκνότητα σε μικρές αποστάσεις απο τον διανυσματικό μέσο, ενώ για μεγαλύτερες αποστάσεις υπερέχει η Student (Φαίνεται να διακρίνεται περισσότερο το μπλέ χρώμα). Λογικό λαμβάνοντας υπόψιν πως η Student διατηρεί μεγαλύτερη πυκνότητα σε ακραίες παρατηρήσεις.

3.3 Κατανομή Kotz

Πέραν των γνωστών κατανομών όπως η Κανονική και η Student, υπάρχει μία τεράστια πληθώρα απο Σφαιρικές και Ελλειπτικές κατανομές που γενικεύουν τις ήδη γνωστές αλλά και αυξάνουν τις δυνατότητες ανάλυσης δεδομένων. Μία τέτοια κατανομή είναι και η Kotz. Συγκεκριμένα στο παρόν υποκεφάλαιο θα μελετηθεί μία ειδική μορφή ευρύτερων οικογενειών Ελλειπτικών κατανομών που ορίζεται ως τύπου Kotz (Kotz type distribution). Προτάθηκαν απο τους Fang et al. (1990). Βέβαια συνδέονται και με τις πολυδιάστατες κατανομές που προτάθηκαν απο τον Simoni (1968) άλλα και με αυτές που παρουσιάστηκαν απο τους Gomez et al. (1998) ως Power Exponential.
Αρχικά θα ξεκινήσουμε με την παρουσίαση του τύπου των γενικών Kotz type κατανομών. Συγκεκριμένα η συνάρτηση πυκνότητας πιθανότητας (pdf) ενός τυχαίου διανύσματος $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$:

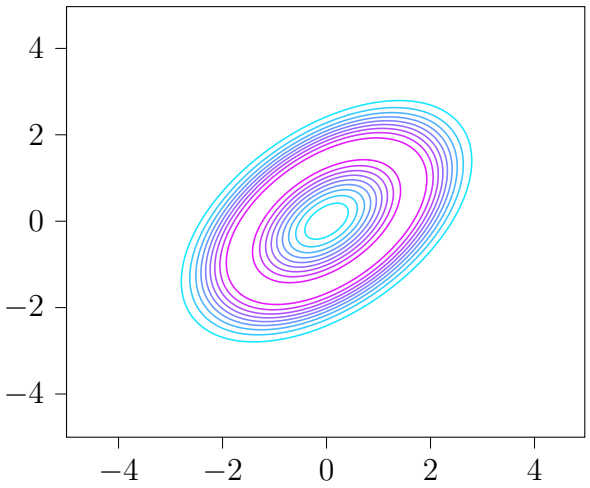
$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = C_p |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot [(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]^{N-1} \cdot e^{-r[(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})]^s} \tag{3.4}$$

όπου $r \geq 0$, $s > 0$, $N \geq \frac{2-p}{2}$, $\boldsymbol{\Sigma}_{p\times p}$ ένας θετικά ορισμένος πίνακας και $\boldsymbol{\mu}_{p\times 1}$ ο διανυσματικός μέσος. Το $C_p = \frac{s\Gamma(\frac{p}{2})}{\pi^{\frac{p}{2}}\Gamma(\frac{2N+p-2}{2s})} \cdot r^{\frac{2N+p-2}{2s}}$ αποτελεί την σταθερά κανονικοποίησης.

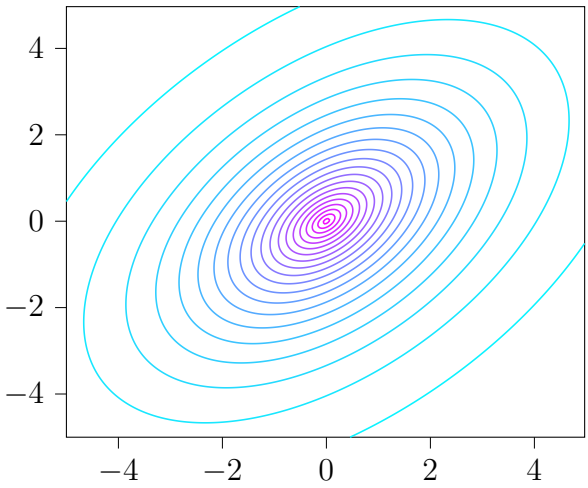
Συνήθως μελετάται μια ειδική περίπτωση της Kotz με $N = 1, s = \frac{1}{2}, r = 1$, της οποίας η πυκνότητα έχει τη μορφή :

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = C_p |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-[(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})]^{\frac{1}{2}}}. \tag{3.5}$$

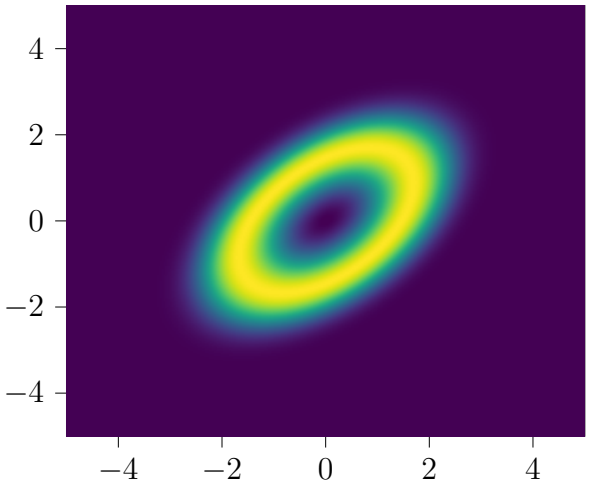
με $\mathrm{Cov}(\boldsymbol{X}) = (p+1)\boldsymbol{\Sigma}$. Παρακάτω θα παρουσιαστούν γραφικές αναπαραστάσεις της Kotz κατανομής δύο διαστάσεων για διάφορες τιμές των παραμέτρων της. Κύριος σκοπός αυτών των γραφημάτων είναι να αποτυπωθούν οι διάφορες μορφές της.



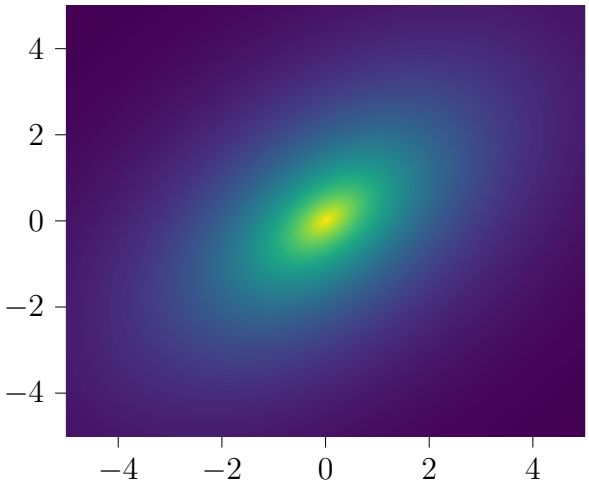
Σχήμα 3.11: Πυκνότητα δισδιάστατης κατανομής Kotz με $p = 2, r = 1, s = 2, N = 2$, και $\mathrm{Cov}(X, Y) = 0.5$



Σχήμα 3.12: Πυκνότητα δισδιάστατης κατανομής Kotz με $p = 2, s = 0.5, r = 1, N = 1$, και $\mathrm{Cov}(X, Y) = 0.5$



Σχήμα 3.13: Εναλλακτική αναπαράσταση δισδιάστατης κατανομής Kotz με $p = 2, r = 1, s = 2, N = 2$, και $\mathrm{Cov}(X, Y) = 0.5$



Σχήμα 3.14: Εναλλακτική αναπαράσταση δισδιάστατης κατανομής Kotz με $p = 2, s = 0.5, r = 1, N = 1$, και $\mathrm{Cov}(X, Y) = 0.5$

Φαίνεται πως το σχήμα της Kotz επηρεάζεται σε μεγάλο βαθμό απο την παράμετρο N όπως φαίνεται απο το Σχήμα (3.13) όπου εκεί η πυκνότητα έχει παράμετρο $N = 1$ ενώ στο Σχήμα (3.14) το $N = 2$.

Όπως προαναφέρθηκε, η Kotz αποτελεί κατηγορία των Ελλειπτικών κατανομών συνεπώς μπορεί να λάβει την χρήσιμη Στοχαστική Αναπαράσταση (Βλέπε Κεφ 1.) $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{R^2}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{S}$, όπου $\boldsymbol{S}_{p\times 1}$ ομοιόμορφα κατανεμημένο στην μοναδιαία σφαίρα και το R (μή αρνητική μεταβλητή ανεξάρτητη της $\boldsymbol{S}$, βλέπε Κεφ 1.) με πυκνότητα (pdf) :

$$f_{R^2}(v) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2})} \cdot v^{\frac{p}{2}-1} \cdot g(v)$$
$$= \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2})} C_p \cdot v^{\frac{p}{2}+N-2} \cdot e^{-rv^s},$$

δηλαδή $g(v) = C_p \cdot v^{N-1} e^{-rv^s}$. Η Kotz είναι ιδιαίτερα χρήσιμη κατανομή καθώς διαθέτει πιο παχιές ουρές απο την Κανονική Κατανομή και μπορεί να χρησιμοποιηθεί όταν δεν μπορεί να προσαρμοστεί ικανοποιητικά η Κανονική στα δεδομένα, όπως στην περίπτωση πολλών ακραίων παρατηρήσεων.

3.3.1 Περιθώριες και Δεσμευμένες Κατανομές

Μπορούμε να ορίσουμε τις περιθώριες κατανομές της Kotz θεωρώντας μια διαμέριση $\boldsymbol{X} = (\boldsymbol{X}_{(1)}^T, \boldsymbol{X}_{(2)}^T)^T$ με $\boldsymbol{X}_{(1)} = (X_1, \ldots, X_k)^T$ και $\boldsymbol{X}_{(2)} = (X_{k+1}, \ldots, X_p)^T$ με $k < p$. Επίσης και ο πίνακας $\boldsymbol{\Sigma}$ διαμερίζεται ως :

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Τότε η $\boldsymbol{X}_{(1)} \sim \mathcal{EC}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, g_1)$. Η θέση (μέσος) και ο πίνακας διακύμανσης-συνδιακύμανσης είναι $\mathbb{E}(\boldsymbol{X}_{(1)}) = \boldsymbol{\mu}_1$, $\mathrm{Cov}(\boldsymbol{X}_{(1)}) = (p+1)\boldsymbol{\Sigma}_{11}$. Αντίστοιχα για τις δεσμευμένες κατανομές η κατανομή του $\boldsymbol{X}_{(2)}|\boldsymbol{X}_{(1)} \sim \mathcal{EC}_{p-k}(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{22|1}, g_{2|1})$ (Ελλειπτική κατανομή) όπου $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_{(1)} - \boldsymbol{\mu}_{(1)})$ και $\boldsymbol{\Sigma}_{22|1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$.

Η σημαντική πληροφορία απο αυτή την ενότητα είναι κυρίως πως οι περιθώριες και οι δεσμευμένες κατανομές της Kotz είναι κι αυτές Ελλειπτικές μιας και ανήκει στην οικογένεια των Ελλειπτικών κατανομών.

3.3.2 Εκτίμηση Παραμέτρων

Στην Στατιστική μία τυπική και ισχυρή μέθοδος εκτίμησης της άγνωστης παραμέτρου είναι με την μέθοδο της Μέγιστης Πιθανοφάνειας. Χρησιμοποιείται ευρέως για την εκτίμηση του μέσου της Κανονικής κατανομής αλλά και για άλλες κατανομές και παραμέτρους. Η ίδια μέθοδος θα χρησιμοποιηθεί και για την εύρεση του μέσου της κατανομής Kotz. Στην ειδική περίπτωση με τύπο :

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = C_p |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot e^{-[(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})]^{\frac{1}{2}}}. \tag{3.6}$$

Ξεκινώντας απο την απο κοινού πιθανοφάνεια της, θα λογαριθμήσουμε έτσι ώστε να την μετατρέψουμε σε μια πιο διαχειρίσιμη μορφή :

$$\ln \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = n \ln C_p - \frac{n}{2} \ln |\boldsymbol{\Sigma}| - \sum_{i=1}^{n} \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}. \tag{3.7}$$

Οι εκτιμητές μέγιστης πιθανοφάνειάς των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ βρίσκονται μεγιστοποιώντας την ποσότητα (3.7). Συγκεκριμένα ο εκτιμητής του μέσου $\boldsymbol{\mu}$ της Kotz κατανομής ονομάζεται Spatial Median

όταν έχουμε σφαιρικά δεδομένα $\boldsymbol{\Sigma} = \boldsymbol{I_p}$ (Haldane (1948)) ή Generalized Spatial Median για οποιοδήποτε άλλο πίνακα $\boldsymbol{\Sigma}$ (Rao (1988)). Όταν ως εκτίμηση του πίνακα $\boldsymbol{\Sigma}$ χρησιμοποιήσουμε τον δειγματικό πίνακα συνδιακυμάνσεων $\boldsymbol{S}$, ο εκτιμητής μέγιστης πιθανοφάνειας (ε.μ.π.) $\hat{\boldsymbol{\mu}}$ ενός συνόλου $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ ελαχιστοποιεί την ποσότητα :

$$\sum_{i=1}^{n} \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}.$$

Εναλλακτικά αν λάβουμε και το $\boldsymbol{\Sigma}$ ως παράγοντα στη βελτιστοποίηση, τότε βρίσκονται τα βέλτιστα $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ ελαχιστοποιώντας την ποσότητα :

$$\frac{n}{2} \ln |\boldsymbol{\Sigma}| + \sum_{i=1}^{n} \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}$$

ως προς $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$. Αν και δεν υπάρχει κάποιος τύπος του ε.μ.π. (εκτιμητή μέγιστης πιθανοφάνειας) του μέσου, έχουν αναπτυχθεί αλγοριθμικές μέθοδοι που οδηγούν σε ε.μ.π. εκτιμητές τόσο για τον μέσο όσο και για τον πίνακα $\boldsymbol{\Sigma}$.

3.3.3 Ασυμπτωτική Συμπεριφορά του ε.μ.π. Εκτιμητή του Μέσου

Βάσει των Huber (1981), Ducharme and Milasevic (1987) και Naik (1993) ορίζεται η ασυμπτωτική κατανομή της ε.μ.π. του μέσου :

Θεώρημα 3.3.1 (*Ασυμπτωτική Κατανομή του ε.μ.π. εκτιμητή του μέσου*)

Έστω $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ ένα δείγμα απο μία p-διάστατη τυχαία μεταβλητή $\boldsymbol{X}$ που ακολουθεί την κατανομή Kotz (3.6) με παραμέτρους $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ και $\hat{\boldsymbol{\mu}}$ ο ε.μ.π. του μέσου. Τότε ισχύει

$$\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to N_p(\boldsymbol{0}, \boldsymbol{\Sigma} \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1} \boldsymbol{\Sigma}).$$

Για τους πίνακες $\boldsymbol{A}$, $\boldsymbol{B}$ μπορούν να χρησιμοποιηθούν οι παρακάτω εκτιμήτριες :

$$\hat{\boldsymbol{B}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T}{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})},$$

$$\hat{\boldsymbol{A}} = \frac{1}{n} \sum_{1}^{n} \frac{1}{\sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})}} \cdot \left[\hat{\boldsymbol{\Sigma}} - \frac{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T}{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})} \right].$$

Απο την ασυμπτωτική κατανομή, εφόσον οδηγεί σε πολυδιάστατη Κανονική κατανομή, μπορεί να δημιουργηθεί έλεγχος υπόθεσης για το μέσο της Kotz :

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$
$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

χρησιμοποιώντας την παρακάτω τετραγωνική μορφή που ακολουθεί χ^2 κατανομή :

$$n(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \hat{\boldsymbol{\Omega}}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \sim \chi^2_p,$$

όπου $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{A}}^{-1} \hat{\boldsymbol{B}} \hat{\boldsymbol{A}}^{-1} \hat{\boldsymbol{\Sigma}}$.

3.3.4 Ιδιότητες

Σε αυτή την ενότητα θα μελετηθούν η λοξότητα και η κύρτωση της κατανομής Kotz όπως ορίστηκε στην (3.6). Αυτά τα δύο μέτρα είναι αρκετά σημαντικά καθώς μπορούν να χρησιμοποιηθούν για τον έλεγχο υπόθεσης οτι τα δεδομένα ακολουθούν Kotz κατανομή. Στην μονοδιάστατη περίπτωση υπάρχουν κλειστοί τύποι για αυτές τις δύο ροπές. Εδώ πλέον οι συντελεστές ορίζονται ως πολυμεταβλητή λοξότητα (Mardia's multiVariate Skewness) συμβολικά ως β_{1p} και πολυμεταβλητή κύρτωση (Mardia's multiVariate Kyrtosis) ως β_{2p}, μιας και βρισκόμαστε σε πολλές διαστάσεις. Ως προς την Kotz, οι παραμετρικές τιμές είναι $\beta_{1p} = 0$, και $\beta_{2p} = \frac{p(p+2)(p+3)}{(p+1)}$. Απο ένα σύνολο δεδομένων, οι εκτιμήσεις είναι οι εξής :

$$\hat{\beta}_{1p} = \frac{\sum_{i=1}^n \sum_{j=1}^n [(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})]^3}{n^2}$$

$$\hat{\beta}_{2p} = \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}{n},$$

όπου $\boldsymbol{S}$ ο δειγματικός πίνακας συνδιακυμάνσεων. Για τα παραπάνω μέτρα έχουν βρεθεί οι συγκλίσεις κατά κατανομή (Baringhaus and Henze (1992)) :

$$n\beta_{1p} \to a_1 \chi^2_p + a_2 \chi^2_{\frac{p(p-1)(p+4)}{6}},$$

$a_1 = \frac{3}{p}[\frac{m_6}{p+2} - 2m_4 + p(p+2)]$ και $a_2 = \frac{6m_6}{p(p+2)(p+4)}$ όπου $m_4 = \frac{p(p+2)(p+3)}{p+1}$, $m_6 = \frac{p(p+2)(p+3)(p+4)(p+5)}{(p+1)^2}$. Απο την άλλη, ασυμπτωτικά για την κύρτωση ισχύει οτι :

$$\sqrt{n}\Big(\beta_{2p} - p(p+1)(p+2)(p+3)\Big) \to N(0, \tau^2),$$

όπου $\tau^2 = r_8 - r_4^2 + \frac{4}{p} r_4 (\frac{r_4^2}{p} - r_6)$. Γενικά για την κατανομή Kotz με πυκνότητα (3.6) ισχύει $r_k = \mathbb{E}(R^k) = p(p+1)(p+2)\cdots(p+(k-1))$, $k \geq 1$. Ο κύριος λόγος που παρουσιάστηκαν αυτές οι ροπές είναι διότι μπορούν να χρησιμοποιηθούν για τον έλεγχο της υπόθεσης αν τα δεδομένα προέρχονται απο την κατανομή Kotz (3.6). Παρόμοιοι έλεγχοι υπάρχουν όπως ο Jarque-Bera που για τον έλεγχο Κανονικότητας χρησιμοποιούνται οι τιμές της λοξότητας και κύρτωσης.

3.4 Πολυδιάστατη Συμμετρική Κατανομή Laplace

Σε αυτή την ενότητα θα μελετηθεί μία άλλη Ελλειπτικής μορφής κατανομή με όνομα Πολυμεταβλητή Laplace ή Ελλειπτική Laplace. Αυτή η κατανομή διαθέτει δύο μορφές, συγκεκριμένα την συμμετρική και την ασύμμετρη. Στην διπλωματική αυτή θα παρουσιαστεί η συμμετρική Laplace. Γενικά μία κατανομή ορίζεται ως Συμμετρική Laplace, συμβολικά $L_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

με μέσο όρο μηδέν και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\boldsymbol{\Sigma}$ θετικά ορισμένο άν η χαρακτηριστική της συνάρτηση δίνεται ως :

$$\phi_{\boldsymbol{X}-\boldsymbol{\mu}}(\boldsymbol{t}) = \frac{1}{1+\frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}}.$$

Απο τον τύπο αυτό προκύπτει άμεσα πως πρόκειται για Ελλειπτική κατανομή διότι είναι συνάρτηση του $\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$ (2.2)

3.4.1 Πυκνότητα Πιθανότητας

Η πυκνότητα πιθανότητας της $\boldsymbol{X} \sim L_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ είναι η εξής :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot \left(\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)^{\frac{\nu}{2}} \cdot K_\nu\left(\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}\right),$$

όπου $\nu = \frac{2-p}{2}$ με p η διάσταση και $K_\nu(\cdot)$ είναι η τροποποιημένη συνάρτηση Bessel τρίτου είδους που ορίζεται ως εξής :

$$K_\nu(u) = \frac{1}{2}\left(\frac{u}{2}\right)^\nu \cdot \int_0^\infty z^{-\nu-1}\, e^{-z-\frac{u^2}{4z}}\, dz.$$

Μία διαφορετική έκφραση της πυκνότητας που προκύπτει απο την εφαρμογή ασυμπτωτικού τύπου για την συνάρτηση Bessel (βλέπε Daojing et al. (2008)) είναι η παρακάτω :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)^{\frac{p}{2}-1}}},$$

με συμβολισμό $L_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ για οποιοδήποτε $\boldsymbol{\mu}$. Αυτός ο τύπος θα χρησιμοποιηθεί στο Κεφάλαιο της Διαχωριστικής Ανάλυσης. Γενικά είναι μία κατανομή που διαθέτει πιό παχιές ουρές απο την Κανονική κατανομή και γι'αυτό το λόγο μπορεί να χρησιμοποιηθεί για την προσαρμογή σε δεδομένα με ακραίες παρατηρήσεις, κάτι που θα ήταν αδύνατο με την Κανονική. Στην ειδική περίπτωση δύο διαστάσεων, ο τύπος της Laplace $L_2(\mathbf{0}, \boldsymbol{\Sigma})$ έχει την μορφή :

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot K_0\left(\sqrt{\frac{2(\frac{x_1^2}{\sigma_1^2} - \frac{2\rho x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2})}{1-\rho^2}}\right)$$

με μέσο $\boldsymbol{\mu} = (0,0)^T$ και πίνακα $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix},$$

όπου ρ ο συντελεστής συσχέτισης των δύο μεταβλητών.

Όσον αφορά την Στοχαστική Αναπαράσταση της πολυδιάστατης συμμετρικής κατανομής Laplace $\boldsymbol{X} \sim L_p(\mathbf{0}, \boldsymbol{\Sigma})$ ισχύει πως γράφεται ως :

$$\boldsymbol{X} = R\boldsymbol{A}\boldsymbol{S}$$

όπου $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$, μεταβλητή $\boldsymbol{S}$ κατανεμημένη στην μοναδιαία υπερσφαίρα και τ.μ. R με κατανομή :

$$f_R(r) = \frac{2r^{\frac{p}{2}} \cdot K_{\frac{p}{2}-1}(r\sqrt{2})}{\sqrt{2^{\frac{p}{2}-1}} \cdot \Gamma(\frac{p}{2})}, r > 0$$

όπου K_ν η συνάρτηση Bessel τρίτου είδους. Παρακάτω θα προσομοιωθούν δεδομένα απο Laplace και Κανονική Κατανομή για να παρουσιαστούν οι διαφορές τους στις δύο διαστάσεις. Για την παραγωγή δεδομένων απο την συμμετρική Laplace $\boldsymbol{X} \sim L_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ θα εφαρμόσουμε το αποτέλεσμα των Kozubowski and Panorska (1999) :

Θεώρημα 3.4.1 (*Βασική Αναπαράσταση της Ελλειπτικής Κατανομής Laplace*)

Έστω $\boldsymbol{Y} \sim L_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ *και* $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. *Επίσης έστω τ.μ.* $W \sim exp(\lambda = 1)$. *Τότε ισχύει*

$$\boldsymbol{Y} = \sqrt{W} \cdot \boldsymbol{X}$$

Συνεπώς ένας τρόπος προσομοίωσης τιμών απο μία τ.μ. $\boldsymbol{Y} \sim L_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ είναι ο εξής :

Ψευδοκώδικας Προσομοίωση τιμών απο $L_2(\boldsymbol{0}, \boldsymbol{\Sigma})$

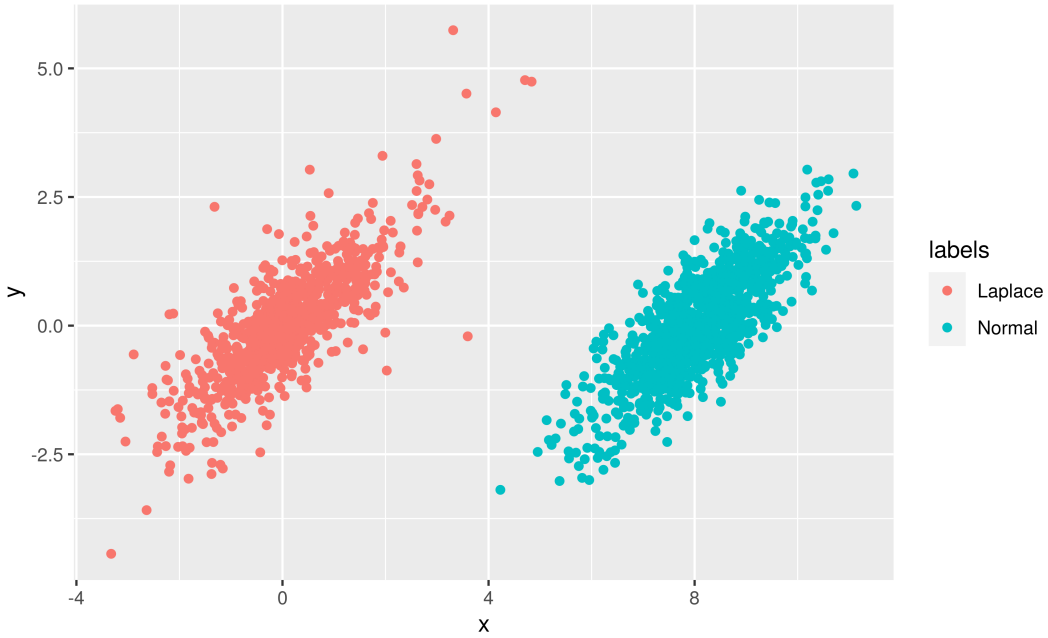
1: Παρήγαγε δεδομένα απο Κανονική κατανομή $\boldsymbol{X} \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$
2: Παρήγαγε δεδομένα απο Τυπική Εκθετική κατανομή $Z \sim exp(\lambda = 1)$
3: Θέσε $\boldsymbol{\Upsilon} \leftarrow \sqrt{Z} \cdot \boldsymbol{X}$
4: Επέστρεψε $\boldsymbol{Y}$

Στη συνέχεια θα παρουσιαστούν γραφήματα της πυκνότητας της δισδιάστατης Laplace και θα συγκριθούν με την Κανονική κατανομή με σκοπό να παρουσιαστούν οι διαφορές τους. Διακρίνονται οι διαφορές στη διακύμανση των 1000 προσομοιωμένων τιμών των δύο κατανομών.

3.4.2 Γραφικές Απεικονίσεις

Σχήμα 3.15: Προσομοίωση και σύγκριση δισδιάστατων σφαιρικών Laplace και Κανονικών κατανομών

Όπως είναι φανερό η κατανομή Laplace διατηρεί πολύ περισσότερες ακραίες τιμές σε σχέση με την Κανονική κατανομή. Αυτό είναι κι ο λόγος που μπορεί να χρησιμοποιηθεί για την προσαρμογή σε δεδομένα με πολλές ακραίες παρατηρήσεις.



Σχήμα 3.16: Προσομοίωση και σύγκριση δισδιάστατων ελλειπτικών Laplace και Κανονικών κατανομών

Όπως στην περίπτωση τυποποιημένων Σφαιρικών δεδομένων, δηλαδή με $\boldsymbol{\Sigma} = \boldsymbol{I}_2$, έτσι και στην Ελλειπτική περίπτωση διακρίνονται σημαντικές διαφορές στις δύο κατανομές (βλέπε Σχήμα (3.16)). Αξιοσημείωτες είναι οι πολλές ακραίες τιμές της Laplace έναντι της Κανονικής που διαθέτει πιο συγκεντρωτικά δεδομένα.

Κεφάλαιο 4

Συσταδοποίηση με Χρήση Ελλειπτικά Συμμετρικών Κατανομών

Σε αυτό το κεφάλαιο θα διερευνηθεί ο κλάδος της Μηχανικής-Στατιστικής μάθησης που ονομάζεται Ανάλυση σε Ομάδες με Χρήση Πιθανοθεωρητικών Μοντέλων (Model Based Clustering) και συγκεκριμένα με Σφαιρικές και Ελλειπτικές κατανομές. Η μέθοδος αυτή μπορεί να θεωρηθεί και γενίκευση πιο κλασικών και ίσως πιο γνωστών μεθόδων όπως της μεθόδου των k-μέσων, της μεθόδου Ward και των τεχνικών ιεραρχικής και μή ιεραρχικής ταξινόμησης. Θα μελετηθεί αρχικά η μέθοδος αυτή θεωρητικά, παρουσιάζοντας ταυτοχρόνως και την σύνδεσή της με παρόμοιες τεχνικές μάθησης και εν τέλει θα διερευνηθεί η αποδοτικότητά της σε πραγματικά δεδομένα. Αναμφίβολα η πρακτική εφαρμογή θα αποτελέσει βοήθημα για τον αναγνώστη που θα θελήσει την βαθύτερη κατανόηση της θεωρίας.

4.1 Εισαγωγή

Οι μέθοδοι Συσταδοποίησης αποτελούν έναν σημαντικό τομέα της μηχανικής και στατιστικής μάθησης στον κλάδο της μή εποπτευόμενης μάθησης. Πρωταρχικός σκοπός της είναι η διαμέριση των δεδομένων σε κλάσεις-κατηγορίες υπο την έννοια ότι κάποια υποσύνολα παρατηρήσεων έχουν παρόμοια χαρακτηριστικά, δηλαδή διακρίνονται απο κάποια έμφυτη συνοχή. Ένα εύλογο ερώτημα είναι πως γίνεται να ανιχνευθούν τα κοινά χαρακτηριστικά των δεδομένων. Γι'αυτό το λόγο, έχουν οριστεί έννοιες απόστασης αλλά και ομοιότητας και διαφοράς οι οποίες προσπαθούν να ποσοτικοποιήσουν την ομοιότητα των παρατηρήσεων. Συγκεκριμένα σε συνεχή δεδομένα, υπάρχουν πολλές μετρικές απόστασης (Ευκλείδια, Manhattan, Minkovsky, Mahalanobis, Canberra κτλ) οι οποίες κατατάσσουν στην ίδια-ομάδα τα σημεία με μικρή απόσταση μεταξύ τους. Περιληπτικά το ίδιο ισχύει και για κατηγορικές μεταβλητές (διατάξιμες και μή) όπου εκεί χρησιμοποιείται η έννοια της ομοιότητας ως προς κάποιο χαρακτηριστικό για την ομαδοποίηση. Έλλειψη ή ύπαρξη αυτού καθορίζει την διαφορά και την ομοιότητα αντίστοιχα. Δημοφιλείς αλγόριθμοι κατάταξης των δεδομένων βάσει των προαναφερθέντων είναι οι ιεραρχικές και μή ιεραρχικές μέθοδοι συσταδοποίησης. Αναλυτικά, απο τους πιό γνωστούς και αποτελεσματικούς αλγορίθμους είναι αυτός των k-μέσων. Συνοπτικά αυτό που κάνει είναι να ξεκινά με ένα γνωστό αριθμό ομάδων και στη συνέχεια προσπαθεί επαναληπτικώς να βρεί τα βέλτιστα κέντρα των ομάδων με βασικό κριτήριο την ελαχιστοποίηση των αποστάσεων των παρατηρήσεων απο τα κέντρα. Κάθε φορά το υποψήφιο ακατάτακτο σημείο θα απορροφηθεί

στην ομάδα απο την οποία έχει την ελάχιστη απόσταση. (η απόσταση σημείου απο ομάδα ορίζεται βάσει κάποιου κριτηρίου όπως ελάχιστης, μέγιστης ή μέσης σύνδεσης [single linkage, complete linkage, average linkage]).

Η πιθανοθεωρητική ομαδοποίηση αποτελεί έναν άλλο παρόμοιο κλάδο της στατιστικής μάθησης. Σε αντίθεση με τις προηγούμενες μεθόδους που αναφέρθηκαν, που χρησιμοποιούν την έννοια της απόστασης για την κατάταξη των παρατηρήσεων σε συστάδες, στην ανάλυση σε ομάδες με χρήση πιθανοθεωρητικού μοντέλου (model based clustering) λαμβάνεται υπόψιν η κατανομή των δεδομένων. Πιο αναλυτικά, υποθέτουμε πως κάθε ομάδα ακολουθεί μία κατανομή η οποία παράγει τα δεδομένα της. Το σύνολο όλων των ομάδων-κατανομών δημιουργεί μία σύνθεση απο πολλές σ.π.π. που ονομάζεται μίξη κατανομών.

4.2 Πεπερασμένες Μίξεις Κατανομών

Ο τρόπος με τον οποίο δομείται η μέθοδος της πιθανοθεωρητικής ομαδοποίησης (model based clustering) είναι ο εξής. Κάθε παρατήρηση προέρχεται απο μόνο μία κατανομή η οποία αποτελεί συνθετικό στοιχείο της κατανομής μίξης. Η πυκνότητα που προκύπτει ως μίξη των κατανομών που την συνθέτουν βασίζεται στο θεώρημα ολικής πιθανότητας.

Θεώρημα 4.2.1 (*Θεώρημα Ολικής Πιθανότητας*)

Έστω ότι $A_1, A_2, \ldots, A_n$ είναι μια διαμέριση ενός δειγματικού χώρου S τέτοια ώστε $P(A_i) \neq 0$, $i = 1, 2, \ldots, n$. Τότε για κάθε ενδεχόμενο E έχουμε :

$$P(E) = \sum_{i=1}^{n} P(A_i) \cdot P(E|A_i) \tag{4.1}$$

Η πιθανότητα $P(E)$ αντιστοιχεί στη μίξη των κατανομών με επιμέρους τις πυκνότητες $P(E|A_i)$, $i = \{1, \ldots, k\}$ όπου k είναι ο αριθμός των ομάδων. Τα $P(A_i)$ αποτελούν τις εκ των προτέρων πιθανότητες επιλογής της κατανομής που θα παράγει την κάθε παρατήρηση. Η (4.1) μπορεί να γραφεί και στην μορφή πυκνότητας πιθανότητας $f(\boldsymbol{x}) = \sum_{i=1}^{k} \pi_i \cdot f(\boldsymbol{x}|\theta_i)$, όπου $f(\boldsymbol{x})$ η μίξη, με $f(\boldsymbol{x}|\theta_i)$ τα συνθετικά και π_i οι εκ των προτέρων πιθανότητες. Ωστόσο ο ερευνητής που δέχεται τα δεδομένα δεν γνωρίζει εκ των προτέρων απο ποιά κατανομή προήλθε η κάθε παρατήρηση. Στην περίπτωση που γνωρίζαμε τις κατηγορίες των δεδομένων, τότε η εκτίμηση των πυκνοτήτων θα ήταν εύκολη, π.χ. εάν γνωρίζαμε ότι οι κατανομές τους είναι Κανονικές με άγνωστο μέσο όρο και πίνακα συνδιακύμανσης, απλά θα εκτιμούσαμε με ε.μ.π. εκτιμητές τις παραμέτρους των κατανομών για κάθε επίπεδο. Αντίστοιχα στην μέθοδο των k-μέσων, αν γνωρίζαμε τις ομάδες τότε τα κέντρα θα βρίσκονταν εύκολα παίρνοντας απλά των διανυσματικό μέσο των παρατηρήσεων για κάθε ομάδα. Στην πραγματικότητα όμως δεν έχουμε στη διάθεσή μας την πληροφορία των ομάδων, γι'αυτό χαρακτηρίζονται και ως missing data, και αποτελούν τις χαμένες ετικέτες (labels) αλλά και τον κύριο στόχο της συσταδοποίησης ως μέθοδο εκτίμησής τους. Στην πράξη ακολουθούμε μία πολύ ισχυρή αλγοριθμική μέθοδο παρόμοια με αυτή απο άλλες μεθόδους όπως των k-μέσων. Αρχικά θέτουμε κάποιες αρχικές συνθήκες (παραμέτρους), εν συνεχεία υπολογίζουμε τις ομάδες δοθείσης της πιο πρόσφατης πληροφορίας και τέλος βάσει των νέων ομάδων επανεκτιμούμε τις αρχικές παραμέτρους. Η διαδικασία αυτή τελειώνει όταν δεν υπάρχει σημαντική αλλαγή μεταξύ των επαναλήψεων των βημάτων. Ώς παράδειγμα μπορεί να θεωρηθεί η περίπτωση των k-μέσων όπου αρχικά ορίζουμε τους μέσους των επιπέδων που θεωρούμε κοντά στους πραγματικούς, στη συνέχεια γίνεται εκτίμηση των συστάδων χρησιμοποιώντας ένα κατάλληλο κριτήριο απόστασης και στο τέλος βά-

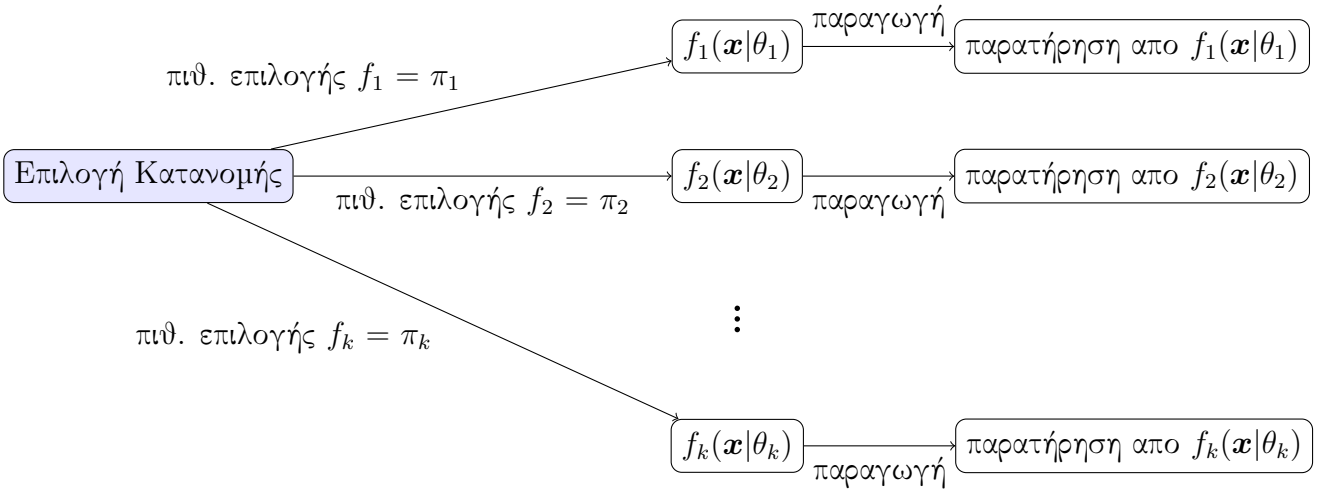
σει των πρόσφατων ομάδων επανεκτιμούνται τα κέντρα τους (επαναλαμβάνοντας τη διαδικασία έως ότου ικανοποιηθεί κάποιο κριτήριο σύγκλισης). Απο την άλλη, το model based clustering διατηρεί κι αυτό μία σχετικά όμοια λειτουργία για την εκτίμηση των ομάδων μόνο που εδώ εφόσον εφαρμόζουμε κατανομές στα δεδομένα, ο βασικός τρόπος εκτίμησης των παραμέτρων είναι μέσω της πιθανοφάνειας. Η μέθοδος αυτή χρησιμοποιεί έναν πολύ ισχυρό αλγόριθμο που ονομάζεται **Expectation Maximization - E.M. εκτίμησης μεγιστοποίησης** και περιλαμβάνει τα εξής δύο βήματα :

- Ξεκινάμε με κάποιες αρχικές εκτιμήσεις για τις παραμέτρους των κατανομών.
- **E-βήμα** : Εκτιμούμε την ομάδα που ανήκει η κάθε παρατήρηση (missing data) δοθέντος των εκτιμήσεων των παραμέτρων.
- **M-βήμα** : Χρησιμοποιούνται οι εκτιμήσεις των ομάδων (missing data) απο το **E-βήμα** με τις οποίες γίνεται επανεκτίμηση των παραμέτρων μέσω της μεθόδου μέγιστης πιθανοφάνειας.
- Επαναλαμβάνονται τα βήματα **E** και **M** μέχρι να μήν παρατηρούνται σημαντικές διαφορές στις εκτιμήσεις των παραμέτρων των κατανομών

Γενικά όταν διαθέτουμε ένα σύνολο δεδομένων αρκετά εύκολα μπορούμε να υποθέσουμε οτι προήλθε απο μία κατανομή. Στην περίπτωση του model based clustering οι παρατηρήσεις αποτελούν το παράγωγο ενός συνόλου k το πλήθος κατανομών οι οποίες συνθέτουν την γενική κατανομή μίξης με εκ των προτέρων πιθανότητες ή βάρη π_j, με $\sum_{j=1}^{k} \pi_j = 1$ (κάτι σαν μία στάθμιση των πολλών διαφορετικών κατανομών). Σχετικά με τα π_j, μία πρακτική ερμηνεία τους είναι πως καθορίζουν πόσο συχνά λαμβάνουμε δεδομένα απο την εκάστοτε κατανομή. Όπως ειπώθηκε στην εισαγωγή του Κεφαλαίου, η μέθοδος θα βασιστεί σε Ελλειπτικές μίξεις κατανομών. Συγκεκριμένα κάθε πυκνότητα που παράγει τις παρατηρήσεις (άρα συνθέτει και τις αντίστοιχες ομάδες) μπορεί να γραφεί ως $f(\boldsymbol{x}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \pi_1, \ldots, \pi_k) = \eta_p|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)]$ η οποία αποτελεί την γενική πυκνότητα Ελλειπτικών κατανομών με η_p μία σταθερά. Συνεπώς, μπορούμε να θεωρήσουμε πως οι παρατηρήσεις προήλθαν απο την σύνθεση των επιμέρους πυκνοτήτων με γενική μορφή βάσει του θεωρήματος (4.2.1) :

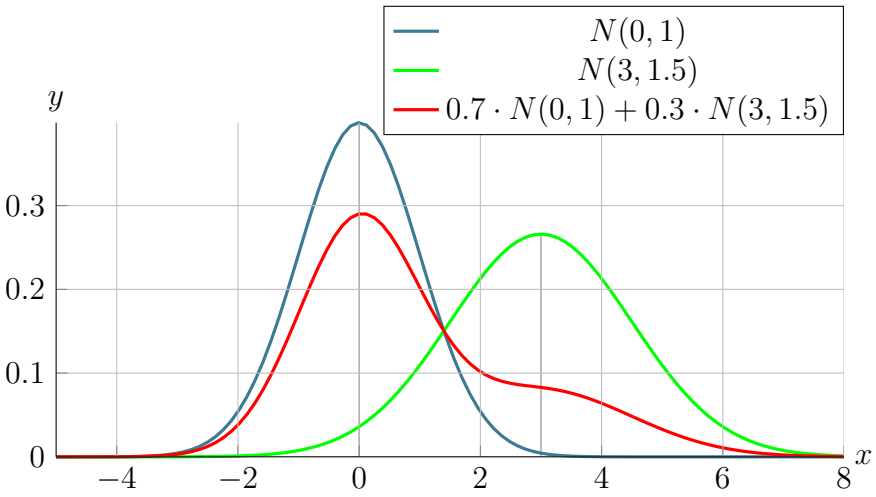
$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \pi_1, \ldots, \pi_k) = \sum_{j=1}^{k} \pi_j \cdot \left[\eta_{pi}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)] \right]. \tag{4.2}$$

Είναι λογικό κάποιος να σκεφτεί την παραγωγή των δεδομένων χωρίζοντας τη διαδικασία σε ορισμένα μέρη. Αρχικά, απο το σύνολο των κατανομών που είναι διαθέσιμες, επιλέγουμε μία απο αυτές κάθε φορά για την παραγωγή μίας παρατήρησης με πιθανότητα επιλογής π_j, $j \in \{1, \ldots, k\}$. Πλέον εφόσον έχει επιλεγεί η κατανομή, παράγεται μία παρατήρηση. Η διαδικασία αυτή (επιλογής πυκνότητας και παραγωγής παρατήρησης) επαναλαμβάνεται μέχρι να συγκέντρωθεί ένα συγκεκριμένο συνολικό πλήθος σημείων.



Πίνακας 4.1: Απεικόνιση μεθόδου παραγωγής παρατηρήσεων απο μίξη κατανομών

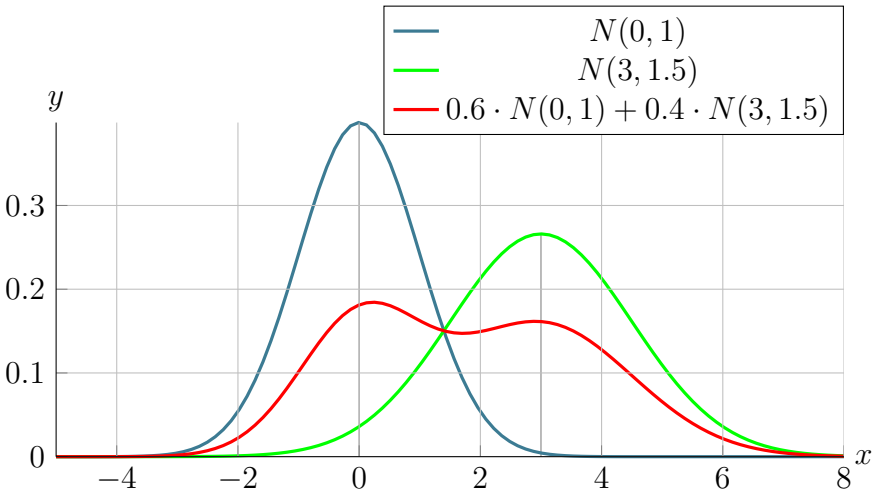
Στον πίνακα (4.1) είναι πιο ευδιάκριτος ο τρόπος δειγματοληψίας απο μίξεις πυκνοτήτων. Σημαντική είναι η λεπτομέρεια πως πρώτα επιλέγεται απο το σύστημα κάποια κατανομή και έπειτα αυτή παράγει με τη σειρά της μία παρατήρηση. Επαναλαμβανόμενο αυτό N φορές παίρνουμε ένα ολόκληρο δείγμα απο την πυκνότητα (4.2). Γενικά σε αυτό το κεφάλαιο θα ασχοληθούμε με τις μίξεις Σφαιρικών και Ελλειπτικών πυκνοτήτων. Για την κατανόηση της κατανομής στάθμισης $f(\boldsymbol{x})$ και πώς αυτή επηρεάζεται απο αυτές που την συνθέτουν, δηλαδή τις $f(\boldsymbol{x}|\theta_i)$, $i \in 1, \ldots, k$. παρατίθενται γραφικά διαγράμματα στις δύο διαστάσεις που βοηθούν στην καλύτερη επεξήγηση.



Σχήμα 4.1: Απόδοση κατανομής στάθμισης και των συνθετικών της $(\pi_1, \pi_2) = (0.7, 0.3)$

Απο το Σχήμα (4.1) διακρίνεται πως οι κατανομές (Στο παράδειγμα Κανονικές) που συνθέτουν την ενιαία μίξη είναι η $f_1(\boldsymbol{x}|\mu = 0, \sigma^2 = 1)$ με βάρος (εκ των προτέρων πιθανότητα) $\pi_1 = 0.7$ και η $f_2(\boldsymbol{x}|\mu = 3, \sigma^2 = 1.5^2)$ με βάρος $\pi_2 = 0.3$. Αξιοσημείωτα είναι τα εξής. Η πυκνότητα στάθμισης $f(\boldsymbol{x})$ επηρεάζεται απο την θέση των συνθετικών κατανομών καθώς βλέπουμε πως στα σημεία συσσώρευσης της κάθε συνιστώσας πυκνότητας η μίξη τους

(κόκκινη γραμμή) παρουσιάζει κι αυτή υψηλή συγκέντρωση. Σημαντική λεπτομέρεια είναι πως η πυκνότητα με κόκκινο χρώμα δίνει μεγαλύτερη βαρύτητα στην κατανομή που αντιστοιχεί στην υψηλότερη εκ των προτέρων πιθανότητα, που είναι η μπλέ καθώς αυτή λογαριάζεται με $\pi_1 = 0.7$ ενώ η δεύτερη κατανομή $\pi_2 = 0.3$. Λογικό μιας και υψηλές τιμές βάρους π σημαίνει πως έχουμε λάβει περισσότερες τιμές απο την κατανομή με τη μεγαλύτερη εκ των προτέρων πιθανότητα, συνεπώς συνεισφέρει κατά πολύ στην δημιουργία της μίξης.



Σχήμα 4.2: Απόδοση κατανομής στάθμισης και των συνθετικών της $(\pi_1, \pi_2) = (0.6, 0.4)$

Όπως και στο παραπάνω γράφημα, έτσι και στο Σχήμα (4.2) η θέση αλλά και η διακύμανση των πυκνοτήτων επηρεάζει την στάθμισή τους. Διακρίνονται δύο τοπικά μέγιστα που μιμούνται το αρχικό σχήμα των συνθετικών καμπυλών (μπλέ και πράσινη). Το διαφορετικό βέβαια σε αυτό το παράδειγμα είναι πως υπάρχει αλλαγή στις πιθανότητες π_1 και π_2 με $\pi_1 = 0.6, \pi_2 = 0.4$. Τώρα που τα βάρη έχουν μικρότερη διαφορά μεταξύ τους συγκριτικά με το προηγούμενο γράφημα, φαίνεται πως τα δύο κόκκινες πυκνότητες έχουν σχεδόν εξισορροπηθεί. Αυτό το γεγονός οφείλεται στο ότι πλέον η δειγματοληψία απο την πράσινη κατανομή $N(3, 1.5^2)$ είναι πιό συχνή άρα συνεισφέρει περισσότερο στο τελικό αποτέλεσμα με υψηλή βαρύτητα καθορίζοντας έτσι σε μεγάλο βαθμό το σχήμα.

Εν συνεχεία, θα ερμηνευτεί ο τρόπος με τον οποίο βρίσκουμε την πιθανοφάνεια του μοντέλου με σκοπό την εκτίμηση των παραμέτρων του. Ως γνωστόν το υπόδειγμα αποτελείται απο μία μίξη Ελλειπτικών κατανομών, συμβολικά ως $\mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ με εκ των προτέρων πιθανότητες η καθεμία. Το διάνυσμα παραμέτρων που οφείλουν να εκτιμηθούν είναι το παρακάτω :

$$\boldsymbol{\Theta} = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{k} = (\pi_1, \ldots, \pi_k, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k),$$

ενώ παράλληλα γνωρίζουμε πως η δεσμευμένη σ.π.π. μιας παρατήρησης να προέρχεται απο την $\mathcal{EC}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ είναι $f(\boldsymbol{x}|z = j) = \mathcal{EC}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g) = \eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)]$. Το z όμως δεν είναι γνωστό καθώς αποτελεί την μεταβλητή που η τιμή της αφορά την κατηγορία στην οποία ανήκει η κάθε παρατήρηση. Για να βρούμε την πιθανότητα $f(\boldsymbol{x})$ θα χρησιμοποιήσουμε την απο κοινού πυκνότητα $f(\boldsymbol{x}, z)$ και θα πάρουμε την περιθώρια ως προς z. Έχουμε :

$$f(\boldsymbol{x}) = \sum_{j=1}^{k} f(\boldsymbol{x}, z=j) = \sum_{j=1}^{k} f(z=j)\cdot f(\boldsymbol{x}|z=j) = \sum_{j=1}^{k} \pi_j \cdot \mathcal{EC}_p(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$$
$$= \sum_{j=1}^{k} \pi_j \cdot \left[\eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)]\right].$$

Όπου $f(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\Theta})$, $\boldsymbol{\Theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^k$, δηλαδή η πυκνότητα που συνδυάζει τις συνθετικές κατανομές με τα αντίστοιχα βάρη τους. Για την εκτίμηση των παραμέτρων αυτών, θα εφαρμοστεί η μέθοδος της εμπ. Η πιθανοφάνεια εκφράζεται ως :

$$\mathbf{L}(\boldsymbol{\Theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(\boldsymbol{x}_j|\boldsymbol{\Theta}),$$

συνεπώς ο λογάριθμος της πιθανοφάνειας έχει την μορφή :

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{x}) = log\,\mathbf{L}(\boldsymbol{\Theta}|\boldsymbol{x}) = log\prod_{i=1}^{n} f(\boldsymbol{x}_j|\boldsymbol{\Theta}) = \sum_{i=1}^{n} log\, f(\boldsymbol{x}_i|\boldsymbol{\Theta})$$
$$= \sum_{i=1}^{n}\{log\sum_{j=1}^{k} \pi_j \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)\}$$
$$= \sum_{i=1}^{n}\{log\sum_{j=1}^{k} \pi_j \cdot \eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)]\}. \tag{4.3}$$

Στη συνέχεια χρειάζεται να μεγιστοποιήσουμε την ποσότητα (4.3) με μερικές παραγώγους ως προς τις παραμέτρους της. Βέβαια εδώ παρουσιάζεται ένα πρόβλημα. Η πιθανοφάνεια γράφεται ως άθροισμα λογαρίθμων αθροισμάτων και η μεγιστοποίηση αυτή είναι δύσκολη. Ωστόσο, εάν γνωρίζαμε τις πραγματικές κατηγορίες των παρατηρήσεων τότε η όλη διαδικασία μεγιστοποίησης της πιθανοφάνειας θα ήταν αρκετά εύκολη. Συγκεκριμένα δεν θα χρειάζονταν το δεύτερο άθροισμα απο την σχέση (4.3), διότι αυτό αναπαριστά την έλλειψη της πληροφορίας σε ποιά κατηγορία ανήκει πραγματικά το δεδομένο. Γι'αυτό χρησιμοποιείται το άθροισμα που λογαριάζει όλες τις πιθανές εκβάσεις για το σημείο απο κάθε μία πυκνότητα των ομάδων (k το σύνολο). Βέβαια επειδή δεν γνωρίζουμε ούτε την προέλευση των δεδομένων γίνεται να οριστεί μία νέα μεταβλητή που θα ανιχνεύει το επίπεδο της κάθε $\boldsymbol{x}_i$. Έστω λοιπόν διάνυσμα $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{ik})^T$ το οποίο μαρτυρά την προέλευση της παρατήρησης $\boldsymbol{x}_i$ (σε ποιό επίπεδο ανήκει). Εάν προέρχεται απο την j ομάδα τότε $\delta_{ij} = 1$ ενώ για οποιοδήποτε άλλο j, συμβολικά $\hat{j}$, έχουμε $\delta_{i\hat{j}} = 0$, πχ για την 3η παρατήρηση προερχόμενη απο την δεύτερη ομάδα, $\boldsymbol{\delta}_3 = (0,1,0,0,\ldots,0)_{1\times k}^T$. Με αυτό τον τρόπο επιτεύχθηκε η γραφή της πιθανοφάνειας αποκλειστικά ως γινόμενο ποσοτήτων πλέον ως :

$$f(\boldsymbol{x}_i|\boldsymbol{\delta}_i, \boldsymbol{\Theta}) = \prod_{j=1}^{k}\left[\eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)]\right]^{\delta_{ij}} \tag{4.4}$$

αναλύοντάς την έτσι σε γινόμενο ως προς όλες τις ομάδες επιτρέποντας να επιβιώσει μόνο η πραγματική πυκνότητα. Παράδειγματος χάρη, έστω j_0 το επίπεδο που προήλθε το $\boldsymbol{x}_i$, τότε $\delta_{ij_0} = 1$ ενώ όλες οι άλλες ακυρώνονται με $\delta_{ij} = 0$ (οπότε στο γινόμενο αφήνει μονάδα (1)).

Συγκεκριμένα η ισότητα (4.4) θα γραφεί ως :

$$f(\boldsymbol{x}_i|\boldsymbol{\delta}_i,\boldsymbol{\Theta}) = [\mathcal{EC}_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,g)]^{\delta_{i1}}\cdot[\mathcal{EC}_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2,g)]^{\delta_{i2}}\cdots[\mathcal{EC}_p(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0},g)]^{\delta_{ij_0}}\cdots[\mathcal{EC}_p(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,g)]^{\delta_{ik}}.$$

Ωστόσο τα $\delta_{i\hat{j}_0}$ θα μηδενιστούν και θα μείνει μόνο το $\delta_{ij_0}=1$, οπότε στο τέλος θα έχουμε ως αποτέλεσμα την εξής πιθανοφάνεια $f(\boldsymbol{x}_i|\boldsymbol{\delta}_i,\boldsymbol{\Theta}) = 1\cdot 1\cdots 1\cdot\mathcal{EC}_p(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0},g)^1\cdot 1\cdots 1 = \mathcal{EC}_p(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0},g)$. Επιπλέον χρήσιμες σχέσεις είναι και οι παρακάτω :

$$f(\delta_{ij}=1|\boldsymbol{\Theta}) = \pi_j,$$

$$f(\boldsymbol{\delta}_i|\boldsymbol{\Theta}) = \prod_{j=1}^{k}\pi_j^{\delta_{ij}}.$$

Οπότε πλέον η πιθανοφάνεια αποχτά την εξής μορφή :

$$\mathcal{L}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\pi_1,\ldots,\pi_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k|\boldsymbol{x},\boldsymbol{\delta}) =$$
$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{x},\boldsymbol{\delta}) =$$
$$\sum_{i=1}^{n} log\, f(\boldsymbol{x}_i,\boldsymbol{\delta}_i,\boldsymbol{\Theta}) = \sum_{i=1}^{n} log\, f(\boldsymbol{x}_i|\boldsymbol{\delta}_i,\boldsymbol{\Theta})\cdot f(\boldsymbol{\delta}_i|\boldsymbol{\Theta}) =$$
$$\sum_{i=1}^{n} log \prod_{j=1}^{k}[f(\boldsymbol{x}_i|\boldsymbol{\delta}_{ij},\boldsymbol{\Theta})\cdot f(\boldsymbol{\delta}_{ij}|\boldsymbol{\Theta})]^{\delta_{ij}}.$$

Η τελευταία σχέση γράφεται σε μορφή αθροισμάτων λόγω του λογαρίθμου ως εξής :

$$\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{x},\boldsymbol{\delta}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\delta_{ij}\cdot log\,[f(\boldsymbol{\delta}_{ij}|\boldsymbol{\Theta})f(\boldsymbol{x}_i|\boldsymbol{\delta}_{ij},\boldsymbol{\Theta})] =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{k}\delta_{ij}\cdot[log\,\pi_j + log\,\mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,g)] =$$
$$\sum_{i=1}^{n}\sum_{j=1}^{k}\delta_{ij}\cdot\{log\,\pi_j + log(\eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}}\cdot g[(\boldsymbol{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_j)])\}. \tag{4.5}$$

4.3 Αλγόριθμος EM (Expectation Maximization)

Η μέθοδος αυτή πρόκειται για μία αρκετά προχωρημένη διαδικασία που στόχο έχει, όπως προαναφέρθηκε, την εκτίμηση του διανύσματος των παραμέτρων $\boldsymbol{\Theta}$. Στην πράξη αυτό που γίνεται είναι πως αρχικά ορίζουμε τιμές των παραμέτρων για να λάβει η διαδικασία μία αρχική πληροφορία $\boldsymbol{\Theta}^{(1)} = (\boldsymbol{\mu}_1^{(1)},\ldots,\boldsymbol{\mu}_k^{(1)},\pi_1^{(1)},\ldots,\pi_1^{(1)},\boldsymbol{\Sigma}_1^{(1)},\ldots,\boldsymbol{\Sigma}_k^{(1)})$, και απο αυτές προσπαθούμε να ανιχνεύσουμε την ομάδα που ανήκει η κάθε παρατήρηση $\boldsymbol{x}$, μέσω της μεταβλητής $\boldsymbol{\delta}$. Για αυτό το λόγο υπολογίζεται το $f(\boldsymbol{\delta}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})$, (στην πρώτη επανάληψη του αλγορίθμου) που αφορά σε ποιά κατανομή αντιστοιχεί το κάθε δεδομένο. Έπειτα παίρνουμε των μέσο όρο των πιθανοφανειών που παρουσιάζονται απο αυτή τη διαδικασία (Ε μέρος).

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) &= \mathbb{E}_{f(\boldsymbol{\delta}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})}\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{x},\boldsymbol{\delta}).\\ &= \sum_{\boldsymbol{\delta}} \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{x},\boldsymbol{\delta}) \cdot f(\boldsymbol{\delta}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})\end{aligned}$$

Εν συνεχεία, έχοντας εκτίμηση της πιθανοφάνειας, μεγιστοποιούμε (M μέρος) και έτσι παράγονται νέες εκτιμήσεις $\boldsymbol{\Theta}^{(2)}$. Γενικά σε κάθε επανάληψη :

$$\boldsymbol{\Theta}^{(n+1)} = argma\boldsymbol{x}_{\boldsymbol{\Theta}}\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})$$

Αυτή η επανάληψη συνεχίζεται μέχρι ο αλγόριθμος να συγκλίνει, δηλαδή μέχρι να μην υπάρχει κάποια σημαντική βελτίωση στις τιμές των παραμέτρων.

4.3.1 Ε μέρος

Ξεκινώντας με το κομμάτι εκτίμησης (E), για λόγους απλότητας θα χρησιμοποιηθεί η συντόμευση $\phi_{ij} = log\,\pi_j + log(\eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)])$ απο τον τύπο (4.5).

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) &= \sum_{\boldsymbol{\delta}} \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{\delta},\boldsymbol{x}) f(\boldsymbol{\delta}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \phi_{ij}\cdot\delta_{ij} f(\delta_{ij}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} \{log\,\pi_j + log(\eta_p|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)])\}\cdot w_{ij}\end{aligned}$$

με :

$$\begin{aligned}w_{ij} &= \mathbb{E}(\delta_{ij}|\boldsymbol{\Theta}^{(n)},\boldsymbol{x}) = 0\times f(\delta_{ij}=0|\boldsymbol{\Theta}^{(n)},\boldsymbol{x}) + 1\times f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)},\boldsymbol{x})\\ &= f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)},\boldsymbol{x}),\end{aligned} \tag{4.6}$$

όπου το w_{ij} συμβολίζει το βάρος ή αλλιώς την πιθανότητα κατά πόσο η τιμή $\boldsymbol{x}_i$ έχει προέλθει απο το επίπεδο j. Κάθε βήμα έχει σκοπό την τροποποίηση του $\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})$ στη νιοστή επανάληψη. Πλέον το μόνο που πρέπει να αλλάξει είναι το $f(\delta_{ij}=1|,\boldsymbol{\Theta}^{(n)},\boldsymbol{x})$ (σχέση (4.6)) για να αποδοθεί στη μορφή (ο συμβολισμός $\hat{i},u$ σημαίνει όλα τα u εκτός του i) :

$$\begin{aligned}f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)},\boldsymbol{x}) &= \frac{f(\boldsymbol{x},\delta_{ij}=1|\boldsymbol{\Theta}^{(n)})}{f(\boldsymbol{x}|\boldsymbol{\Theta}^{(n)})} = \frac{f(\boldsymbol{x}|\delta_{ij}=1,\boldsymbol{\Theta}^{(n)})\cdot f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)})}{\prod_i f(\boldsymbol{x}_i|\boldsymbol{\Theta}^{(n)})}\\ &= \frac{\prod_{l=1}^{n}[f(x_l|\delta_{ij}=1,\boldsymbol{\Theta}^{(n)})]\cdot f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)})}{f(x_i|\boldsymbol{\Theta}^{(n)})\cdot\prod_{\hat{i},u} f(x_u|\boldsymbol{\Theta}^{(n)})}\\ &= \frac{f(x_i|\delta_{ij}=1,\boldsymbol{\Theta}^{(n)})\cdot f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)})\cdot\prod_{\hat{i},u} f(x_u|\delta_{ij}=1,\boldsymbol{\Theta}^{(n)})}{\sum_{l=1}^{k}[f(x_i,\delta_{il}=1|\boldsymbol{\Theta}^{(n)})]\cdot\prod_{\hat{i},u} f(x_u|\boldsymbol{\Theta}^{(n)})}\\ &= \frac{f(x_i|\delta_{ij}=1,\boldsymbol{\Theta}^{(n)})\cdot f(\delta_{ij}=1|\boldsymbol{\Theta}^{(n)})}{\sum_{l=1}^{k}[f(x_i|\delta_{il}=1,\boldsymbol{\Theta}^{(n)})\cdot f(\delta_{il}=1|\boldsymbol{\Theta}^{(n)})]}.\end{aligned} \tag{4.7}$$

Γνωρίζουμε ότι $f(\boldsymbol{x}_i, \delta_{ij} = 1|\boldsymbol{\Theta}^{(n)}) = f(\delta_{ij} = 1|\boldsymbol{\Theta}^{(n)}) \cdot f(\boldsymbol{x}_i|\delta_{ij} = 1, \boldsymbol{\Theta}^{(n)}) = \pi_j \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ και το $\sum_l f(\boldsymbol{x}_i, \delta_{il} = 1|\boldsymbol{\Theta}^{(n)}) = \sum_l [f(\delta_{il} = 1, \boldsymbol{\Theta}^{(n)}) \cdot f(\boldsymbol{x}_i|\delta_{il} = 1, \boldsymbol{\Theta}^{(n)})] = \sum_{l=1}^{k} \pi_l \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, g)$, τα οποία είναι γνωστά στην κάθε επανάληψη, οπότε :

$$f(\delta_{ij} = 1|\boldsymbol{\Theta}^{(n)}, \boldsymbol{x}) = \frac{\pi_j \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)}{\sum_{l=1}^{k} \pi_l \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, g)}.$$

4.3.2 M μέρος

Ανακαλώντας την εκτίμηση πιθανοφάνειας στο n-βήμα του αλγορίθμου :

$$\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\{log\,\pi_j + log(\eta_{pj}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \cdot g[(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)])\} \cdot w_{ij},$$

οι ποσότητες που περιέχονται είναι όλες γνωστές εφόσον $w_{ij} = \frac{\pi_j \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)}{\sum_{l=1}^{k} \pi_l \cdot \mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, g)}$ και τα $\boldsymbol{x}_i, \boldsymbol{\Sigma}_j, \pi_j$ είναι εκτιμημένα στην νιοστή επανάληψη. Συνεπώς είναι γνωστή η εκτίμηση της πιθανοφάνειας στο n-βήμα ως $\mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})$. Για να πάρουμε τις νέες εκτιμήσεις των παραμέτρων για να προχωρήσουμε στην $n+1$ επανάληψη θα χρησιμοποιήθούν οι ε.μ.π. που βρίσκονται με τη λύση του συστήματος :

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})}{\partial \boldsymbol{\mu}_j} &= \frac{\partial}{\partial \boldsymbol{\mu}_j}\sum_{i=1}^{N} w_{ij} \cdot log\,\mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g) = 0 \\
\frac{\partial \mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})}{\partial \boldsymbol{\Sigma}_j} &= \frac{\partial}{\partial \boldsymbol{\Sigma}_j}\sum_{i=1}^{N} w_{ij} \cdot log\,\mathcal{EC}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g) = 0 \\
\frac{\partial \mathcal{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(n)})}{\partial \pi_j} &= \frac{\partial}{\partial \pi_j}\sum_{i=1}^{N} w_{ij} \cdot log\,\pi_j = 0,
\end{aligned}$$

υπό τον περιορισμό $\sum_{j=1}^{k} \pi_j = 1$. Επαναλαμβανεται η διαδικασία μέχρι να ικανοποιηθεί κάποιο κριτήριο σύγκλισης όπως για παράδειγμα μέχρις ότου οι διαφορές των πρόσφατων με των προηγούμενων εκτιμήσεων να γίνουν μικρότερες απο ένα φράγμα, π.χ. φράγμα $= 10^{-5}$.

4.4 Αριθμητικά Αποτελέσματα Συσταδοποίησης με Χρήση Ελλειπτικής Κατανομής

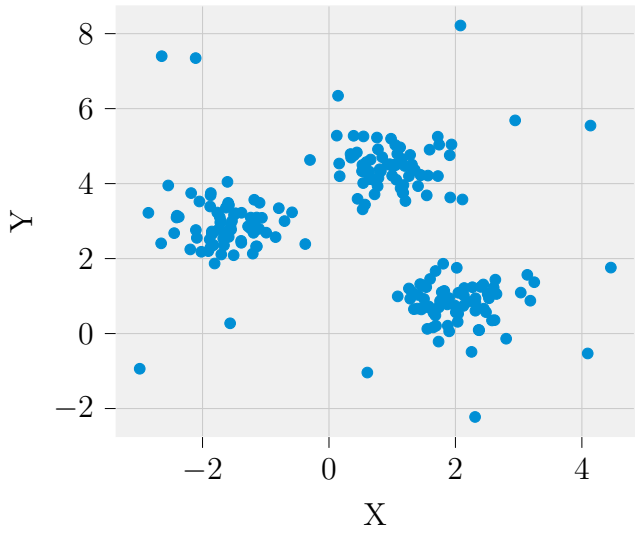
Έχοντας παρουσιάσει την θεωρία που θεμελιώνει την διαδικασία, θα προχωρήσουμε πλέον σε κάποια παραδείγματα συσταδοποίησης σε δύο διαστάσεις για να γίνει πιο κατανοητός ο τρόπος εφαρμογής της μεθόδου. Η ιδέα του μηχανισμού της αβίαστα δύναται να γενικευτεί και στις περισσότερες διαστάσεις καθώς παραμένει ίδια η τεχνική αλλά αυξάνεται μόνο η πολυπλοκότητα.

Αρχικά θα προσομοιωθούν δεδομένα απο ένα σύνολο τριών Ελλειπτικών κατανομών δύο διαστάσεων με διαφορετικούς μέσους αλλά και πίνακες Σ. Βασικό χαρακτηριστικό τους θα είναι η ύπαρξη αρκετών ακραίων τιμών, δηλαδή παρατηρήσεις που θα απέχουν πολύ απο τους διανυσματικούς μέσους των πυκνοτήτων. Αυτές οι πυκνότητες αποτελούν και τις ομάδες των δεδομένων που στην πράξη είναι άγνωστες. Στο Σχήμα (4.3) δίνεται η γραφική τους απεικόνιση.

Αλγόριθμος 4.4.1: Δισδιάστατες Σφαιρικες Ελλειπτικές Κατανομές

```
#Απαραίτητες Βιβλιοθήκες
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
import sklearn.datasets
import tikzplotlib

sns.set()

# Προσομοίωση Δεδομένων με ακραίες τιμές
X, y_true = sklearn.datasets.make_blobs(n_samples=200, centers=3,
                        cluster_std=0.50, random_state=False)

x_noise = np.random.uniform(low=np.min(X),
                            high=np.max(X),
                           size=(20,2))
X = np.vstack([X, x_noise])
plt.scatter(X[:,0], X[:,1], s=10)
plt.title("Unclustered data")
plt.show()
```

Προσομοιωμένα Δεδομένα με Ακραίες Τιμές



Σχήμα 4.3: Δεδομένα προερχόμενα απο σφαιρικές και ελλειπτικές κατανομές με άγνωστες τις ετικέτες (labels) των ομάδων

Η μέθοδος που αποτελεί την εφαρμογή της πιθανοθεωρητικής συσταδοποίησης θα μας δώσει τις σημαντικές πληροφορίες σχετικά με την προέλευση των σημείων αλλά και την κατανομή που τα παρήγαγε (τόσο την καθεμιά αλλα και την ένωσή τους). Συγκεκριμένα οι κατανομές που θα

προσαρμοστούν είναι δισδιάστατες Student-t, κυρίως γιατί αποτελούν πυκνότητες πιθανοτήτων που εφαρμόζονται όταν διαθέτουμε δεδομένα με αρκετές ακραίες τιμές. Ο αλγόριθμος EM θα προσπαθήσει (βάσει των μεθόδων που αναλύθηκαν) να εκτιμήσει τις παραμέτρους των κατανομών της καθεμιά συστάδας αλλα προφανώς και της μίξης αυτών. Αναλυτικότερα το σύνολο των παραμέτρων που θα εκτιμηθούν είναι το :

$$\boldsymbol{\Theta} = \{\nu_1, \nu_2, \ldots, \nu_k, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_k\}.$$

Με Student-t σ.π.π είναι η εξής :

$$f_i(\boldsymbol{x}; \nu_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{\Gamma[\frac{\nu_i+p}{2}]}{\Gamma(\frac{\nu_i}{2})\nu_i^{\frac{p}{2}}\pi^{\frac{p}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \cdot \left[1 + \frac{1}{\nu_i}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right]^{-\frac{\nu_i+p}{2}}$$

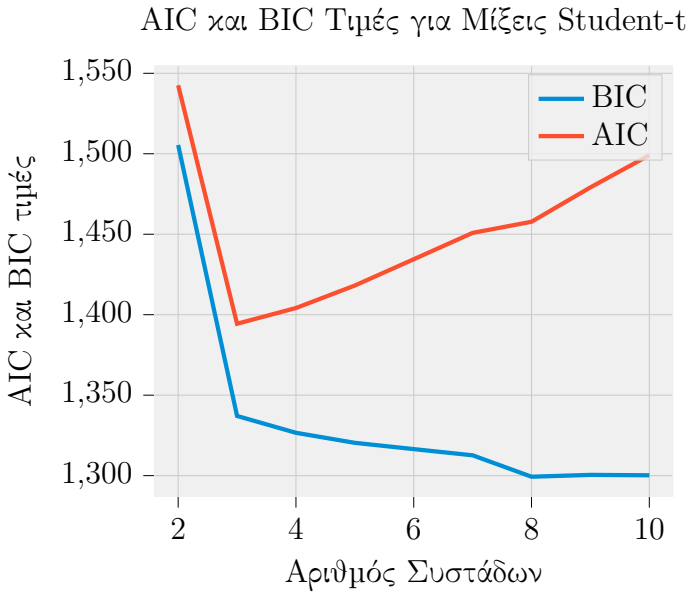
και μίξη αυτών $f(\boldsymbol{x}) = \sum_i \pi_i \cdot f_i(\boldsymbol{x}; \nu_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Ωστόσο για να λειτουργήσει ο αλγόριθμος, πρέπει αρχικά να δοθεί ο πραγματικός αριθμός των συστάδων. Στην πραγματικότητα αυτό είναι άγνωστο, ειδικά σε μεγαλύτερες διαστάσεις όπου χάνεται πλέον η γεωμετρική εποπτεία. Στην περίπτωση των k-μέσων δοκιμάζουμε διάφορες αρχικές τιμές ομάδων στον αλγόριθμο και επιλέγουμε αυτόν που οδηγεί στην βελτιστοποίηση ενός κριτηρίου επιλογής όπως οι αποστάσεις εντός και εκτός των συστάδων. Ωστόσο στην περίπτωση του Model Based Clustering, εφόσον εφαρμόζουμε κατανομές στις ομάδες μπορούμε να χρησιμοποιήσουμε πολύ πιο ισχυρές μεθόδους για την εύρεση του βέλτιστου πλήθους ομάδων όπως τα γνωστά στατιστικά κριτήρια **AIC**, **BIC**. Μεταξύ δύο η περισσότερων μοντέλων, την καλύτερη προσαρμογή θεωρείται οτι έχει αυτό με το χαμηλότερο **AIC** ή **BIC**. Γενικά το **BIC** είναι πιό αυστηρό καθώς δίνει μεγαλύτερη ποινή στις περισσότερες ομάδες. Στον αλγόριθμο 4.4.2 θα χρησιμοποιηθεί το πακέτο της Python studenttmixture[1] που θα μας δώσει το Σχήμα (4.4) στο οποίο θα δούμε το διάγραμμα των καταγεγραμμένων τιμών αυτών των κριτηρίων για διάφορες τιμές των ομάδων που έτρεξε ο αλγόριθμος.

Αλγόριθμος 4.4.2: Γραφικό AIC και BIC Τιμών ως προς Αριθμό Συστάδων

```
# Επιλογή βέλτιστου πλήθους Συστάδων βάσει AIC, BIC. Για την εφαρμογή
↪ Student-t μίξεων θα χρησιμοποιηθεί το πακέτο της Python :
↪ studenttmixture

import studenttmixture
from studenttmixture import EMStudentMixture
aics, bics = [], []
for ncomp in range(2,10):
    mix_model = EMStudentMixture(n_components=ncomp, fixed_df=True,
    ↪ df=1.0)
    mix_model.fit(X)
    aics.append(mix_model.aic(X))
    bics.append(mix_model.bic(X))

plt.plot(np.arange(2,10), [aic for aic in aics], label='BIC')
plt.plot(np.arange(2,10), [bic for bic in bics], label='AIC')
```

[1]Για το πακέτο της Python βλέπε https://pypi.org/project/studenttmixture/

```
plt.legend(loc='best')
plt.xlabel('Αριθμός Συστάδων')
plt.xlabel("Αριθμός Συστάδων")
plt.ylabel("AIC και BIC τιμές")
plt.show()
```



Σχήμα 4.4: AIC και BIC τιμές μίξης κατανομών Student-t

Παρατηρείται πως το AIC έδωσε ελάχιστο όταν ο αριθμός των συστάδων είναι τρείς. Απο την άλλη το BIC δεν φαίνεται να δίνει κάποια βέλτιστη τιμή σε αυτό το εύρος αριθμού συστάδων οπότε συνεχίζουμε την ανάλυση με την τιμή του AIC που είναι τρείς συστάδες.

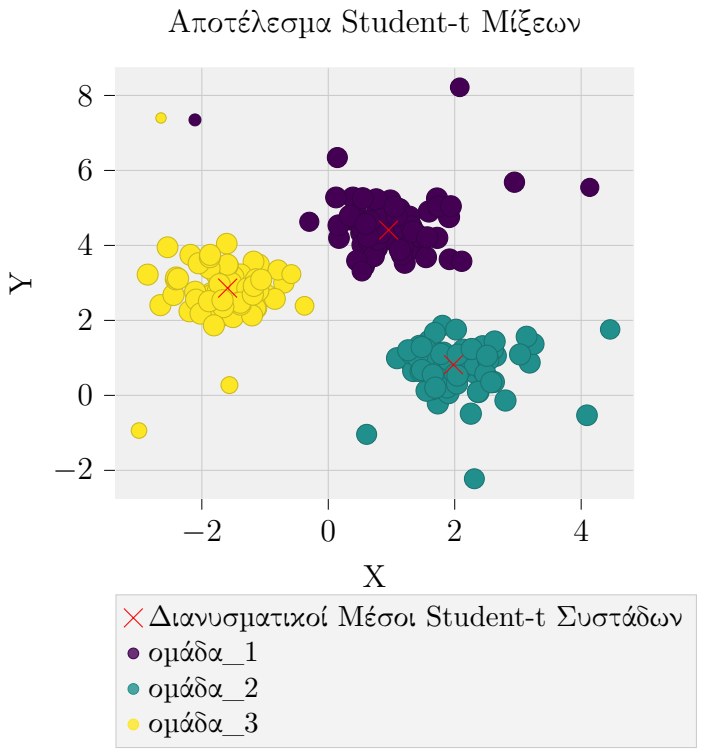
Στη συνέχεια θα εφαρμοστεί η μέθοδος **EMmixtureModels**[2] με των αριθμό συστάδων που βρέθηκε απο το παραπάνω διάγραμμα (τρείς το πλήθος) ο οποίος θα ανιχνεύσει τα επίπεδα των δεδομένων χρησιμοποιώντας τον αλγόριθμο **EM**. Αναλυτικότερα αυτό που πράττει είναι να εκτιμήσει τις παραμέτρους $\{\pi_i, \nu_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ για $i = \{1, 2, 3\}$ των κατανομών Student-t ή αλλιώς των ομάδων μιας και οι ομάδες είναι πλέον πυκνότητες πιθανοτήτων.

Αλγόριθμος 4.4.3: Εφαρμογή Αλγορίθμου EM Συσταδοποίησης με Χρήση Πιθανοθεωρητικού Μοντέλου Μίξεων Student-t

```
# Προσαρμογή EMStudentMixture χρησιμοποιώντας 3 συστάδες

stm = EMStudentMixture(n_components=3, n_init=5, fixed_df=False, df=1,
 ↪  random_state=11)
stm.fit(X)
labels_stm = stm.predict(X)
probs_stm = stm.predict_proba(X)
size_stm = 50 * probs_stm.max(1) ** 2

```

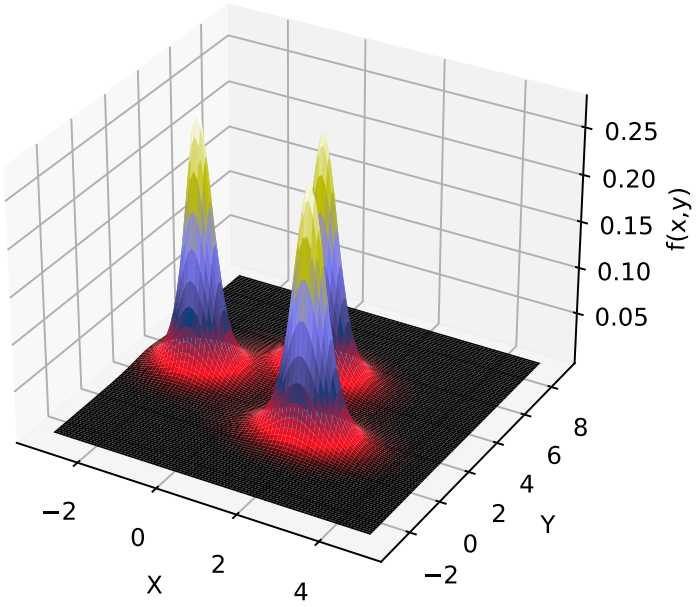
[2]Βλέπε ιστοσελίδα https://github.com/jlparkI/mix$_T$

```
import cmasher as cmr
color = cmr.take_cmap_colors('viridis',3,return_fmt='hex')
sustada = ['ομάδα_1','ομάδα_2','ομάδα_3']
plt.scatter(X[:,0], X[:,1], s= size_stm, c= labels_stm, cmap =
↪  "viridis")
plt.show()
stm_clusters = stm.location
plt.scatter(stm_clusters[:,0], stm_clusters[:,1], color="purple",
↪  marker="x",
                label="Διανυσματικοί Μέσοι\nStudent-t Συστάδων", s=100)

[plt.scatter([],[],label=sustada[i], c = color[i]) for i in range(3)]
plt.legend()
plt.title("Αποτέλεσμα Student-t Μίξεων")
plt.xlabel("X")
plt.ylabel("Y")
plt.show()
```



Σχήμα 4.5: Αποτέλεσμα μίξεων κατανομών Student-t

Απο το παραπάνω γράφημα που αποτελεί και αποτέλεσμα της μεθόδου **EMmixtureModels**, φαίνεται πως δημιουργήθηκαν τρείς συστάδες (απεικόνιση με διαφορετικά χρώματα) οι οποίες αποτελούν ταυτοχρόνως και τις κατανομές Student-t. Εναλλακτική αναπαράσταση παρουσιάζεται στο Σχήμα (4.6) όπου εκεί διακρίνεται η μίξη των κατανομών Student-t.



Σχήμα 4.6: Γράφημα της μίξης των κατανομών Student-t

Συγκεκριμένα, για τις εκτιμήσεις των παραμέτρων έχουμε :

$$f_{\text{ομάδα\_1}}(x,y) = t_{\nu=3}\Big(\boldsymbol{\mu}_1 = (0.954342274.40935493)^T, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.19825526 & -0.01758858 \\ -0.01758858 & 0.20660431 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_2}}(x,y) = t_{\nu=3}\Big(\boldsymbol{\mu}_2 = (-1.591608192.86127231)^T, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.2191887 & 0.01129342 \\ 0.01129342 & 0.17755 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_3}}(x,y) = t_{\nu=3}\Big(\boldsymbol{\mu}_3 = (-1.59519087, 2.88384863)^T, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.16106262 & 0.00297376 \\ 0.00297376 & 0.21787599 \end{bmatrix}\Big).$$

Ας εξηγήσουμε τώρα την διαδικασία του αλγορίθμου αλλά και κάποια βασικά στοιχεία της. Είπαμε προηγουμένως πως για την εκκίνηση του αλγορίθμου πρέπει να δοθούν τυχαία κάποιες αρχικές τιμές των παραμέτρων όπως θα κάναμε και στους k-μέσους. Συγκεκριμένα το πακέτο `EMmixtureModels` χρησιμοποιεί μέθοδο k-μέσων για μία αρχική εκτίμηση των παραμέτρων, που θα χρησιμοποιηθούν ως εναρκτήριες τιμές για την διαδικασία EM.

```
>>> stm.init_type
'kmeans'
```

`kmeans` : υποδηλώνοντας αρχική εκτίμηση παραμέτρων μέσω αλγορίθμου k-μέσων. Κάποια σημαντικά αριθμητικά αποτελέσματα του EM αλγορίθμου παραδίδονται παρακάτω :

```
>>> stm.converged_
True
```

υποδεικνύοντας ότι η διαδικασία έχει συγκλίνει.

```
>>> stm.mix_weights_
array([0.33493489, 0.3368138 , 0.32825131])
```

Αυτό το διάνυσμα αποτελεί τις εκτιμώμενες εκ των προτέρων πιθανότητες της διαδικασίας για κάθε μία κατανομή $\{\pi_j\}_{j=1}^3$. Όλες είναι σχεδόν $\frac{1}{3}$ διότι κάθε ομάδα περιέχει περίπου 100 παρατηρήσεις (όλες μαζί είναι 300). Αναλυτικότερα, είναι η πιθανότητα να πάρουμε παρατήρηση απο κάποιο συγκεκριμένο επίπεδο, που είναι περίπου $\frac{100}{300}$.

```
>>> stm.location_
array([[ 0.95434227,  4.40935493],
       [ 1.97984696,  0.81817873],
       [-1.59160819,  2.86127231]])
```

Ο πίνακας `stm.location` έχει αποθηκεύσει τους εκτιμώμενους διανυσματικούς μέσους απο τις τρείς κανονικές κατανομές. Αναλυτικά έχουμε :

ομάδες	$\hat{\mu}$
ομάδα 1	$(0.95434227, 4.40935493)^T$
ομάδα 2	$(1.97984696, 0.81817873)^T$
ομάδα 3	$(-1.59160819, 2.86127231)^T$

Πίνακας 4.2: Εκτιμήσεις δειγματικών μέσων απο EMmixtureModels

```
>>> stm.scale_
array([[[ 0.19825526,  0.2191887 ,  0.16106262],
        [-0.01758858,  0.01129342,  0.00297376]],

       [[-0.01758858,  0.01129342,  0.00297376],
        [ 0.20660431,  0.17755   ,  0.21787599]]])
```

Πίνακες Σ		X	Y
Ομάδα 1	X	0.19825526	-0.01758858
	Y	-0.01758858	0.20660431
Ομάδα 2	X	0.2191887	0.01129342
	Y	0.01129342	0.17755
Ομάδα 3	X	0.16106262	0.00297376
	Y	0.00297376	0.21787599

Πίνακας 4.3: Αποτελέσματα των πινάκων Σ απο EMmixtureModels

Μέχρι στιγμής, όσον αφορά τις εκτιμήσεις των παραμέτρων φάνηκε αρκετά ικανοποιητική εκτίμηση των πραγματικών τιμών τόσο ως προς την θέση (διανυσματικοί μέσοι) αλλά και προς την μεταβλητότητα (πίνακες scale Σ) των ομάδων-κατανομών. Κλείνοντας το αριθμητικό κομμάτι, θα παρατεθούν και κάποια αποτελέσματα σχετικά με την εκτίμηση των ετικετών των δεδομένων, δηλαδή την προσπάθεια εύρεσης της ομάδας προέλευσης της κάθε παρατήρησης.

```
>>> stm.n_components
3
```

Ο αριθμός των επιπέδων προφανώς όπως παρουσιάστηκε απο τα παραπάνω αποτελέσματα είναι τρία.

```
>>> stm.predict_proba(X[:5,:])
array([[2.47595427e-03, 3.05340439e-04, 9.97218705e-01],
       [9.98617141e-01, 2.78017101e-04, 1.10484149e-03],
       [9.99049375e-01, 2.57837986e-04, 6.92787329e-04],
       [2.00188719e-03, 2.78868051e-04, 9.97719245e-01],
       [9.98215325e-01, 4.39712865e-04, 1.34496244e-03]])
```

Πιθανοτητες Κατατάξεων	ομάδα 1	ομάδα 2	ομάδα 3
0	2.47595427e-03	3.05340439e-04	9.97218705e-01
1	9.98617141e-01	2.78017101e-04	1.10484149e-03
2	9.99049375e-01	2.57837986e-04	6.92787329e-04
3	2.00188719e-03	2.78868051e-04	9.97719245e-01
4	9.98215325e-01	4.39712865e-04	1.34496244e-03

Πίνακας 4.4: Πιθανότητες κατατάξεως των πρώτων πέντε δεδομένων ανα επίπεδο για Student-t μίξεις

Οι πιθανότητες κατάταξης των πρώτων πέντε παρατηρήσεων υπολογισμένες για κάθε ομάδα. Αυτές υπολογίζονται βάσει του τύπου (4.7) απο την θεωρία που αποδίδει το πηλίκο της πιθανοφάνειας του επιπέδου ως προς όλα τα άλλα, δηλαδή μία στάθμιση των πιθανοφανειών. Όσο πιο μεγάλη είναι η πιθανοφάνεια της παρατήρησης σε μία κατανομή, τόσο μεγαλύτερη πιθανότητα να προήλθε απο αυτήν.

```
>>> stm.predict(X[:5,:])
array([2, 0, 0, 2, 0])
```

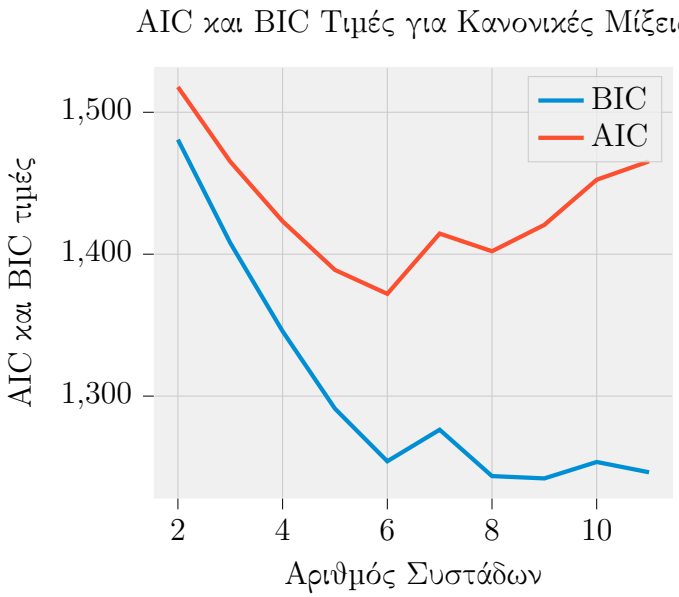
Εν τέλει, απο το διάνυσμα `stm.predict(X[:5,:])` διακρίνεται η πρόβλεψη των ομάδων για τα πρώτα πέντε σημεία. Το συμπέρασμα είναι πως η πρώτη και η τέταρτη παρατήρηση προέρχονται απο την ομάδα 3, ενώ η δεύτερη, τρίτη και πέμπτη απο την ομάδα 1 (οι κωδικοί των ομάδων είναι 0,1,2 που αντιστοιχούν σε ομάδες 1,2,3 αντίστοιχα). Αυτό απορρέει άμεσα απο τις πιθανότητες κατάταξης απο τον προηγούμενο πίνακα. Φαίνεται δηλαδή πως για το πρώτο και το τέταρτο σημείο η πιθανότητα να προέρχονται απο την ομάδα 3 είναι σχεδόν 1, αντίστοιχα το ίδιο συμβαίνει για τις παρατηρήσεις 2,3 και 5 που έχουν κατηγοριοποιηθεί στην ομάδα 1.

Ένα ερώτημα που θα μπορούσε να θέσει κάποιος είναι ποιά θα ήταν η εικόνα και η ποιότητα ομαδοποίησης αν χρησιμοποιούνταν μίξεις Κανονικών κατανομών αντί για Student-t. Γνωρίζουμε απο την γραφική αναπαράσταση που έγινε στην αρχή πως λόγω των ακραίων τιμών σε κάποια δεδομένα δεν θα ήταν καλή η επιλογή Κανονικών πυκνοτήτων πιθανοτήτων ως κατανομές των συστάδων. Ωστόσο για να δειχθεί η ικανότητα της Student-t να προσαρμόζεται σε σύνολα δεδομένων όπου η Κανονική αδυνατεί, παρακάτω παρατίθεται μία σύγκριση των δύο διαδικασιών εφαρμόζοντας την ίδια μέθοδο αλλά αυτή τη φορά για Κανονικές μίξεις με την ονομασία Gaussian Mixture Models GMM[3] :

Αλγόριθμος 4.4.4: Βέλτιστο Πλήθος Συστάδων Κανονικών Μίξεων GMM Βάσει AIC και BIC

```
# Μοντέλο GMM (Model Based Gaussian Clustering)
# Επιλογή βέλτιστου πλήθους Συστάδων βάσει AIC, BIC

from sklearn.mixture import GaussianMixture as GMM

aics, bics = [], []
for ncomp in range(2,10):
    mix_model = GMM(n_components=ncomp).fit(X)
    mix_model.fit(X)
    aics.append(mix_model.aic(X))
    bics.append(mix_model.bic(X))

plt.plot(np.arange(2,10), [aic for aic in aics], label='BIC')
plt.plot(np.arange(2,10), [bic for bic in bics], label='AIC')
plt.legend(loc='best')
plt.xlabel('Αριθμός Συστάδων')
plt.xlabel("Αριθμός Συστάδων")
plt.ylabel("AIC και BIC τιμές")
plt.show()
```

[3]Για το πακέτο Κανονικών μίξεων, βλέπε https://scikit-learn.org/stable/modules/mixture.html



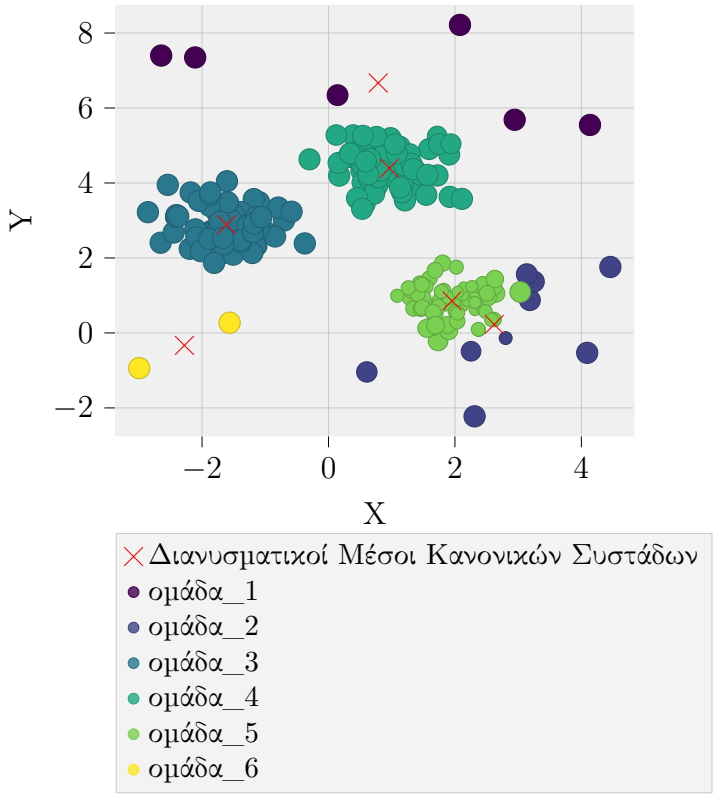
Σχήμα 4.7: AIC και BIC τιμές μίξης Κανονικών κατανομών

Για προσαρμογή Κανονικών μίξεων οι BIC τιμές δεν έδωσαν κάποια σαφή εικόνα για το βέλτιστο πλήθος των ομάδων. Απο την άλλη το κριτήριο AIC υποδεικνύει βέλτιστο πλήθος ομάδων έξι γι'αυτό και θα προσαρμοστούν έξι Κανονικές μίξεις.

Αλγόριθμος 4.4.5: Εφαρμογή Κανονικών Μίξεων GMM

```
# Μοντέλο GMM (Model Based Gaussian Clustering)
import cmasher as cmr
color_gmm = cmr.take_cmap_colors('viridis',5,return_fmt='hex')
from sklearn.mixture import GaussianMixture as GMM
gmm = GMM(n_components=5).fit(X)
labels_gmm = gmm.predict(X)

probs_gmm = gmm.predict_proba(X)

sustada_gmm = ['ομάδα_1','ομάδα_2','ομάδα_3','ομάδα_4','ομάδα_5']
size_gmm = 50 * probs_gmm.max(1) ** 2
```

Αποτέλεσμα μεθόδου GMM (Gaussian Mixture Model)



Σχήμα 4.8: Αποτέλεσμα μίξεων Κανονικών κατανομών

Απο το Σχήμα (4.7) που αποτελεί και αποτέλεσμα της μεθόδου **GaussianMixtureModels**[4], δημιουργήθηκαν 6 συστάδες (απεικόνιση με διαφορετικά χρώματα) που αποτελούν τις Κανονικές κατανομές. Συγκεκριμένα έχουμε :

$$f_{\text{ομάδα\_1}}(x,y) = N_2\Big(\boldsymbol{\mu}_1 = (0.7863992, 6.66834997)^T, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 6.07517962 & -1.33024060 \\ -1.33024060 & 1.04766014 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_2}}(x,y) = N_2\Big(\boldsymbol{\mu}_2 = (2.62165319, 0.22755444)^T, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.02210496 & 0.495592654 \\ 0.495592654 & 1.29092211 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_3}}(x,y) = N_2\Big(\boldsymbol{\mu}_3 = (-1.61333044, 2.88765908)^T, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.240530106 & -0.0166543148 \\ -0.0166543148 & 0.235752530 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_4}}(x,y) = N_2\Big(\boldsymbol{\mu}_4 = (0.96391098, 4.39392407)^T, \boldsymbol{\Sigma}_4 = \begin{bmatrix} 0.239940390 & -0.0270686602 \\ -0.0270686602 & 0.239548111 \end{bmatrix}\Big)$$

[4]Για τον κώδικα, βλέπε https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture

$$f_{\text{ομάδα\_5}}(x,y) = N_2\Big(\boldsymbol{\mu}_5 = (1.95285377, 0.85588405)^T, \boldsymbol{\Sigma}_5 = \begin{bmatrix} 0.206880088 & 0.00323522226 \\ 0.00323522226 & 0.178111628 \end{bmatrix}\Big)$$

$$f_{\text{ομάδα\_6}}(x,y) = N_2\Big(\boldsymbol{\mu}_6 = (-2.2784369, -0.33230685)^T, \boldsymbol{\Sigma}_6 = \begin{bmatrix} 0.511224823 & 0.433323307 \\ 0.433323307 & 0.367294307 \end{bmatrix}\Big)$$

Απο ότι φαίνεται, οι Κανονικές μίξεις δημιούργησαν 3 επιπλέον ομάδες αποκλειστικά για τις ακραίες τιμές, συγκριτικά με τις μίξεις κατανομών Student-t. Στη συνέχεια παρατίθονται κάποια αριθμητικά αποτελέσματα των Κανονικών μίξεων.

```
>>> gmm.init_type
'kmeans'
```

`kmeans` : υποδηλώνει αρχική εκτίμηση παραμέτρων μέσω αλγορίθμου k-μέσων. Κάποια σημαντικά αριθμητικά αποτελέσματα του EM αλγορίθμου παραδίδονται παρακάτω :

```
>>> gmm.converged_
True

>>> gmm.weights_
array([0.02936621, 0.05620409, 0.3162742 , 0.31015096, 0.27870221,
       0.00930232])
```

Αυτό το διάνυσμα αποτελεί τις εκτιμώμενες εκ των προτέρων πιθανότητες της διαδικασίας για κάθε μία κατανομή $\{\pi_j\}_{j=1}^3$. Εδώ βλέπουμε διαφορές στις πιθανότητες για την κάθε μία ομάδα. Αναμενόμενο μιας και διακρίνονται ομάδες με μεγάλες διαφορές στο πλήθος των σημείων που κατέχουν.

```
>>> gmm.means_
array([[ 0.7863992 ,  6.66834997],
       [ 2.62165319,  0.22755444],
       [-1.61333044,  2.88765908],
       [ 0.96391098,  4.39392407],
       [ 1.95285377,  0.85588405],
       [-2.2784369 , -0.33230685]])
```

Ο πίνακας `gmm.means_` έχει αποθηκεύσει τους εκτιμώμενους διανυσματικούς μέσους απο τις 6 κανονικές κατανομές. Αναλυτικά έχουμε :

ομάδες	$\hat{\mu}$
ομάδα 1	$(0.7863992, 6.66834997)^T$
ομάδα 2	$(2.62165319, 0.22755444)^T$
ομάδα 3	$(-1.61333044, 2.88765908)^T$
ομάδα 4	$(0.96391098, 4.39392407)^T$
ομάδα 5	$(1.95285377, 0.85588405)^T$
ομάδα 6	$(-2.2784369, -0.33230685)^T$

Πίνακας 4.5: Εκτιμήσεις δειγματικών μέσων απο GaussianMixtureModels

```
>>> gmm.covariances_
array([[[ 6.07517962e+00, -1.33024060e+00],
        [-1.33024060e+00,  1.04766014e+00]],

       [[ 1.02210496e+00,  4.95592654e-01],
        [ 4.95592654e-01,  1.29092211e+00]],

       [[ 2.40530106e-01, -1.66543148e-02],
        [-1.66543148e-02,  2.35752530e-01]],

       [[ 2.39940390e-01, -2.70686602e-02],
        [-2.70686602e-02,  2.39548111e-01]],

       [[ 2.06880088e-01,  3.23522226e-03],
        [ 3.23522226e-03,  1.78111628e-01]],

       [[ 5.11224823e-01,  4.33323307e-01],
        [ 4.33323307e-01,  3.67294307e-01]]])
```

Πίνακες $\boldsymbol{\Sigma}$		X	Y
Ομάδα 1	X	6.07517962e+00	-1.33024060e+00
	Y	-1.33024060e+00	1.04766014e+00
Ομάδα 2	X	1.02210496e+00	4.95592654e-01
	Y	4.95592654e-01	1.29092211e+00
Ομάδα 3	X	2.40530106e-01	-1.66543148e-02
	Y	-1.66543148e-02	2.35752530e-01
Ομάδα 4	X	2.39940390e-01	-2.70686602e-02
	Y	-2.70686602e-02	2.39548111e-01
Ομάδα 5	X	2.06880088e-01	3.23522226e-03
	Y	3.23522226e-03	1.78111628e-01
Ομάδα 6	X	5.11224823e-01	4.33323307e-01
	Y	4.33323307e-01	3.67294307e-01

Πίνακας 4.6: Πίνακες $\boldsymbol{\Sigma}$ απο GaussianMixtureModels

```
>>> gmm.n_components
6
```

Ο αριθμός των ομάδων προφανώς όπως παρουσιάστηκε απο τα παραπάνω αποτελέσματα είναι τρία.

```
>>> gmm.predict_proba(X[:5,:])
array([[5.96103493e-07, 2.51845197e-11, 9.99999373e-01, 3.03781683e-08,
        6.08979181e-22, 0.00000000e+00],
       [3.54981592e-04, 4.98701240e-08, 1.62385833e-08, 9.99644952e-01,
        2.31573537e-17, 0.00000000e+00],
       [5.63582708e-04, 7.85832906e-08, 4.37093931e-10, 9.99436338e-01,
        9.13318932e-18, 0.00000000e+00],
       [2.86347033e-07, 1.12883200e-10, 9.99999657e-01, 5.67872024e-08,
        5.42337749e-20, 0.00000000e+00],
       [3.98803719e-03, 2.06156520e-08, 6.54617786e-12, 9.96011942e-01,
        8.52199694e-22, 0.00000000e+00]])
```

	ομάδα 1	ομάδα 2	ομάδα 3	ομάδα 4	ομάδα 5	ομάδα 6
0	5.9e-07	2.5e-11	9.9e-01	3.0e-08	6.0e-22	0.0e+00
1	3.5e-04	4.9e-08	1.6e-08	9.9e-01	2.3e-17	0.0e+00
2	5.6-04	7.8e-08	4.3e-10	9.9e-01	9.1e-18	0.00e+00
3	2.8e-07	1.1e-10	9.9e-01	5.6e-08	5.49e-20	0.00e+00
4	3.9e-03	2.0e-08	6.5e-12	9.9e-01	8.5e-22	0.0e+00

Πίνακας 4.7: Πιθανότητες κατατάξεως των πρώτων πέντε δεδομένων ανα επίπεδο για Κανονικές μίξεις

Οι πιθανότητες κατάταξης των πρώτων πέντε παρατηρήσεων υπολογισμένες για κάθε ομάδα υπολογισμένα βάσει του τύπου (4.7).

```
>>> gmm.predict(X[:5,:])
array([2, 3, 3, 2, 3])
```

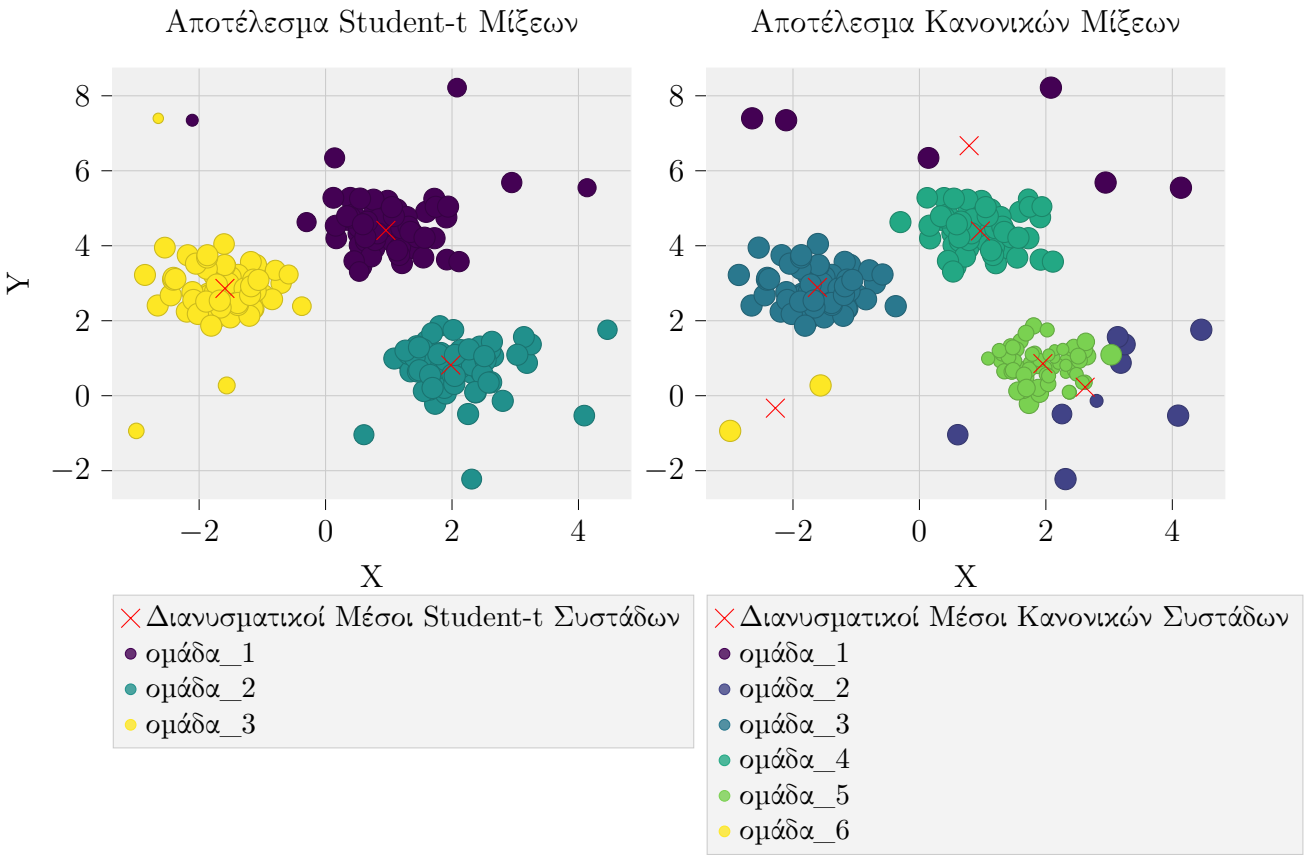
Απο το `gmm.predict(X[:5,:])`, το πρώτο και τέταρτο σημείο κατατάχθηκαν στην ομάδα 3, ενώ τα υπόλοιπα στην ομάδα 4.

Για σύγκριση των δύο διαφορετικών μίξεων θα παρουσιαστούν και τα γραφήματα με τα αποτελέσματα ομαδοποίησης και για τις δύο επιλεγμένες κατανομές (Student-t και Κανονική).

Αλγόριθμος 4.4.6: Γραφική Αναπαράσταση των Αποτελεσμάτων Ομαδοποίησης για τις δύο Κατανομές Student-t και Κανονική

```
# Γράφημα-Σύγκριση Student και Κανονικής Συσταδοποίησης
fig, ax = plt.subplots(1,2)
ax[0].scatter(X[:,0], X[:,1], s= size_stm, c= labels_stm, cmap =
 ↪  "viridis")
```

```
ax[0].scatter(stm_clusters[:,0], stm_clusters[:,1], color="violet",
    marker="x",
                label="Διανυσματικοί Μέσοι\nStudent-t Συστάδων", s=100)
ax[0].set_title("Αποτέλεσμα Student-t\nΜίξεων")
ax[0].set_xlabel('X')
ax[0].set_ylabel('Y')
[ax[0].scatter([],[],label=sustada[i], c = color[i]) for i in range(3)]
ax[0].legend()

ax[1].scatter(X[:, 0], X[:, 1], s= size_gmm ,c=labels_gmm,
    cmap='viridis');
ax[1].scatter(gmm.means_[:,0], gmm.means_[:,1], color="violet",
    marker="x",
                label="Διανυσματικοί Μέσοι\nΚανονικών Συστάδων", s=100)
ax[1].set_title("Αποτέλεσμα Κανονικών\nΜίξεων")
ax[1].set_xlabel('X')
ax[1].set_ylabel('Y')
[ax[1].scatter([],[],label=sustada_gmm[i], c = color_gmm[i]) for i in
    range(5)]
ax[1].legend()
plt.show()
```



Σχήμα 4.9: Αποτέλεσμα και σύγκριση Student-t και Κανονικών μίξεων

Όπως φαίνεται στο Σχήμα (4.9), χρησιμοποιώντας Κανονικές μίξεις δεν λαμβάνεται καλό αποτέλεσμα καθώς δεν φαίνεται να έχει επιτευχθεί καλή κατηγοριοποίηση. Συγκεκριμένα αυτό που έχει γίνει στις Κανονικές μίξεις είναι να δημιουργηθούν τρείς επιπλέον συστάδες (ομάδα_1, ομάδα_2, ομάδα_6) για να καλυφθούν τα ακραία δεδομένα. Διακρίνεται όμως πως στην περίπτωση των Student-t μίξεων οι επιπλέον ομάδες δεν χρειάστηκαν, γιατί εν αντιθέσει με την Κανονική, η Student-t διαθέτει πιο παχιές ουρές με αποτέλεσμα να μπορεί να καλύψει αυτές τις ακραίες τιμές. Το συμπέρασμα δηλαδή είναι πως δεν χρειάζεται να σπάσει η δομή δημιουργώντας επιπλέον τρείς ανεπιθύμητες συστάδες. Γι'αυτό το λόγο η εφαρμογή Student-t κατανομών έδωσε μία γενικά καλύτερη εικόνα.

Κεφάλαιο 5

Διαχωριστική Ανάλυση με Σφαιρικές και Ελλειπτικές Κατανομές

Σε αυτό το Κεφάλαιο θα μελετηθεί η τεχνική της Διαχωριστικής (Διακριτικής) Ανάλυσης για Σφαιρικές και Ελλειπτικές κατανομές. Συνήθως εφαρμόζουμε αυτή την τεχνική κατηγοριοποίησης χρησιμοποιώντας Κανονική κατανομή. Ωστόσο σε πραγματικά δεδομένα, όπως εξηγήθηκε σε προηγούμενα Κεφάλαια, το να περιορίζουμε το είδος των δεδομένων μόνο σε Κανονικά είναι πολλές φορές μη ρεαλιστικό. Πολλά πραγματικά δεδομένα διαθέτουν ασυμμετρίες στις πυκνότητές τους αλλά και αρκετές ακραίες τιμές. Συνεπώς η χρήση πιο ευέλικτων και γενικών πυκνοτήτων είναι αναγκαία για την σωστή προσαρμογή μιας κατανομής στα δεδομένα. Στο παρόν Κεφάλαιο Θα μελετηθούν οι κατανομές που αναλύθηκαν στα προηγούμενα Κεφάλαια τόσο θεωρητικά όσο και στην πράξη παρουσιάζοντας και μερικά παραδείγματα με αυτές. Ως εφαρμογή θα διερευνηθεί το πλήθος δεδομένων Iris απο το οποίο θα επιλεγούν τρείς μεταβλητές και δύο πληθυσμοί φυτών (Versicolour, Virginica) για ανάλυση με σκοπό την καλύτερη γραφική εποπτεία των αποτελεσμάτων.

5.1 Διαχωριστική Ανάλυση

Σκοπός της διαχωριστικής ανάλυσης είναι να κατατάξει μία νέα παρατήρηση που λαμβάνεται $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^T$ σε έναν απο τους υποψήφιους πληθυσμούς απο τους οποίους προέρχονται όλα τα δεδομένα. Στην περίπτωση δύο πληθυσμών, κάθε παρατήρηση μπορεί να προέρχεται απο μία εκ των δύο κατανομών $f_1(\boldsymbol{x}; \theta_1)$ ή $f(\boldsymbol{x}; \theta_2)$ με θ_1 και θ_2 τις παραμέτρους απο την πρώτη και δεύτερη κατανομή αντίστοιχα. Γενικότερα όσους πληθυσμούς έχουμε τόσες και οι κατανομές απο τις οποίες προέρχονται τα δεδομένα. Ο κανόνας του διαχωρισμού βασίζεται στο να ελαχιστοποιήσει την πιθανότητα λανθασμένης κατάταξης. Συγκεκριμένα (για δύο πληθυσμούς) δημιουργούνται περιοχές $R_1 : G(\boldsymbol{x}) > c$ και $R_2 : G(\boldsymbol{x}) \leq c$ όπου $G(\boldsymbol{x})$ μία συνάρτηση που καθορίζει την απόφαση. Εάν $\boldsymbol{x} \in R_1$ τότε η παρατήρηση κατατάσσεται στον πληθυσμό 1. Διαφορετικά όταν $\boldsymbol{x} \in R_2$ θεωρείται οτι προήλθε απο τον πληθυσμό 2. Όπως αναφέρθηκε, αυτές οι περιοχές επιλέγονται ώστε να ελαχιστοποιηθεί η ποσότητα (συνολική πιθανότητα λανθασμένης ταξινόμησης) :

$$TPM = P(2|1) + P(1|2) = \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x},$$

όπου $P(i|j)$ πιθανότητα να ανήκει στον j ενώ τοποθετείται στον i πληθυσμό. Ωστόσο συνήθως επειδή η λάθος κατάταξη συνεπάγεται και κάποιο κόστος, η βελτιστοποίηση του Διαχωρισμού βασίζεται στην ελαχιστοποίηση του μέσου κόστους λανθασμένης κατάταξης :

$$ECM = c(1|2)P(1|2)p_2 + c(2|1)P(2|1)p_1,$$

όπου το $c(i|j)$ είναι το κόστος κατηγοριοποίησης παρατήρησης στον i πληθυσμό ενώ ανήκει στον j και p_i η εκ των προτέρων πιθανότητα του i συνόλου. Εν τέλει, ο κανόνας απόφασης λαμβάνει την μορφή :

$$\begin{aligned} &R_1 : G(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)p_2}{c(2|1)p_1} \\ &R_2 : \text{Διαφορετικά.} \end{aligned}$$

Στην παρακάτω ανάλυση για λόγους απλοποίησης θα ληφθεί η περίπτωση ίσων εκ των προτέρων πιθανοτήτων καθώς και ίσων σφαλμάτων, συνεπώς το $R_1 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1$ με f πλέον τη συνάρτηση της Ελλειπτικής κατανομής. Υποθέτουμε αρχικά δύο Ελλειπτικού τύπου κατανομές με συναρτήσεις πυκνοτήτων :

$$\begin{aligned} f_1(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) &= \eta_{p1}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)) \\ f_2(\boldsymbol{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) &= \eta_{p2}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)). \end{aligned}$$

Για να ορίσουμε τον κανόνα απόφασης κατάταξης θα διακρίνουμε δύο περιπτώσεις. Η μία αφορά πληθυσμούς-κατανομές με ίσους πίνακες διακυμάνσεων και η άλλη με άνισους.
Σε αυτή την περίπτωση οι δύο πληθυσμοί Π_1, Π_2 διαμορφώνονται ως εξής :

$$f_i(\boldsymbol{x}) = \eta_{pi}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i))$$

όπου η_{pi} μία σταθερά ορισμένη για κάθε ένα πληθυσμό i και $\boldsymbol{\mu}_i \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_i$ θετικά ορισμένος και g η συνάρτηση που δημιουργεί την πυκνότητα του Ελλειπτικού με $i = \{1, 2\}$. Αν υποθέσουμε πως διαθέτουμε ενα πλήθος δεδομένων $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}$ και $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}$ απο τους πληθυσμούς Π_1, Π_2 αντίστοιχα, με n_1, n_2 τα πλήθοι των παρατηρήσεων, γνωρίζοντας τις παραμέτρους των κατανομών $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ και $\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$ (αν δεν είναι γνωστές πρέπει να εκτιμηθούν), τότε οι περιοχές κατάταξης αλλά και η συνάρτηση διαχωρισμού $G(\boldsymbol{x})$ διαμορφώνονται με τον τρόπο :

$$\begin{aligned} &R_1 : G(\boldsymbol{x}) = \frac{\eta_{p1}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1))}{\eta_{p2}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} g((\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2))} \geq 1 \\ &R_2 : \text{Διαφορετικά.} \end{aligned}$$

Στην ειδική περίπτωση $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ και με σταθερές ίσες $\eta_{p1} = \eta_{p2}$ η συνάρτηση $G(\boldsymbol{x})$ που

καθορίζει τις περιοχές κατάταξης είναι η εξής :

$$\frac{g((\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1))}{g((\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2))} \geq 1.$$

Η συνάρτηση g και η σταθερά η_p καθορίζει και το είδος της Ελλειπτικής κατανομής. Παρακάτω θα περιγραφούν οι κατανομές Kotz, Laplace και Student-t σχετικά με το πώς διαμορφώνεται το κριτήριο απόφασης κατάταξης των δεδομένων.

5.1.1 Κατανομή Kotz

Σε αυτή την περίπτωση θα παρουσιαστεί ο κανόνας απόφασης όταν οι κατανομές είναι Kotz, δηλαδή :

$$f_i(\boldsymbol{x}) = c_p|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \cdot e^{-[(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)]^{\frac{1}{2}}}$$

με $\boldsymbol{\mu}_i \in \mathbb{R}^p$, $\boldsymbol{\Sigma}_i$ θετικά ορισμένοι για $i = \{1, 2\}$ και $c_p = \frac{\Gamma(\frac{p}{2})}{2\pi^{\frac{p}{2}}\Gamma(p)}$. Έστω τώρα οτι έχουμε συγκεντρώσει ένα δείγμα n_1 παρατηρήσεων απο τον Π_1 πληθυσμό $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}$ και n_2 απο τον Π_2 με δείγμα $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}$. Επιπλέον έχουμε υπολογίσει τους εκτιμητές των παραμέτρων των δύο κατανομών ως $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1$ και $\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_2$. Τότε η συνάρτηση απόφασης $G(\boldsymbol{x})$ που ορίζει και τους χώρους R_1, R_2 δίνεται στην μορφή :

$$\hat{R}_1 : \left(\frac{|\hat{\boldsymbol{\Sigma}}_1|}{|\hat{\boldsymbol{\Sigma}}_2|}\right)^{-\frac{1}{2}} \cdot e^{-\sqrt{(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)}+\sqrt{(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_2)}} \geq 1$$

$$\hat{R}_2 : \text{Διαφορετικά.}$$

Χρησιμοποιώντας λογάριθμο για απλοποίηση του τύπου της συνάρτησης κατάταξης, ο γενικός κανόνας κατηγοριοποίησης μιας παρατήρησης $\boldsymbol{x}_0$ στον πληθυσμό Π_1 υπο τις συνθήκες $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ και ίσες εκ των προτέρων πιθανότητες των πληθυσμών είναι ο εξής :

$$\sqrt{(\boldsymbol{x}_0-\hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}_0-\hat{\boldsymbol{\mu}}_2)} - \sqrt{(\boldsymbol{x}_0-\hat{\boldsymbol{\mu}}_1)^T\hat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{x}_0-\hat{\boldsymbol{\mu}}_1)} - \frac{1}{2}\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}_1|}{|\hat{\boldsymbol{\Sigma}}_2|}\right) \geq 0.$$

Απο την άλλη ο γενικός κανόνας κατηγοριοποίησης στον πληθυσμό Π_1 υπο τις συνθήκες $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ και ίσες εκ των προτέρων πιθανότητες των πληθυσμών είναι ο εξής :

$$(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2) \geq 0.$$

5.1.2 Κατανομή Laplace

Στην περίπτωση που οι δύο πληθυσμοί Π_1, Π_2 ακολουθούν Πολυδιάστατη κατανομή Laplace, αναλυτικά έχουμε δύο κατανομές :

$$f_1(\boldsymbol{x}) = \frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right)^{\frac{p}{2}-1}}} \tag{5.1}$$

$$f_2(\boldsymbol{x}) = \frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right)^{\frac{p}{2}-1}}}. \tag{5.2}$$

Η γενική μορφή των περιοχών R_1, R_2 για την κατάταξη παρατηρήσεων στους πληθυσμούς Π_1, Π_2 αντίστοιχα ορίζεται με τον παρακάτω τρόπο :

$$R_1 : \frac{\frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right)^{\frac{p}{2}-1}}}}{\frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right)^{\frac{p}{2}-1}}}} \geq 1$$

R_2 : Διαφορετικά.

Με συνέπεια η συνάρτηση κατάταξης να διαμορφώνεται ως :

$$G(\boldsymbol{x}) = \frac{\frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}{2}\right)^{\frac{p}{2}-1}}}}{\frac{2}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \cdot \frac{\left[\frac{\pi}{2\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}\right]^{\frac{1}{2}} e^{-\sqrt{2(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}}}{\sqrt{\left(\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}{2}\right)^{\frac{p}{2}-1}}}}.$$

Για να απλοποιηθούν οι τύποι (5.1), (5.2) μπορεί να λογαριθμηθεί το παραπάνω πηλίκο οπότε ο έλεγχος δίνεται σε μια απλοποιημένη μορφή. Επιπρόσθετα για διευκόλυνση χρησιμοποιείται και η συντόμευση $\boldsymbol{u} = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}$, οπότε παίρνουμε :

$$\frac{1}{2}\ln|\boldsymbol{\Sigma}_2| + \frac{1}{2}\ln\boldsymbol{u}_2 + \boldsymbol{u}_2\sqrt{2} + (\frac{p}{2}-1)\ln\boldsymbol{u}_2 - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{1}{2}\ln\boldsymbol{u}_1 - \boldsymbol{u}_1\sqrt{2} - (\frac{p}{2}-1)\ln\boldsymbol{u}_1 \geq 0.$$

Αν τα $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ είναι άγνωστα, εκτιμούνται ως $\hat{\boldsymbol{\mu}}_i$ και $\hat{\boldsymbol{\Sigma}}_i$ $i = \{1, 2\}$ (ε.μ.π. των παραμέτρων) κάνοντας χρήση κάποιας αριθμητικής μεθόδου (πχ Newton-Raphson). Μπορεί να γίνουν απλοποιήσεις του τύπου, παραδείγματος χάριν στην περίπτωση ίσων πινάκων $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

5.1.3 Κατανομή Student-t

Σε αυτή την περίπτωση οι πυκνότητες των πληθυσμών ορίζονται ως :

$$f_1(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = \frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}) n^{\frac{p}{2}} \pi^{\frac{p}{2}} |\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)]^{-\frac{n+p}{2}}$$

$$f_2(\boldsymbol{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}) n^{\frac{p}{2}} \pi^{\frac{p}{2}} |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)]^{-\frac{n+p}{2}}$$

με αποτέλεσμα να διαμορφώνονται οι περιοχές κατάταξης στους πληθυσμούς ως εξής :

$$R_1 : \frac{\frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}) n^{\frac{p}{2}} \pi^{\frac{p}{2}} |\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)]^{-\frac{n+p}{2}}}{\frac{\Gamma(\frac{n+p}{2})}{\Gamma(\frac{n}{2}) n^{\frac{p}{2}} \pi^{\frac{p}{2}} |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)]^{-\frac{n+p}{2}}} \geq 1$$

R_2 : Διαφορετικά.

Απο τον παραπάνω τύπο, δοθέντος ενός σημείου $\boldsymbol{x}$, μας ενδιαφέρει η ποσότητα :

$$G(\boldsymbol{x}) = \frac{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)]^{-\frac{n+p}{2}}}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}} \cdot [1 + \frac{1}{n}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)]^{-\frac{n+p}{2}}}$$

Εναλλακτική μορφή μπορεί να δωθεί μέσω λογαρίθμησης, χρησιμοποιώντας την συντόμευση $u_i = (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$:

$$\ln G(\boldsymbol{x}) = \frac{1}{2} \ln |\boldsymbol{\Sigma}_2| - \frac{n+p}{2} \ln[1 + \frac{1}{n} \cdot u_2] - \frac{1}{2} \ln |\boldsymbol{\Sigma}_1| + \frac{n+p}{2} \ln[1 + \frac{1}{n} \cdot u_1].$$

Για $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ άγνωστα, αντικαθιστούμε στους τύπους τις εκτιμήσεις τους $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ αντίστοιχα.

5.2 Παράδειγμα Εφαρμογής Ελλειπτικών Κατανομών στη Διαχωριστική Ανάλυση

Στην ενότητα αυτή θα αναλυθούν τα γνωστά δεδομένα Iris του Fisher που αφορούν τις διαστάσεις διαφορετικών ειδών φυτών Setoca, Versicolour, Virginica. Συγκεκριμένα η Διαχωριστική Ανάλυση θα αφορά τα δύο είδη φυτών Versicolour και Virginica. Τα δεδομένα είναι τεσσάρων διαστάσεων καθώς αφορούν τις μεταβλητές sepal length, sepal width, petal length, petal width. Για λόγους εύκολης γεωμετρικής αναπαράστασης θα χρησιμοποιηθούν οι πρώτες τρείς μεταβλητές. Συνεπώς τα δεδομένα είναι τρισδιάστατα με δύο πληθυσμούς (Versicolour, Virginica).

Αλγόριθμος 5.2.1: Πίνακας των Δεδομένων του Iris Συνόλου

```
### Python Διαχωριστική Ανάλυση ###
from sklearn import datasets
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import scipy as sc
import seaborn as sns
import math
from scipy.stats import chi2

iris = datasets.load_iris()
x = iris.data[:,:3]
y = iris.target

dfx = pd.DataFrame(x)
dfy = pd.DataFrame(y)
df = pd.concat([dfx,dfy], axis = 1)

# Επιλογή μόνο των ετικετών 1 και 2, δηλαδή Virginica και Vercicolour
df = df[df['labels'] != 0]
#Αλλαγή των κωδικών των ετικετών για να ξεκινάνε απο το μηδέν
df['labels'].replace({1:0,2:1}, inplace = True)
df.columns = ["Sepal Length","Sepal Width", "Petal Length","labels"]
df.reset_index(drop = True, inplace = True)

#Επαναυπολογισμός των dfx, dfy βάσει των νέων ετικετών
dfx = df[['Sepal Length','Sepal Width', 'Petal Length']]
dfy = df[['labels']]
```

	Sepal Length	Sepal Width	Petal Length	labels
0	7.0	3.2	4.7	0
1	6.4	3.2	4.5	0
2	6.9	3.1	4.9	0
3	5.5	2.3	4.0	0
4	6.5	2.8	4.6	0

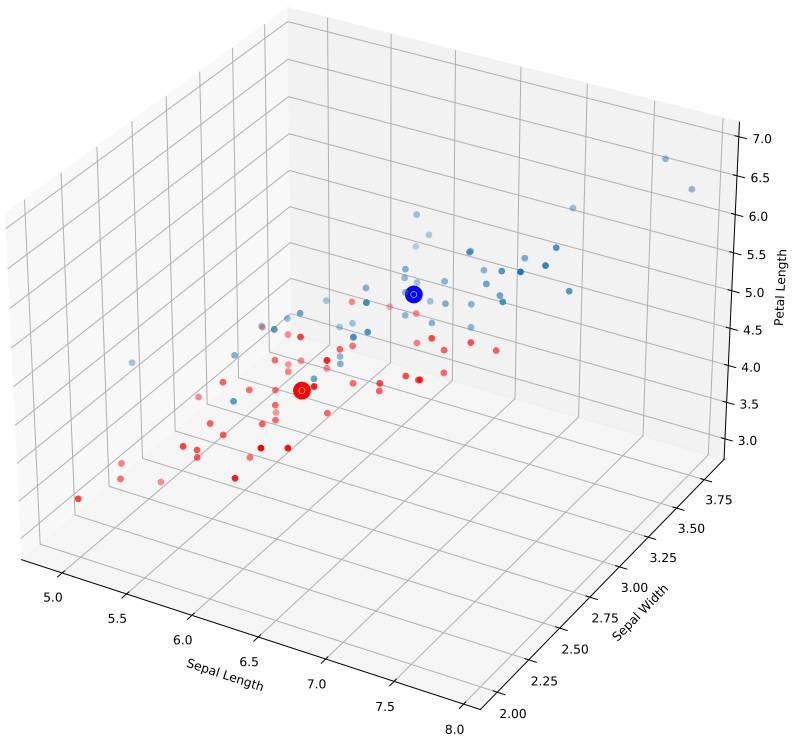
Πίνακας 5.1: Περιληπτικός πίνακας δεδομένων Iris

Διακρίνονται οι στήλες των παρατηρήσεων αλλά και το είδος (labels) των φυτών με κωδικούς 0 και 1. Ένας τρόπος γεωμετρικής αναπαράστασης των δεδομένων είναι το τρισδιάστατο διάγραμμα νέφους σημείων (scatter-plot). Αυτό αρχικά θα μας δώσει μία εικόνα για το άν οι δύο πληθυσμοί Π_1, Π_2 ειδών φυτών (Versicolour, Virginica) διαφέρουν.

Αλγόριθμος 5.2.2: Τρισδιάστατο Scatter Plot

```
#labels των Versicolour (lab1) και Virginica (lab2)
lab1 = (dfy[dfy==0].dropna()).index
lab2 = (dfy[dfy==1].dropna()).index

```

```
#dataframes των μεταβλητών για το είδος Versicolour (df1) και Virginica
↪ (df2)
df1 = dfx.iloc[lab1,:]
df2 = dfx.iloc[lab2,:]

#Πίνακες scale Σ και διανυσματικοί μέσοι για κάθε είδος φυτού (label)
sigma1 = np.cov([df1.iloc[:,0], df1.iloc[:,1], df1.iloc[:,2]])
sigma2 = np.cov([df2.iloc[:,0], df2.iloc[:,1], df2.iloc[:,2]])
sigma = [sigma1, sigma2]

mean1 = df1.aggregate("mean")
mean2 = df2.aggregate("mean")
mean = [mean1, mean2]

# 3D Plot
fig = plt.figure(figsize=(12, 12))
ax = fig.add_subplot(projection='3d')
ax.scatter(df1.iloc[:,0],df1.iloc[:,1],df1.iloc[:,2], color = "red")
ax.scatter(df2.iloc[:,0],df2.iloc[:,1],df2.iloc[:,2])
ax.scatter(mean1[0], mean1[1], mean1[2], linewidths = 10, c = "red")
ax.scatter(mean2[0], mean2[1], mean2[2], linewidths = 10, c = "blue")
ax.set_xlabel("Sepal Length")
ax.set_ylabel("Sepal Width")
ax.set_zlabel("Petal Length")
plt.show()
```



Σχήμα 5.1: Τρισδιάστατο Γράφημα Iris Δεδομένων των φυτών Versicolour και Virginica

Απο το Σχήμα (5.1) διακρίνουμε πως τα δύο είδη φυτών παρουσιάζουν διαφορές στις τρείς επιλεγμένες μεταβλητές. Πιό αναλυτικά οι πληθυσμοί είναι τοποθετημένοι σε τέτοιες αποστάσεις που μπορούν να φανούν τα μπλέ απο τα κόκκινα δεδομένα, δηλαδή δεν συγχέονται τελείως μεταξύ τους. Αυτή είναι η πραγματική κατάταξη των δεδομένων. Στην Διαχωριστική Ανάλυση, το είδος φυτού που ανήκει η κάθε νέα παρατήρηση είναι άγνωστο. Η προσπάθεια σε αυτή την περίπτωση είναι να εκτιμηθεί το είδος του φυτού για το κάθε ενα νέο δεδομένο με όσο το δυνατόν μικρότερο σφάλμα κατάταξης συγκριτικά με την πραγματικότητα. Θα υποθέσουμε άνισους πίνακες $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ για τους ελέγχους που θα ακολουθήσουν.

Για να προχωρήσουμε παρακάτω και να προσαρμόσουμε τις Ελλειπτικές κατανομές στα δεδομένα, εφόσον αναφερόμαστε σε μία διαδικασία εποπτευόμενης μάθησης, αρχικά θα χωρίσουμε τα δεδομένα σε εκπαιδευτικά (train set) και πειραματικά (test set). Στα εκπαιδευτικά θα γίνει η εκτίμηση των παραμέτρων για τις κατανομές των πληθυσμών και στη συνέχεια το μοντέλο θα εκτιμήσει για καθένα πειραματικό δεδομένο την ετικέτα του (label) που αποτελεί το είδος του φυτού. Έτσι στο τέλος μπορεί να γίνει μία εκτίμηση της πιθανότητας σωστής και λανθασμένης κατάταξης που θα καθορίσει και την αποτελεσματικότητα του μοντέλου στο να ανιχνεύει επιτυχώς την κατηγορία των Iris φυτών. Αναλυτικά, στον αλγόριθμο (5.2.3) που ακολουθεί χωρίζουμε το σύνολο των δεδομένων σε 70% εκπαιδευτικά και 30% πειραματικά.

Αλγόριθμος 5.2.3: Δημιουργία Εκπαιδευτικών και Πειραματικών Συνόλων

```
# Διαχωρισμός Δεδομένων σε Εκπαιδευτικά και Πειραματικά
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(xn, yn, test_size =
↪ 0.3)

#Τα dataframes των εκπαιδευτικών δεδομένων (τα dfx, dfy αφορούν πλέον
↪ τα εκπαιδευτικά δεδομένα μιας και με αυτά θα συνεχιστεί η ανάλυση)
dfx = pd.DataFrame(X_train).reset_index(drop = True)
dfy = pd.DataFrame(y_train).reset_index(drop = True)
df = pd.concat([dfx,dfy], axis = 1)
df.columns = ["Sepal Length","Sepal Width", "Petal Length","labels"]

loc1 = (dfy[dfy==0].dropna()).index
loc2 = (dfy[dfy==1].dropna()).index

#Τα labels των εκπαιδευτικών δεδομένων
lab1 = (dfy[dfy==0].dropna()).index
lab2 = (dfy[dfy==1].dropna()).index
labs = np.concatenate([lab1,lab2])

#Τα dataframes των εκπαιδευτικών δεδομένων
df1 = dfx.iloc[lab1,:]
df2 = dfx.iloc[lab2,:]
```

Έχοντας πλέον τα εκπαιδευτικά δεδομένα που αποτελούν το 70% του συνόλου του δείγματος (τυχαία επιλεγμένα), η εφαρμογή για κάθε μία Ελλειπτική κατανομή θα γίνει χρησιμοποιώντας αυτές τις παρατηρήσεις. Στη συνέχεια, η αποτελεσματικότητα του μοντέλου θα διερευνηθεί στο πειραματικό σύνολο (30% του συνόλου του δείγματος) για να βρούμε το ποσοστό των σωστών ταξινομήσεων.

Γενικά. θα ακολουθήσουμε την εξής μέθοδο για την εφαρμογή Διαχωριστικής Ανάλυσης στο σύνολο δεδομένων :

- Βάσει της κατανομής που έχει επιλεγεί για την αναπαράσταση των δύο πληθυσμών των φυτών (Versicolour, Virginica), αρχικά θα εκτιμηθεί η απο κοινού πιθανοφάνεια του πρώτου και δεύτερου πληθυσμού ξεχωριστά.

- Αφού έχουμε τις ε.μ.π. εκτιμήτριες των παραμέτρων για κάθε μία ομάδα (είδη φυτών), γνωρίζουμε πλέον τις κατανομές των πληθυσμών, συνεπώς μέσω της συνάρτησης απόφασης G μπορούμε να εκτιμήσουμε την ομάδα των πειραματικών δεδομένων (test data).

- Τέλος, υπολογίζουμε τις πιθανότητες σωστών ταξινομήσεων και ερμηνεύουμε την αποτελεσματικότητα του μοντέλου βάσει της επιλεγμένης κατανομής.

Μπορούμε να επεξηγήσουμε την μέθοδο, αναλύοντας και τον κώδικα που θα χρησιμοποιηθεί στους αλγορίθμους (5.2.4, 5.2.8, 5.2.12). Αρχικά θα υπολογιστεί η απο κοινού πιθανοφάνεια για καθεμιά ομάδα (Versicolour:1, Virginica:2) ως $lik1 = \prod_{i=1}^{n} f_1(\boldsymbol{x}_i|\theta)$, $lik2 = \prod_{i=1}^{n} f_1(\boldsymbol{x}_i|\theta)$ ή το λογάριθμό τους, $loglik1 = \sum_{i=1}^{n} \log f_1(\boldsymbol{x}_i|\theta)$, $loglik2 = \sum_{i=1}^{n} \log f_2(\boldsymbol{x}_i|\theta)$ για ευκολία

πράξεων, ως προς την Ελλειπτική κατανομή που έχει επιλεγεί για την ανάλυση. Για να βρεθούν οι ε.μ.π. των κατανομών των πληθυσμών, θα εφαρμοστεί αλγόριθμος βελτιστοποίησης των πιθανοφανειών. Συγκεκριμένα απο την Python είναι το πακέτο `scipy.optimize.minimize` με το οποίο θα ελαχιστοποιήσουμε την αρνητική πιθανοφάνεια (ελαχιστοποίηση της αρνητικής είναι ισοδύναμο με την μεγιστοποίηση της θετικής πιθανοφάνειας, διότι το πακέτο αυτό μέχρι στιγμής δεν διαθέτει αλγόριθμο μεγιστοποίησης). Στον κώδικα βελτιστοποίησης εισάγεται το $\boldsymbol{x}$ που περιέχει όλες τις απαραίτητες τιμές που χρειάζονται (αρχικές εκτιμήσεις των παραμέτρων για την έναρξη του κώδικα) γράφοντας έτσι `scipy.optimize.minimize`$(l'(\theta), \boldsymbol{x}0 = \boldsymbol{x})$, όπου $\boldsymbol{x} = [x_1, \ldots, x_l]$ με l το πλήθος των παραμέτρων ελαχιστοποίησης και $l'(\theta) = -l(\theta)$ η αρνητική απο κοινού πιθανοφάνεια (ή αλλιώς χρησιμοποιούμε τον αρνητικό λογάριθμο της πιθανοφάνειας για ευκολία πράξεων). Οπότε το $\boldsymbol{x}$ αποτελεί μία λίστα με το σύνολο των αγνώστων παραμέτρων. Συνεπώς στην δική μας περίπτωση, κάθε ομάδα-κατανομή διαθέτει παραμέτρους το $\boldsymbol{\mu}$ και το $\boldsymbol{\Sigma}$. Το $\boldsymbol{\mu}$ και το $\boldsymbol{\Sigma}$ όμως επειδή είναι σε μορφή διανύσματος και πίνακα αντίστοιχα, θα εισέλθουν στο πακέτο `scipy.optimize.minimize` σε μορφή λίστας $\boldsymbol{x} = [\mu_1, \mu_2, \mu_3, \sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{21}, \sigma_{22}, \sigma_{23}, \sigma_{31}, \sigma_{32}, \sigma_{33}]$. Κρίσιμο κομμάτι όμως σε αυτή την φάση είναι πως προσπαθώντας να βρεί τα βέλτιστα σ_{ij}, ο αλγόριθμος μπορεί να βρεί τέτοιες τιμές που η ορίζουσα του πίνακα $\boldsymbol{\Sigma}$ να είναι αρνητική, δίνοντας σφάλμα διότι την θέλουμε μή αρνητική (οι κατανομές περιέχουν την ποσότητα $\sqrt{|\boldsymbol{\Sigma}|}$). Για αυτό το λόγο αντί να χρησιμοποιηθεί στον αλγόριθμο το $\boldsymbol{\Sigma}$, θα βάλουμε τον κάτω τριγωνικό πίνακα $\boldsymbol{A}$ απο την διάσπαση Cholesky. Αφού βρούμε τις βέλτιστες τιμές του πίνακα αυτού, πλέον μπορούμε να πάρουμε τον πίνακα $\boldsymbol{\Sigma}$ ως $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. Συνεπώς πλεόν το $\boldsymbol{x}$ διαμορφώνεται ως $\boldsymbol{x} = [\mu_1, \mu_2, \mu_3, \alpha_{11}, \alpha_{21}, \alpha_{22}, \alpha_{31}, \alpha_{32}, \alpha_{33}]$ με α_{ij} τα κάτω διαγώνια στοιχεία του πίνακα Cholesky $\boldsymbol{A}$ (τα άλλα στοιχεία είναι μονίμως μηδέν άρα δεν βελτιστοποιούνται).

5.2.1 Κατανομή Kotz

Οι ε.μ.π. εκτιμητές των μέσων των πληθυσμών Π_1, Π_2 θα γράφονται ως $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ ενώ των πινάκων $\boldsymbol{\Sigma}$ ως $\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2$ και θα εκτιμηθούν με αριθμητική μέθοδο. Γι'αυτό πρώτα θα κατασκευαστούν οι πιθανοφάνειες για καθένα επίπεδο και στη συνέχεια θα χρησιμοποιηθεί αλγόριθμος μεγιστοποίησης της πιθανοφάνειας για την εύρεση του ε.μ.π. εκτιμητή του μέσου και του $\boldsymbol{\Sigma}$ υπο την προϋπόθεση ότι οι πληθυσμοί ακολουθούν κατανομή Kotz. Για την μεγιστοποίηση της πιθανοφάνειας αρκεί να ελαχιστοποιηθεί η ποσότητα $\frac{n}{2}\ln|\boldsymbol{\Sigma}| + \sum_{i=1}^{n}\sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}$ για καθεμιά κατανομή ως προς $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ όπως δόθηκε στο κεφάλαιο της κατανομής Kotz που θα μας οδηγήσει στους ε.μ.π. εκτιμητές των παραμέτρων.

Αλγόριθμος 5.2.4: Συναρτήσεις απο Κοινού Πιθανοφανειών για κάθε Επίπεδο Κατανομής Kotz

```
loc1 = (dfy[dfy==0].dropna()).index
loc2 = (dfy[dfy==1].dropna()).index

#Συναρτήσεις loglik1, loglik2 (για πληθυσμούς Versicolour, Virginica
 ↪  αντίστοιχα) με σκοπό την ελαχιστοποίηση της
 ↪  (n/2)ln|Σ| + Σ_{i=1}^{n} sqrt((x_i − μ)^T Σ^{-1} (x_i − μ))  ως προς μ, Σ. Η ελαχιστοποίησή
 ↪  της σημαίνει μεγιστοποίηση της απο κοινού πιθανοφάνειας της
 ↪  κατανομής Kotz
```

```
5
6  #Το x αποτελεί την λίστα με όλες τις παραμέτρους της Kotz σε ανοικτή
   ↪  μορφή, δηλαδή η είσοδος είναι x = [μ1, μ2, μ3, a11, a12, a13, a22, a23, a33]. Τα
   ↪  στοιχεία a είναι του πίνακα cholesky A οπότε απο αυτά βρίσκεται το
   ↪  Σ = AA^T
7  def loglik1(x)  :
8      n = len(loc1)
9      m1 = np.array(x[:3])
10     chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshap
       ↪  e(3,3)
11     s1 = chol@chol.T
12     sq = []
13     for i in loc1 :
14         x = dfx.iloc[i,:]
15         sq.append(np.sqrt((x-m1).T @ np.linalg.inv(s1) @ (x-m1)))
16     result = n/2*np.log(np.linalg.det(s1)) + np.sum(sq)
17     return result
18
19
20 def loglik2(x)  :
21     n = len(loc2)
22     m2 = x[:3]
23     chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshap
       ↪  e(3,3)
24     s2 = chol@chol.T
25     nu = 3
26     sq = []
27     for i in loc2 :
28         x = dfx.iloc[i,:]
29         sq.append(np.sqrt((x-m2).T @ np.linalg.inv(s2) @ (x-m2)))
30     result = n/2*np.log(np.linalg.det(s2)) + np.sum(sq)
31     return result
32
33 ch1 = np.linalg.cholesky(sigma1)
34 ch2 = np.linalg.cholesky(sigma2)
35
36 idx1 = np.flatnonzero(ch1)
37 idx2 = np.flatnonzero(ch2)
38
39 #Αρχικές τιμές των παραμέτρων για κάθε ομάδα
40 x01 = list(mean1) + list(ch1.flatten()[idx1])
41 x02 = list(mean2) + list(ch2.flatten()[idx2])
```

Στη συνέχεια θα εφαρμοστεί ο αλγόριθμος (5.2.5) ελαχιστοποίησης των συναρτήσεων του αλγορίθμου (5.2.4) που θα οδηγήσει στους ε.μ.π. εκτιμητές. Η μέθοδος ελαχιστοποίησης ορίζεται στην python απο το πακέτο scipy.optimize.

Αλγόριθμος 5.2.5: Αριθμητική Μέθοδος Εύρεσης ε.μ.π. Εκτιμητών Kotz

```
#Διαδικασία εύρεσης ελαχίστων για τις δύο κατανομές-ομάδες
opt1 = sc.optimize.minimize(loglik1, x0 = x01, method = "Nelder-Mead",
 ↪  tol = 1e-3)
opt2 = sc.optimize.minimize(loglik2, x0 = x02, method = "Nelder-Mead",
 ↪  tol = 1e-3)

#Εύρεση των μέσων και των πινάκων Σ απο τα κελιά του πίνακα διάσπασης
 ↪  Cholesky
ch1_mle = np.zeros([3,3])
idx1 =  np.tril_indices(3)
ch1_mle[idx1] = opt1.x[3:9]
mean1_mle = opt1.x[:3]
sigma1_mle = ch1_mle@ch1_mle.T

ch2_mle = np.zeros([3,3])
idx2 =  np.tril_indices(3)
ch2_mle[idx2] = opt2.x[3:9]
mean2_mle = opt2.x[:3]
sigma2_mle = ch2_mle@ch2_mle.T
```

Αρχικά, το αποτέλεσμα της μεθόδου για τον πληθυσμό Versicolour είναι το εξής :

Πίνακας 5.2: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Kotz στον πληθυσμό Versicolour

```
>>> opt1.success
True

>>> opt1.message
'Optimization terminated successfully.'

>>> opt1.x
array([5.86866707, 2.78697622, 4.21364358, 0.2441447 , 0.06820048,
       0.13722081, 0.14634634, 0.07133742, 0.14493925])
```

Βάσει του πίνακα (5.2), η διαδικασία εκτελέστηκε με επιτυχία (opt1.success = True). Απο το opt1.x εξάγονται όλες οι τιμές των παραμέτρων της κατανομής Kotz της ομάδας Versicolour. Συγκεκριμένα απο τις πρώτες τρείς θέσεις του opt1.x παίρνουμε τον διανυσματικό μέσο ενώ οι τιμές στις θέσεις τέσσερα έως εννιά αφορούν τους μή μηδενικούς αριθμούς απο τον πίνακα διάσπασης Cholesky $\boldsymbol{A}$. Συγκεκριμένα, για τον μέσο η ε.μ.π. εκτιμήτρια είναι $\hat{\boldsymbol{\mu}}_1 = (5.86866707, 2.78697622, 4.21364358)^T$ και για τον πίνακα $\hat{\boldsymbol{\Sigma}}_1$ (που προκύπτει απο την εκτίμηση που λάβαμε του πίνακα Cholesky) :

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.2441447 & 0 & 0 \\ 0.06820048 & 0.13722081 & 0 \\ 0.14634634 & 0.07133742 & 0.14493925 \end{pmatrix}$$

απο τον οποίο προκύπτει ο πίνακας $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.05960663 & 0.01665079 & 0.03572968 \\ 0.01665079 & 0.02348086 & 0.01976987 \\ 0.03572968 & 0.01976987 & 0.04751366 \end{pmatrix}.$$

Απο την άλλη, οι αντίστοιχες τιμές των παραμέτρων που έδωσε η διαδικασία για τον πληθυσμό Virginica είναι :

Πίνακας 5.3: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Kotz στον πληθυσμό Virginica

```
>>> opt2.success
True

>>> opt2.message
'Optimization terminated successfully.'

>>> opt2.x
array([ 6.55548066,  2.98365443,  5.52902695,  0.36802256,  0.08929001,
        0.14592935,  0.26306576, -0.01133036,  0.12983812])
```

Απο πίνακα (5.3), ο ε.μ.π. του μέσου είναι $\boldsymbol{\mu}_2 = (6.55548066, 2.98365443, 5.52902695)^T$, ενώ για τον πίνακα διάσπασης έχουμε :

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.36802256 & 0 & 0 \\ 0.08929001 & 0.14592935 & 0 \\ 0.26306576 & -0.01133036 & 0.12983812 \end{pmatrix}$$

με ε.μ.π. του $\boldsymbol{\Sigma}_2$:

$$\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.1354406 & 0.03286074 & 0.09681413 \\ 0.03286074 & 0.02926808 & 0.02183571 \\ 0.09681413 & 0.02183571 & 0.08618991 \end{pmatrix}.$$

Εφόσον πλέον έχουν βρεθεί οι ε.μ.π., διαθέτουμε όλες τις απαραίτητες εκτιμήτριες $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2$ για να δημιουργηθεί η συνάρτηση απόφασης $G(x)$ που θα κατατάσσει παρατηρήσεις στους πληθυσμούς Π_1, Π_2 (αλγόριθμος 5.2.6).

Αλγόριθμος 5.2.6: Δημιουργία Συνάρτησης Διαχωρισμού G

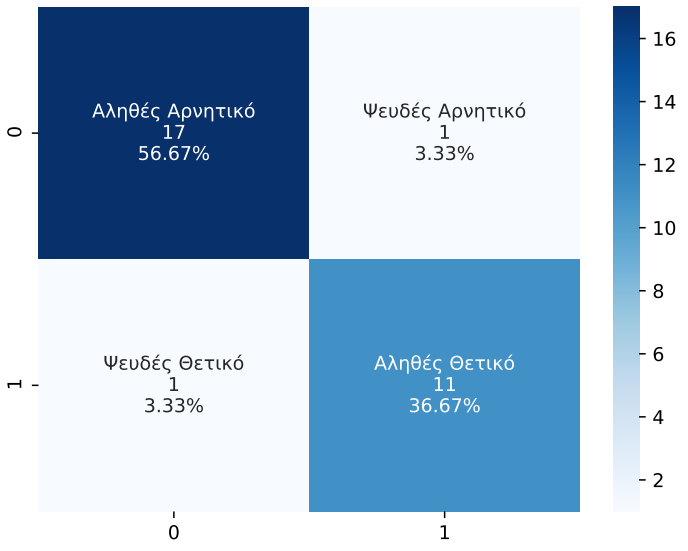
```
1 def g(x, mu1, mu2, sigma1, sigma2) :
```

```
    num = np.sqrt((x-mu2).T @ np.linalg.inv(sigma2) @ (x-mu2)) -
    ↪ np.sqrt((x-mu1).T @ np.linalg.inv(sigma1) @ (x-mu1))
    ↪ -0.5*np.log(np.linalg.det(sigma1)/np.linalg.det(sigma2))
    return num
```

όπου mu1,mu2,sigma1,sigma2 οι ε.μ.π. των $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ και $\boldsymbol{x}$ η παρατήρηση που πρέπει να καταταχθεί σε έναν απο τους δύο πληθυσμούς : Είδη φυτών Versicolour και Virginica. Η συνάρτηση `g(x, mu1, mu2, sigma1, sigma2)` επιστρέφει την πρόβλεψη του επιπέδου για το δεδομένο $\boldsymbol{x}$. Αυτή τη συνάρτηση θα χρησιμοποιήσουμε και θα την εφαρμόσουμε στο πειραματικό πλήθος (test set) για την εκτίμηση των ομάδων της κάθε μιας παρατήρησης. Το αποτέλεσμα είναι ένας πίνακας που μας δείχνει πόσο ικανοποιητικά εκτίμησε η μέθοδος τα πειραματικά δεδομένα.

Αλγόριθμος 5.2.7: Πίνακας Σύγχυσης για Ομάδες Κατανομών Kotz

```
es = lambda x : g(x, mean1_mle, mean2_mle, sigma1_mle, sigma2_mle)

y_pred = []
for i in range(X_test.shape[0]) :
    y_pred.append(res(X_test.iloc[i,:]))

y_pred = pd.Series(y_pred)
y_pred = (y_pred < 0)*1

#Εκτίμηση σωστών και εσφαλμένων κατατάξεων
from sklearn import metrics
metrics.confusion_matrix(y_test, y_pred)
```



Πίνακας 5.4: Πίνακας σύγχυσης για ομάδες κατανομών Kotz

Παρατηρείται πως το μοντέλο χρησιμοποιώντας κατανομές Kotz κατέγραψε 17 παρατηρήσεις σωστά ως Versicolour (0) και 11 σωστά ως Virginica (1). Επιπρόσθετα εκτίμησε 1 τιμή ως Veriscolour ενω ήταν Virginica και 1 τιμή Virginica ενώ ήταν Versicolour. Συνολικά έκανε δύο λάθη σε ένα σύνολο 30 πειραματικών δεδομένων. Άρα το ποσοστό σωστών ταξινομήσεων είναι 96.6%. Απο την άλλη το ποσοστό των λανθασμένων ταξινομήσεων είναι 3.3%.

5.2.2 Κατανομή Student-t

Μπορούμε να προσαρμόσουμε και Student-t Κατανομή για καθένα απο τα επίπεδα. Σε αυτή την περίπτωση εκτελούμε ακριβώς την ίδια διαδικασία με την περίπτωση της Kotz Κατανομής, με μόνη διαφορά τις εκτιμήσεις των μέσων και των πινάκων $\boldsymbol{\Sigma}$ για κάθε κατηγορία. Στους παραπάνω αλγορίθμους το μοναδικό στοιχείο που θα αλλάξει είναι ο τύπος της πιθανοφάνειας απο την οποία με κάποια αριθμητική μέθοδο θα εκτιμηθούν οι ε.μ.π. εκτιμητές του $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$. Απο τον αλγόριθμο (5.2.8) στις σειρές 9-19 και 22-31 θα προσαρμοστούν οι πιθανοφάνειες Student-t κατανομής (αντί της Kotz), οπότε προκύπτει πλέον ο κώδικας που φαίνεται φαίνεται στον αλγόριθμο (5.2.8).

Αλγόριθμος 5.2.8: Συναρτήσεις απο Κοινού Πιθανοφανειών για κάθε Επίπεδο Κατανομής Student-t

```
1  # Συνάρτηση πυκνότητας Student-t Κατανομής
2  def student_lhd(x, mu, Sigma, p, n) :
3      x, mu = np.array(x), np.array(mu)
4      Sigma = np.array(Sigma)
5      lhd = math.gamma((n+p)/2)/(math.gamma(n/2)*n**(p/2)*math.pi**(p/2)*
       ↪  np.linalg.det(Sigma)**(1/2)) * (1+ 1/n *
       ↪  (x-mu).T@np.linalg.inv(Sigma)@(x-mu))**(-(n+p)/2)
6      return lhd
7
8  # Συναρτήσεις εύρεσης (αρνητικής) απο κοινού λογαριθμικής πιθανοφάνειας
   ↪  βάσει Student-t Κατανομής για πληθυσμούς Versicolour (loglik1) και
   ↪  Virginica (loglik2)
9  def loglik1(x)  :
10     n = len(loc1)
11     m1 = np.array(x[:3])
12     chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshap
       ↪  e(3,3)
13     s1 = chol@chol.T
14     nu = 3
15     sq = []
16     for i in loc1 :
17         x = dfx.iloc[i,:]
18         sq.append(student_lhd(dfx.iloc[i,:], mu = m1, Sigma = s1, p =
           ↪  3, n = nu))
19     result = -np.prod(sq)
20     return result
21
22
```

```
def loglik2(x)  :
    n = len(loc2)
    m2 = x[:3]
    chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshape(3,3)
    s2 = chol@chol.T
    nu = 3
    sq = []
    for i in loc2 :
        x = dfx.iloc[i,:]
        sq.append(student_lhd(x = dfx.iloc[i,:], mu = m2, Sigma = s2, p = 3, n = nu))
    result = -np.prod(sq)
    return result


ch1 = np.linalg.cholesky(sigma1)
ch2 = np.linalg.cholesky(sigma2)

idx1 = np.flatnonzero(ch1)
idx2 = np.flatnonzero(ch2)
```

Έχοντας πλεόν ορίσει τις απαραίτητες συναρτήσεις εύρεσης των πιθανοφανειών για Student-t πληθυσμούς, παρακάτω βρίσκονται οι εκτιμητές ε.μ.π. των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$.

Αλγόριθμος 5.2.9: Αριθμητική Μέθοδος Εύρεσης ε.μ.π. Εκτιμητών για Student-t

```
#Εναρκτήριες τιμές τις διαδικασίας βελτιστοποίησης
x01 = list(mean1) + list(ch1.flatten()[idx1])
x02 = list(mean2) + list(ch2.flatten()[idx2])

#Εύρεση ε.μ.π. εκτιμητών
opt1 = sc.optimize.minimize(loglik1, x0 = x01, method = "Nelder-Mead", tol = 1e-3)
opt2 = sc.optimize.minimize(loglik2, x0 = x02, method = "Nelder-Mead", tol = 1e-3)
```

Συνοπτικά κάποια αποτελέσματα :

Πίνακας 5.5: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Student-t στον Πληθυσμό Versicolour

```
>>> opt1.success
True

>>> opt1.message
'Optimization terminated successfully.'

>>> opt1.x
array([5.87677804, 2.79954705, 4.2245815 , 0.39672803, 0.11328478,
       0.2135621 , 0.23322631, 0.11805561, 0.22913096])
```

Βάσει του πίνακα (5.5) βλέπουμε πως η διαδικασία εκτελέστηκε με επιτυχία. Απο το opt1.x μπορούμε να αντλήσουμε όλες τις τιμές των παραμέτρων της κατανομής Student-t της ομάδας Versicolour. Συγκεκριμένα απο τις πρώτες τρείς τιμές παίρνουμε τον διανυσματικό μέσο ενώ οι τιμές στις θέσεις τέσσερα έως εννιά αφορούν τους μή μηδενικούς αριθμούς απο τον πίνακα διάσπασης Cholesky $\boldsymbol{A}$. Ο ε.μ.π. του μέσου είναι $\hat{\boldsymbol{\mu}}_1 = (5.87677804, 2.79954705, 4.2245815)^T$, οι βαθμοί ελευθερίας είχαν προσαρμοστεί στους $\nu = 3$ και για τον πίνακα $\hat{\boldsymbol{\Sigma}}_1$:

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.39672803 & 0 & 0 \\ 0.11328478 & 0.2135621 & 0 \\ 0.23322631 & 0.11805561 & 0.22913096 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^T = \begin{pmatrix} 0.15739313 & 0.04494325 & 0.09252742 \\ 0.04494325 & 0.05844221 & 0.0516332 \\ 0.09252742 & 0.0516332 & 0.12083263 \end{pmatrix}.$$

Παρακάτω παρατίθονται οι ε.μ.π. εκτιμήσεις των παραμέτρων του δεύτερου πληθυσμού :

Πίνακας 5.6: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Student-t στον Πληθυσμό Virginica

```
>>> opt2.success
True

>>> opt2.message
'Optimization terminated successfully.'

>>> opt2.x
array([6.53735455, 2.96722638, 5.51281288, 0.58091001, 0.14014411,
       0.22671577, 0.41085935, 0.0068804 , 0.20204103])
```

Απο το αποτέλεσμα του πίνακα (5.6), πάλι φαίνεται πως η διαδικασία εκτελέστηκε με επιτυχία. Οι εκτιμήσεις ε.μ.π. των παραμέτρων της κατανομής Student-t για την ομάδα Virginica είναι $\hat{\boldsymbol{\mu}}_2 = (6.53735455, 2.96722638, 5.51281288)^T$, ν = 3 και $\hat{\boldsymbol{\Sigma}}_2$ με :

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.58091001 & 0 & 0 \\ 0.14014411 & 0.22671577 & 0 \\ 0.41085935 & 0.0068804 & 0.20204103 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^T = \begin{pmatrix} 0.33745644 & 0.08141112 & 0.23867231 \\ 0.08141112 & 0.07104041 & 0.05913941 \\ 0.23867231 & 0.05913941 & 0.20967332 \end{pmatrix}.$$

Μέχρι στιγμής αυτό που έγινε είναι να εκτιμηθούν οι παράμετροι των δύο πληθυσμών. Το σημαντικό κομμάτι παρατίθεται παρακάτω καθώς τώρα που γνωρίζουμε τις παραμέτρους εύκολα δημιουργείται πλέον ο κανόνας απόφασης κατάταξης παρατηρήσεων σε μια απο τις δύο κατηγορίες.

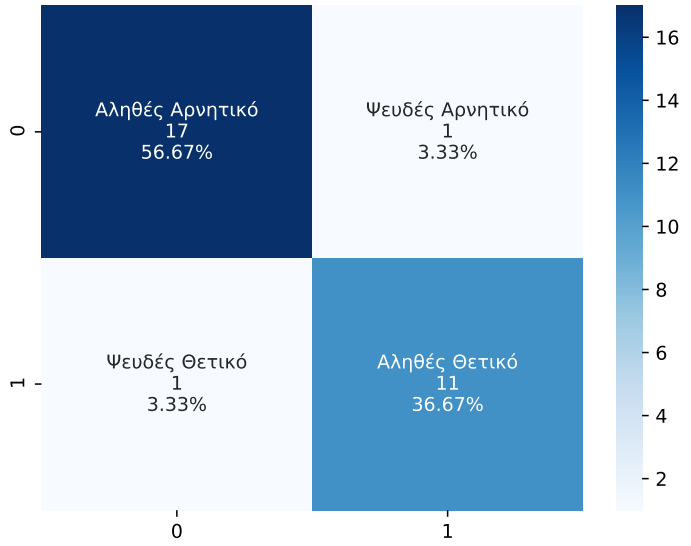
Αλγόριθμος 5.2.10: Δημιουργία Συνάρτησης Διαχωρισμού G

```
def g(x, mu1, mu2, sigma1, sigma2) :
    num = student_lhd(x,mu1,sigma1,p=3,n=3)/student_lhd(x,mu2,sigma2,p=
    ↪  3,n=3)
    return np.log(num)
```

Όπου mu1,mu2,sigma1,sigma2 είναι τα $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2$ (οι εκτιμήσεις των παραμέτρων) και $\boldsymbol{x}$ η παρατήρηση που πρέπει να καταταχθεί σε έναν απο τους δύο πληθυσμούς : Είδη φυτών Versicolour και Virginica. Η συνάρτηση `g(x, mu1, mu2, sigma1, sigma2)` δίνει την πρόβλεψη του επιπέδου για το δεδομένο $\boldsymbol{x}$. Αυτή τη συνάρτηση θα χρησιμοποιήσουμε και θα την εφαρμόσουμε στο πειραματικό πλήθος (test set) για την εκτίμηση των ομάδων της κάθε μιας παρατήρησης. Το αποτέλεσμα είναι ένας πίνακας που μας δείχνει πόσο ικανοποιητικά εκτίμησε η μέθοδος τα πειραματικά δεδομένα.

Αλγόριθμος 5.2.11: Πίνακας Σύγχυσης για Ομάδες Κατανομών Student-t

```
res = lambda x : g(x, mean1_mle, mean2_mle, sigma1_mle, sigma2_mle)

y_pred = []
for i in range(X_test.shape[0]) :
    y_pred.append(res(X_test.iloc[i,:]))

y_pred = pd.Series(y_pred)
y_pred = (y_pred < 0)*1

#Εκτίμηση σωστών και εσφαλμένων κατατάξεων
from sklearn import metrics
metrics.confusion_matrix(y_test, y_pred)
```



Πίνακας 5.7: Πίνακας σύγχυσης για ομάδες κατανομών Student-t

Παρατηρείται πως το μοντέλο χρησιμοποιώντας κατανομές Student-t (μία για κάθε πληθυσμό φυτών, συγκεκριμένα για Versicolour και Virginica) κατέγραψε 17 παρατηρήσεις σωστά ως Versicolour (0) και 11 σωστά ως Virginica (1). Επιπρόσθετα εκτίμησε 1 τιμή ως Veriscolour ενω ήταν Virginica και 1 τιμή Virginica ενώ ήταν Versicolour. Συνολικά έκανε δύο λάθη σε ένα σύνολο 30 πειραματικών δεδομένων. Συνεπώς το ποσοστό σωστών ταξινομήσεων είναι 96.6%. Απο την άλλη το ποσοστό των λανθασμένων ταξινομήσεων είναι 3.3%. Εν κατακλείδι, παρατηρείται πως τα ποσοστά ορθών και λανθασμένων ταξινομήσεων στην περίπτωση εφαρμογής κατανομών Student-t είναι τα ίδια με αυτά της κατανομής Kotz.

5.2.3 Συμμετρική Κατανομή Laplace

Συνεχίζοντας στην τελευταία περίπτωση, για εφαρμογή Ελλειπτικών Laplace κατανομών στις δύο κατηγορίες, η μόνη αλλαγή στον κώδικα θα ήταν οι γραμμές που αφορούν την πιθανοφάνεια δοθέντος του συνόλου των δεδομένων. Οπότε ο αλγόριθμος θα γραφόταν ως :

Αλγόριθμος 5.2.12: Συναρτήσεις απο Κοινού Πιθανοφανειών για κάθε Επίπεδο Κατανομής Laplace

```
# Συνάρτηση Πυκνότητας Πιθανότητας κατανομής Laplace
def laplace_lhd(x, mu, Sigma, k):
    x, mu = np.array(x), np.array(mu)
    Sigma = np.array(Sigma)
    lhd = 2/((2*math.pi)**(k/2)*np.linalg.det(Sigma)**0.5) * (math.pi/(
    ↪  2*np.sqrt(2*(x-mu).T@np.linalg.inv(Sigma)@(x-mu)))**0.5 *
    ↪  np.exp(-np.sqrt(2*(x-mu).T@np.linalg.inv(Sigma)@(x-mu))) *
    ↪  1/(np.sqrt((x-mu).T@np.linalg.inv(Sigma)@(x-mu)/2))**((k-2)/2)
    return lhd

```

```
# Υπολογισμός (αρνητικής) απο κοινού Πιθανοφάνειας για πληθυσμούς Versicolour (lik1) και Virginica (lik2)
def lik1(x)  :
    n = len(loc1)
    m1 = np.array(x[:3])
    chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshape(3,3)
    s1 = chol@chol.T
    sq = []
    for i in loc1 :
        sq.append(laplace_lhd(x = dfx.iloc[i,:], mu = m1, Sigma = s1, k = 3))
    result = -(np.log(sq)).sum()
    return result


def lik2(x)  :
    n = len(loc2)
    m2 = x[:3]
    chol = np.array([[x[3],0,0],[x[4],x[5],0],[x[6],x[7],x[8]]]).reshape(3,3)
    s2 = chol@chol.T
    sq = []
    for i in loc2 :
        sq.append(laplace_lhd(x = dfx.iloc[i,:], mu = m2, Sigma = s2, k = 3))
    result = - (np.log(sq)).sum()
    return result


ch1 = np.linalg.cholesky(sigma1)
ch2 = np.linalg.cholesky(sigma2)

idx1 = np.flatnonzero(ch1)
idx2 = np.flatnonzero(ch2)
```

Τα αποτελέσματα της μεθόδου δίνονται παρακάτω :

Αλγόριθμος 5.2.13: Αριθμητική Μέθοδος Εύρεσης ε.μ.π. Εκτιμητών για την Κατανομή Laplace

```
x01 = list(mean1) + list(ch1.flatten()[idx1])
x02 = list(mean2) + list(ch2.flatten()[idx2])

opt1 = sc.optimize.minimize(lik1, x0 = x01, method = "Nelder-Mead", tol = 1e-3)
```

```
opt2 = sc.optimize.minimize(lik2, x0 = x02, method = "Nelder-Mead", tol
↪  = 1e-3)
```

Αποτελέσματα του κώδικα για το πρώτο επίπεδο :

Πίνακας 5.8: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Laplace στον Πληθυσμό Versicolour

```
>>> opt1.success
True

>>> opt1.message
'Optimization terminated successfully.'

>>> opt1.x
array([5.79992735, 2.70012331, 4.10002897, 0.47292418, 0.11210666,
       0.28376019, 0.2719401 , 0.12220991, 0.30873124])
```

Απο τον πίνακα (5.8), $\hat{\boldsymbol{\mu}}_1 = (5.79992735, 2.70012331, 4.10002897)^T$ και $\hat{\boldsymbol{\Sigma}}_1$ με :

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.47292418 & 0 & 0 \\ 0.11210666 & 0.28376019 & 0 \\ 0.2719401 & 0.1222099 & 0.30873124 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^T = \begin{pmatrix} 0.22365728 & 0.05301795 & 0.12860705 \\ 0.05301795 & 0.09308775 & 0.0651646 \\ 0.12860705 & 0.0651646 & 0.18420166 \end{pmatrix}$$

Ενώ για το δεύτερο επίπεδο :

Πίνακας 5.9: Αριθμητικά αποτελέσματα εφαρμογής του αλγορίθμου για κατανομή Laplace στον Πληθυσμό Virginica

```
>>> opt2.success
True

>>> opt2.message
'Optimization terminated successfully.'

>>> opt2.x
array([ 6.50003031,  3.00016542,  5.49998208,  0.742026680,
        0.166297333,  0.298450632,  0.506672093, -0.00497396462,
        0.263083027])
```

Απο το opt2.x του πίνακα (5.9), $\hat{\boldsymbol{\mu}}_2 = (6.50003031, 3.00016542, 5.49998208)^T$ και $\hat{\boldsymbol{\Sigma}}_2$ με :

$$\hat{\boldsymbol{A}} = \begin{pmatrix} 0.742026680 & 0 & 0 \\ 0.166297333 & 0.298450632 & 0 \\ 0.506672093 & -0.00497396462 & 0.263083027 \end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^T = \begin{pmatrix} 0.55060359 & 0.12339706 & 0.37596421 \\ 0.12339706 & 0.11672758 & 0.08277373 \\ 0.37596421 & 0.08277373 & 0.32595403 \end{pmatrix}$$

Ο κανόνας απόφασης εφόσον έχουμε εκτιμήσει τις παραμέτρους Laplace διαμορφώνεται ως :

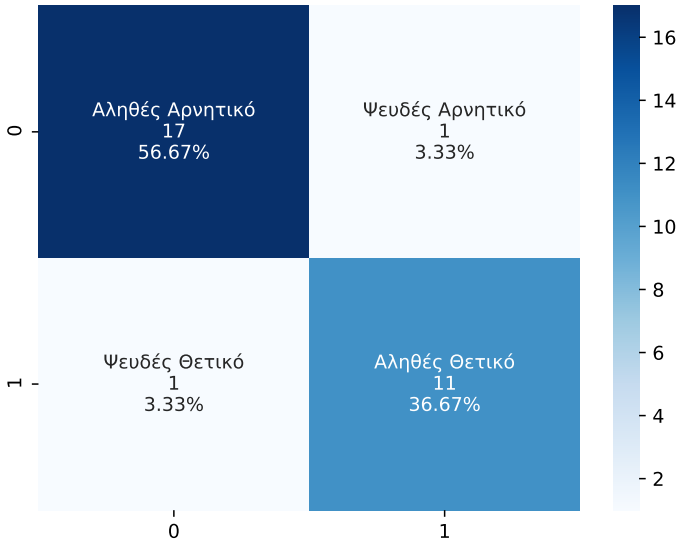
Αλγόριθμος 5.2.14: Δημιουργία Συνάρτησης Διαχωρισμού G

```
def g(x, mu1, mu2, sigma1, sigma2) :
    num = laplace_lhd(x = x, mu = mu1, sigma = sigma1, k =
    ↪  3)/laplace_lhd(x = x, mu = mu2, sigma = sigma2, k = 3)
    return np.log(num)
```

Όπου mu1,mu2,sigma1,sigma2 είναι τα $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2$ (οι εκτιμήσεις των παραμέτρων) και $\boldsymbol{x}$ η παρατήρηση που πρέπει να καταταχθεί σε έναν απο τους δύο πληθυσμούς : Είδη φυτών Versicolour και Virginica. Η συνάρτηση `g(x, mu1, mu2, sigma1, sigma2)` δίνει την πρόβλεψη του επιπέδου για το δεδομένο $\boldsymbol{x}$. Αυτή τη συνάρτηση θα χρησιμοποιήσουμε και θα την εφαρμόσουμε στο πειραματικό πλήθος (test set) για την εκτίμηση των ομάδων της κάθε μιας παρατήρησης. Το αποτέλεσμα είναι ένας πίνακας που μας δείχνει πόσο ικανοποιητικά εκτίμησε η μέθοδος τα πειραματικά δεδομένα.

Αλγόριθμος 5.2.15: Πίνακας Σύγχυσης για Ομάδες Κατανομών Laplace

```
es = lambda x : g(x, mean1_mle, mean2_mle, sigma1_mle, sigma2_mle)

y_pred = []
for i in range(X_test.shape[0]) :
    y_pred.append(res(X_test.iloc[i,:]))

y_pred = pd.Series(y_pred)
y_pred = (y_pred < 0)*1

#Εκτίμηση σωστών και εσφαλμένων κατατάξεων
from sklearn import metrics
metrics.confusion_matrix(y_test, y_pred)
```



Πίνακας 5.10: Πίνακας σύγχυσης για ομάδες κατανομών Ελλειπτικών Laplace

Παρατηρείται πως το μοντέλο χρησιμοποιώντας κατανομές Laplace (μία για κάθε πληθυσμό φυτών, συγκεκριμένα για Versicolour και Virginica) κατέγραψε 17 παρατηρήσεις σωστά ως Versicolour (0) και 11 σωστά ως Virginica (1). Επιπρόσθετα εκτίμησε 1 τιμή ως Veriscolour ενω ήταν Virginica και 1 τιμή Virginica ενώ ήταν Versicolour. Συνολικά έκανε δύο λάθη σε ένα σύνολο 30 πειραματικών δεδομένων. Συνεπώς το ποσοστό σωστών ταξινομήσεων είναι 96.6%. Απο την άλλη το ποσοστό των λανθασμένων ταξινομήσεων είναι 3.3%. Εν κατακλείδι, παρατηρείται πως τα ποσοστά ορθών και λανθασμένων ταξινομήσεων στην περίπτωση εφαρμογής κατανομών Laplace είναι τα ίδια με αυτά της κατανομής Kotz και Student-t με συνέπεια να μην παρουσιάζονται διαφορές στην συμπερασματολογία των τριών μοντέλων.

5.2.4 Τελικά Συμπεράσματα

Όπως είδαμε, προσαρμόζοντας Ελλειπτικές κατανομές στους πληθυσμούς των δεδομένων Iris, τα μοντέλα Διαχωριστικής Ανάλυσης για την κάθε μία κατανομή δεν έδωσαν διαφορετικά αποτελέσματα στις πιθανότητες σωστών και λανθασμένων κατηγοριοποιήσεων στα πειραματικά δεδομένα (test set). Ωστόσο οι εκτιμήσεις των διανυσματικών μέσων αλλά και των $\boldsymbol{\Sigma}$ έδωσαν διαφορετικές τιμές μεταξύ των Ελλειπτικών κατανομών. Αν και παρουσιάστηκαν οι ε.μ.π. των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ σε καθεμιά κατανομή και πληθυσμό, θα ήταν καλύτερο να δούμε τα αποτελέσματα συγκεντρωτικά για να διακρίνουμε ευκολότερα τις διαφορές στις εκτιμήσεις των παραμέτρων στα τρία συνολικά μοντέλα Kotz, Laplace και Student-t. Συγκεκριμένα θα παρουσιαστούν οι πίνακες των μέσων αλλά και των συνδιακυμάνσεων (που προκύπτουν απο τους πίνακες $\boldsymbol{\Sigma}$).

		Sepal L.	Sepal W.	Petal L.
Laplace	Versicolour	5.799 927	2.700 123	4.100 028
	Virginica	6.500 030	3.000 165	5.499 982
Student	Versicolour	5.876 778	2.799 547	4.224 582
	Virginica	6.537 355	2.967 226	5.512 813
Kotz	Versicolour	5.800 044	2.699 997	4.099 980
	Virginica	6.500 017	3.000 049	5.500 031

Πίνακας 5.11: Διανυσματικοί μέσοι για κάθε ελλειπτική κατανομή και πληθυσμό

Χρησιμοποιώντας χρώματα για την ανάδειξη των τιμών στους πίνακες (κλειστόχρωμα : μικρές τιμές, ανοιχτόχρωμα : μεγάλες τιμές), απο τον πίνακα (5.11) με τα διανύσματα των μέσων για κάθε κατανομή και πληθυσμό, βλέπουμε πως παρόμοια αποτελέσματα έχουμε για όλες τις κατανομές που προσαρμόστηκαν στους πληθυσμούς. Γενικά δεν διακρίνονται αξιοσημείωτες διαφορές.

			Sepal L.	Sepal W.	Petal L.
Laplace	Versicolour	Sepal L.	0.223 657	0.053 018	0.128 607
		Sepal W.	0.053 018	0.093 088	0.065 165
		Petal L.	0.128 607	0.065 165	0.184 202
	Virginica	Sepal L.	0.550 604	0.123 397	0.375 964
		Sepal W.	0.123 397	0.116 728	0.082 774
		Petal L.	0.375 964	0.082 774	0.325 954
Student	Versicolour	Sepal L.	0.472 179	0.134 830	0.277 582
		Sepal W.	0.134 830	0.175 327	0.154 900
		Petal L.	0.277 582	0.154 900	0.362 498
	Virginica	Sepal L.	1.012 369	0.244 233	0.716 017
		Sepal W.	0.244 233	0.213 121	0.177 418
		Petal L.	0.716 017	0.177 418	0.629 020
Kotz	Versicolour	Sepal L.	0.238 427	0.066 603	0.142 919
		Sepal W.	0.066 603	0.093 923	0.079 079
		Petal L.	0.142 919	0.079 079	0.190 055
	Virginica	Sepal L.	0.541 762	0.131 443	0.387 257
		Sepal W.	0.131 443	0.117 072	0.087 343
		Petal L.	0.387 257	0.087 343	0.344 760

Πίνακας 5.12: Πίνακες Διακυμάνσεων Συνδιακυμάνσεων για κάθε κατανομή και επίπεδο

Χρησιμοποιώντας χρώματα για την ανάδειξη των τιμών στους πίνακες (κλειστόχρωμα συνεπάγεται μικρές τιμές ενώ ανοιχτόχρωμα σημαίνει μεγάλες τιμές), απο τον πίνακα (5.12) φαίνεται

πως οι περισσότερες σχετικά μεγάλες εκτιμήσεις των συνδιακυμάνσεων βρέθηκαν στην ελλειπτική κατανομή Student και συγκεκριμένα στον πληθυσμό Virginica (κυρίως στην διακύμανση της Sepal L.), χωρίς όμως να παρουσιάζεται κάποια αρκετά σημαντική διαφορά απο τις άλλες κατανομές.

Βιβλιογραφία

L. Baringhaus and N. Henze. Limit distributions for mardia's measure of multivariate skewness. *Ann. Statist.*, 20:1889–1902, 1992.

N.H. Bingham and R. Kiesel. Semi-parametric modelling in finance: theoretical foundation. *Quantitative Finance*, 2:241–250, 2005.

T. Cacoulos and M. Koutras. Quadratic forms in spherical random variables : Generalized noncentral χ^2 distribution. *Naval Research Logistics Quarterly*, 31:447–461, 1984.

S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.

W Daojing, C. Zhang, and X. Zhao. Multivariate laplace filter: a heavy-tailed model for target tracking. *Proceedings of the 19th International Conference on Pattern Recognition: FL*, 2008.

G. R. Ducharme and P. Milasevic. Spatial median and directional data. *Biometrika*, 74: 212–215, 1987.

K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London, 1990.

E. Gomez, M. A. Gomez-Villegas, and J. M. Marin. A multivariate generalization of the power exponential family of distributions. *Commun. Statist.-Theory Meth*, 27:589–600, 1998.

W. S. Gosset. The probable error of a mean. *Biometrica*, 6:1–25, 1908.

J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35:414–415, 1948.

P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

D. Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya A*, 32:419–430, 1970.

M. Koutras. On the generalized noncentral chi-squared distribution induced by an elliptical gamma law. *Biometrica*, 73:528–532, 1986.

M. Koutras. On the performance of the linear discriminant function for spherical distributions. *Journal of Multivariate Analysis*, 21:1–12, 1987.

T.J. Kozubowski and A.K. Panorska. Multivariate geometric stable distributions in financial applications. *Math. Comput. Modelling*, 29:83–92, 1999.

D. N. Naik. *Multivariate medians: a review. In Probability and Statistics.* (Edited by S. K. Basu and B. K. Sinha). Publishing House, New Delhi, India, 1993.

C. R. Rao. Methodology based on the l_1-norm in statistical inference. *Sankhya Ser.*, 50: 289–313, 1988.

S. Simoni. *Miscellaneous real multivariate distributions. In Distributions in Statistics: Continuous Multivariate Distributions.* (Edited by N. L. Johnson and S. Kotz). Wiley, New York, 1968.

A. G. M. Steerneman and F. van Perlo-ten Kleij. Spherical distributions: Schoenberg (1938) revisited. *Expositiones Mathematicae*, 2005.

K. H. Wolfgang and S. Leopold. *Applied Multivariate Statistical Analysis.* 4η Έκδοση. Springer, Berlin Heidelberg, 2014. ISBN 978-3-662-45171-7.