ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΚΑΜΠΥΛΕΣ ΤΟΥ ANDREWS ΚΑΙ ΓΕΝΙΚΕΥΣΕΙΣ

Χαράλαμπος Α. Καπίδης

Διατριβή

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

Πειραιάς
Ιούνιος 2021

UNIVERSITY OF PIRAEUS
School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

ANDREWS CURVES AND THEIR GENERALIZATIONS

By

Charalampos A. Kapidis

Thesis

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

Piraeus, Greece
June 2021

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

Τα μέλη της Επιτροπής ήσαν:

- Μ. Κούτρας (Επιβλέπων)
- Σ. Μπερσίμης
- Β. Σεβρόγλου

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

Στο θείο μου, Παναγιώτη

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της διπλωματικής εργασίας μου, κ. Μ. Κούτρα, για τη συνεχή καθοδήγηση και έμπνευση που μου παρείχε καθ' όλη τη διάρκεια των σπουδών μου αλλά και της συγγραφής της εργασίας. Εκτιμώ ιδιαίτερα τις πολύτιμες συμβουλές και τις συστάσεις του, οι οποίες έχουν συνεισφέρει σημαντικά στην ποιότητα της διπλωματικής μου εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τους αναπληρωτές καθηγητές κ. Σ. Μπερσίμη και κ. Β. Σεβρόγλου για τις χρήσιμες παρεμβάσεις τους. Θα ήθελα επίσης να ευχαριστήσω τους καθηγητές του προγράμματος μεταπτυχιακών σπουδών για όλες τις πολύτιμες γνώσεις που με μετέδωσαν καθώς επίσης και τους φίλους και συμφοιτητές μου, οι οποίοι υπήρξαν σημαντικοί παράγοντες στην προσπάθειά μου, υποστηρίζοντάς με σε κάθε βήμα μου. Τέλος, ευχαριστώ θερμά τους γονείς μου για την στήριξη, την ενθάρρυνση και την πίστη τους σε μένα καθ' όλη τη διάρκεια των σπουδών μου.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως θέμα την παρουσίαση μίας σχετικά νέας τεχνικής για την αναπαράσταση πολυδιάστατων δεδομένων που παρουσίασε για πρώτη φορά ο Andrews το 1972. Η τεχνική αυτή ονομάζεται Καμπύλες Andrews και αφορά την αναπαράσταση κάθε παρατήρησης ενός συνόλου μέσω μιας καμπύλης. Η συγκεκριμένη τεχνική, λόγω της μαθηματικής φύσης της συνάρτησης που χρησιμοποιείται για τη χάραξη των καμπυλών, έχει κάποιες καλές ιδιότητες που αξιοποιούνται στις διάφορες χρήσεις της.

Στο πρώτο κεφάλαιο, γίνεται μία σύντομη εισαγωγή στην πολυμεταβλητή ανάλυση δεδομένων και στην οπτική αναπαράσταση πολυδιάστατων παρατηρήσεων. Επίσης γίνεται μία εισαγωγή στην τεχνική των Καμπυλών Andrews.

Στο δεύτερο κεφάλαιο δίνεται ο ορισμός της συνάρτησης του Andrews καθώς και των καμπυλών που προκύπτουν από τη χάραξη της συγκεκριμένης συνάρτησης. Ακόμα δίνονται κάποιες από τις πλέον σημαντικές και χρήσιμες ιδιότητες των Καμπυλών Andrews. Τέλος, γίνεται αναφορά στην κριτική που έχει δεχθεί η συγκεκριμένη τεχνική και προτείνονται λύσεις για τα διάφορα προβλήματα που προκύπτουν.

Στο τρίτο κεφάλαιο, ο αναγνώστης θα βρει διάφορες παραλλαγές της μεθόδου που έχουν δοθεί στη βιβλιογραφία. Πιο συγκεκριμένη γίνεται ένας διαχωρισμός σε τέσσερις κατηγορίες. Η πρώτη κατηγορία παραλλαγών αφορά εκείνες στις οποίες γίνεται χρήση της βάσης Fourier, όπως ακριβώς και στις Καμπύλες Andrews. Η δεύτερη κατηγορία αφορά εκείνες στις οποίες χρησιμοποιούνται οι τριγωνομετρικές συναρτήσεις χωρίς να γίνεται χρήση της βάσης Fourier. Στην τρίτη κατηγορία περιγράφονται οι παραλλαγές στις τρεις διαστάσεις και στην τέταρτη εκείνες στις οποίες δεν γίνεται χρήση των τριγωνομετρικών συναρτήσεων.

Στο τέταρτο κεφάλαιο δίνονται μερικές από τις πιο γνωστές χρήσεις της μεθόδου. Πιο συγκεκριμένα, η ομαδοποίηση παρατηρήσεων, η ταξινόμηση παρατηρήσεων, η εύρεση ακραίων τιμών, η εύρεση σημαντικών και μη μεταβλητών και η εύρεση της περιόδου σε δεδομένα χρονοσειρών.

Στο πέμπτο και τελευταίο κεφάλαιο, δίνονται τρεις εφαρμογές της μεθόδου των Καμπυλών Andrews. Αρχικά δίνεται η εφαρμογή σε μη ομαδοποιημένα δεδομένα, στη συνέχεια σε ομαδοποιημένα δεδομένα και τέλος σε δεδομένα χρονοσειρών. Στο κεφάλαιο αυτό δίνονται επίσης και συμπεράσματα σχετικά με τα τρία σύνολα δεδομένων που χρησιμοποιούνται.

Abstract

The aim of the current thesis is to present a relatively new technique in the representation of multidimensional data, which was first presented by Andrews in 1972. This technique is named Andrews Curves and concerns the representation of each observation of a dataset through a curve. Due to the mathematical nature of the function that is used to draw the curves, this technique presents some good qualities that can be reclaimed in various uses.

In the first chapter, a short introduction to the multidimensional data analysis and the visual representation of multidimensional data are presented. Also, there is an introduction to the Andrews Curves technique.

In the second chapter, the definition of the function of the Andrews Curves is given, as well as the curves that result from the visualization of this function. Additionally, some of the most important and useful properties of the Andrews Curves are presented. The criticism that this technique has received is also shown in this chapter and solutions to the issues arose are given.

In the third chapter, the reader can find multiple variations of the technique, that can be found in related papers. More specifically, four different categories are shown. The first category of variations is related to the usage of Fourier base, exactly like in the Andrews Curves. The second category concerns the variations in which the trigonometric functions are used without using the Fourier base. In the third category, the variations in three dimensions are described and in the fourth category the variations without the usage of trigonometric functions are shown.

In the fourth chapter, the most common uses of the method are described. More specifically, the clustering and classification of observations, the determination of outliers, the specification of significant and non-significant variables and the finding of period in time series data.

In the fifth chapter, three applications of the Andrews Curves are presented. The application in non-classified data, in classified data and time series data are given. In this chapter, the conclusions of these three applications are drawn.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Γραφημάτων

Κ Ε Φ Α Λ Α Ι Ο 1

Εισαγωγή

Όταν ένας ερευνητής μελετά ένα σύνολο δεδομένων, η πρώτη εργασία που πρέπει να ολοκληρώσει είναι η διερευνητική ανάλυση. Στο πλαίσιο αυτής της ανάλυσης καλείται να υπολογίσει κάποια από τα αριθμητικά περιγραφικά μέτρα για το σύνολο των δεδομένων καθώς και να προβεί σε κάποιας μορφής οπτικοποίηση των παρατηρήσεων.

Αν στο σύνολο περιέχεται μόνο μία μεταβλητή για κάθε άτομο (υπό μελέτη μονάδα) τότε αρκεί η χρήση τεχνικών της μονοδιάστατης ανάλυσης των δεδομένων. Αυτό σημαίνει πως ο ερευνητής πρέπει να υπολογίσει τα διάφορα μέτρα και να πραγματοποιήσει την αντίστοιχη οπτικοποίηση των παρατηρήσεων μόνο για μία μεταβλητή.

Αν, όμως, στο σύνολο δεδομένων υπάρχουν παραπάνω από μία μεταβλητές, τότε ο ερευνητής θα πρέπει να αποφασίσει με ποιο τρόπο θέλει να προχωρήσει στην ανάλυσή του. Οι διαθέσιμοι τρόποι είναι δύο. Στην πρώτη περίπτωση θα αντιμετωπίσει κάθε μεταβλητή ξεχωριστά και ανεξάρτητα από τις υπόλοιπες και επομένως θα ακολουθήσει τόσες φορές τη διαδικασία της μονοδιάστατης ανάλυσης δεδομένων όσες είναι και οι μεταβλητές. Στην περίπτωση όμως που ο ερευνητής θέλει να αναλύσει περαιτέρω πως αλληλεπιδρούν μεταξύ τους οι μεταβλητές καθώς επίσης και ποιες σχέσεις αναπτύσσονται μεταξύ αυτών, τότε θα πρέπει να προχωρήσει σε πολυδιάστατη ανάλυση δεδομένων. Η πολυδιάστατη ανάλυση δεδομένων περιλαμβάνει ένα μεγάλο πλήθος από απλούς αριθμητικούς δείκτες αλλά και ένα μεγάλο πλήθος από συστηματικές μεθόδους για την αναπαράσταση των πολυδιάστατων δεδομένων.

Είναι εξαιρετικά σημαντικό να μπορεί κανείς να δει οπτικοποιημένα τα προς ανάλυση δεδομένα ώστε να μπορεί να καταλήξει σε συμπεράσματα που δεν θα ήταν εύκολο να βρεθούν με αριθμητικές μεθόδους. Παραδείγματα γραφικών μεθόδων αναπαράστασης πολυδιάστατων δεδομένων είναι τα διαγράμματα διασποράς (Scatter Plots) στις δύο ή ακόμα και στις τρεις διαστάσεις, οι χάρτες θερμότητας (Heatmaps), οι παράλληλες συντεταγμένες (Parallel Coordinate Plots), τα Star Plots και τα πρόσωπα του Chernoff (Chernoff Faces).

Στην παρούσα εργασία παρουσιάζεται μία σχετικά πιο νέα μέθοδος αναπαράστασης πολυδιάστατων δεδομένων που είναι οι Καμπύλες Andrews. Η μέθοδος των καμπυλών

Andrews, όπως την παρουσίασε ο Andrews το 1972, αφορά την αναπαράσταση κάθε παρατήρησης (μονάδας) ενός συνόλου δεδομένων με μία καμπύλη μέσω μιας συνάρτησης που ονομάζεται συνάρτηση του Andrews. Μέσα από το γράφημα με το σύνολο των καμπυλών, μία για κάθε πολυδιάστατη παρατήρηση, μπορεί κανείς να εξάγει πολλά και χρήσιμα συμπεράσματα για τα δεδομένα.

Στα επόμενα Κεφάλαια δίνονται ο ορισμός των καμπυλών μαζί με τις ιδιαίτερα χρήσιμες ιδιότητες που αυτές έχουν, οι οποίες κατατάσσουν τη μέθοδο σε μία από τις πλέον χρήσιμες για την αναπαράσταση πολυδιάστατων δεδομένων. Γίνεται απόδειξη κάποιων από τις ιδιότητες καθώς επίσης γίνεται αναφορά στην κριτική που έχει δεχθεί η μέθοδος των καμπυλών Andrews και παρουσιάζονται προτάσεις και λύσεις για τα προβλήματα που προκύπτουν κατά την χρήση τους.

Επιπρόσθετα δίνονται διάφορες παραλλαγές που έχουν εμφανιστεί στη βιβλιογραφία για τη μέθοδο των καμπυλών του Andrews. Η πιο ενδιαφέρουσα παραλλαγή είναι αυτή της χάραξης των καμπυλών στις τρεις διαστάσεις, μιας και έχει τη δυνατότητα να οδηγήσει στην εξαγωγή περισσότερων, και ίσως πιο δύσκολων να βρεθούν, συμπερασμάτων.

Παρουσιάζονται επίσης και οι κυριότερες χρήσεις της μεθόδου των καμπυλών Andrews, όπως είναι η ομαδοποίηση και ταξινόμηση παρατηρήσεων, ο εντοπισμός ακραίων τιμών και ο εντοπισμός των σημαντικών και μη μεταβλητών ενός συνόλου δεδομένων.

Τέλος, γίνεται εφαρμογή των καμπυλών Andrews σε τρία σύνολα δεδομένων με στόχο τόσο την αξιολόγηση της χρησιμότητας της μεθόδου όσο και την εξαγωγή συμπερασμάτων σχετικά με τα τρία σύνολα. Τα σύνολα δεδομένων αφορούν τόσο ομαδοποιημένα και μη ομαδοποιημένα πολυδιάστατα δεδομένα όσο και δεδομένα χρονοσειρών.

Κ Ε Φ Α Λ Α Ι Ο 2

Ορισμός, Ιδιότητες και Κριτική των Καμπυλών Andrews

2.1 Εισαγωγή

Σε αυτό το Κεφάλαιο δίνεται ο ορισμός της συνάρτησης του Andrews καθώς και ο ορισμός των αντίστοιχων καμπυλών. Επίσης, γίνεται απόδειξη μερικών από τις πλέον σημαντικές ιδιότητες των καμπυλών. Τέλος, παρουσιάζονται μερικά από τα σημεία στα οποία έχουν δεχθεί κριτική οι καμπύλες Andrews μαζί με προτάσεις για την αντιμετώπιση των προβλημάτων που εμφανίζονται.

2.2 Ορισμός των Καμπυλών Andrews

Έστω $\boldsymbol{x} = (x_1, x_2, x_3, x_4, \dots)$ μία πολυδιάστατη παρατήρηση. Ως συνάρτηση του Andrews ορίζεται το εσωτερικό γινόμενο του διανύσματος $\boldsymbol{x}$ με το διάνυσμα της βάσης Fourier που περιέχει τόσους όρους όσα είναι και τα στοιχεία του διανύσματος $\boldsymbol{x}$. Ο τύπος που δίνει τη συνάρτηση του Andrews είναι ο εξής

$$f_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \boldsymbol{\Phi_t} \rangle = \frac{x_1}{\sqrt{2}} + x_2 \cdot \sin(t) + x_3 \cdot \cos(t) + x_4 \cdot \sin(2t) + x_5 \cdot \cos(2t) + \dots \quad (2.1)$$

όπου $-\pi < t < \pi$.

Η συνάρτηση (2.1) ορίστηκε για πρώτη φορά από τον Andrews (1972). Στην ίδια εργασία γίνεται και η πρώτη αναφορά στις ιδιότητες της συνάρτησης καθώς και σε εφαρμογές της.

Ως βάση Fourier ορίζεται το διάνυσμα

$$\boldsymbol{\Phi_t} = \left(\frac{1}{\sqrt{2}}, \sin(t), \cos(t), \sin(2t), \cos(2t), \dots\right)$$

όπου $-\pi < t < \pi$.

Σημειώνεται ότι, κάθε περιοδική συνάρτηση, με περίοδο T, μπορεί να γραφεί ως γραμμικός συνδυασμός των στοιχείων του συνόλου

$$B = \left\{1, sin\left(\frac{2\pi}{T}x\right), cos\left(\frac{2\pi}{T}x\right), sin\left(\frac{4\pi}{T}x\right), cos\left(\frac{4\pi}{T}x\right), \dots, sin\left(\frac{n\pi}{T}x\right), cos\left(\frac{n\pi}{T}x\right)\right\}.$$

Επομένως, το σύνολο B αποτελεί μία βάση για το σύνολο των περιοδικών συναρτήσεων, με περίοδο T.

Το γινόμενο μεταξύ δύο περιοδικών συναρτήσεων f και g, με περίοδο T, ορίζεται ως εξής

$$f(x) \cdot g(x) = \int_{-\frac{T}{2}}^{\frac{T}{2}} f(x)g(x)\,dx.$$

Υπολογίζοντας το γινόμενο ανά δύο όλων των στοιχείων της βάσης B, προκύπτουν οι εξής σχέσεις αν $k \neq m$

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} cos\left(\frac{2k\pi}{T}x\right) cos\left(\frac{2m\pi}{T}x\right) dx = 0, \int_{-\frac{T}{2}}^{\frac{T}{2}} sin\left(\frac{2k\pi}{T}x\right) sin\left(\frac{2m\pi}{T}x\right) dx = 0$$

και

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} cos\left(\frac{2k\pi}{T}x\right) sin\left(\frac{2m\pi}{T}x\right) dx = 0.$$

Αντίστοιχα, αν $k = m$, οι σχέσεις που προκύπτουν είναι

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} cos^2\left(\frac{2k\pi}{T}x\right) dx = \frac{T}{2} \text{ και } \int_{-\frac{T}{2}}^{\frac{T}{2}} sin^2\left(\frac{2k\pi}{T}x\right) dx = \frac{T}{2}.$$

Από τις παραπάνω σχέσεις γίνεται φανερό ότι τα στοιχεία της βάσης B είναι μεταξύ τους κάθετα και επομένως η βάση B είναι μία ορθογώνια βάση. Γίνεται επίσης φανερό ότι τα στοιχεία της βάσης δεν έχουν μέτρο ίσο με ένα και επομένως η βάση δεν είναι ορθοκανονική. Για να μετατραπεί η βάση B σε ορθοκανονική θα πρέπει να διαιρεθεί κάθε στοιχείο της με το μέτρο του. Ο πρώτος όρος θα πρέπει να διαιρεθεί με $\sqrt{T}$ και οι υπόλοιποι όροι με $\sqrt{\frac{T}{2}}$.

Μετά τον παραπάνω μετασχηματισμό η βάση B παίρνει την εξής μορφή

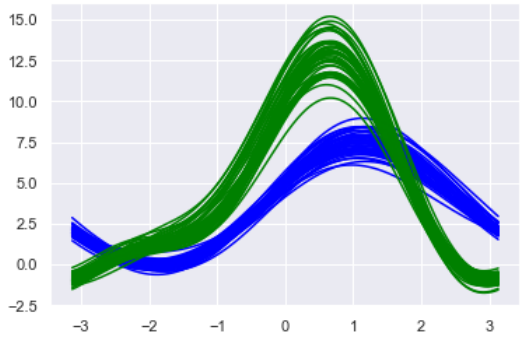
$$B' = \{\frac{1}{\sqrt{T}}, \sqrt{\frac{2}{T}} sin\left(\frac{2\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{2\pi}{T}x\right), \sqrt{\frac{2}{T}} sin\left(\frac{4\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{4\pi}{T}x\right), \dots,$$
$$\sqrt{\frac{2}{T}} sin\left(\frac{n\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{n\pi}{T}x\right)\}.$$

Στην περίπτωση των καμπυλών Andrews η περίοδος που μας ενδιαφέρει είναι $T = 2$ και επομένως παράγεται το διάνυσμα $\boldsymbol{\Phi_t}$ το οποίο αποτελεί ορθοκανονική βάση. Η ιδιότητα αυτή είναι εξαιρετικά χρήσιμη για την απόδειξη των ιδιοτήτων που φέρουν οι καμπύλες Andrews.

Η γραφική παράσταση της συνάρτησης του Andrews $f_x(t)$, που είναι κυματοειδούς μορφής, ορίζεται ως Καμπύλη Andrews. Δεδομένου ενός συνόλου πολυδιάστατων παρατηρήσεων, μπορεί κανείς να αναπαραστήσει κάθε μία από αυτές μέσω μιας καμπύλης και να παραχθεί το σύνολο των καμπυλών Andrews για όλες τις παρατηρήσεις.

Επομένως, οι καμπύλες Andrews αποτελούν μία μέθοδο για την γραφική αναπαράσταση πολυδιάστατων δεδομένων. Από την γραφική αναπαράσταση του συνόλου των δεδομένων, μπορούν να εξαχθούν χρήσιμα συμπεράσματα είτε για το σύνολο είτε για μέρος των δεδομένων αυτών.

Οι καμπύλες του Andrews για το σύνολο δεδομένων Iris, όπως αυτό περιγράφεται στο Παράρτημα 1, φαίνεται στο Γράφημα 2.1.



Γράφημα 2.1
Καμπύλες Andrews για το σύνολο δεδομένων Iris

Οι καμπύλες χρώματος πράσινου αντιστοιχούν στις παρατηρήσεις που ανήκουν στον τύπο Iris-virginica, ενώ εκείνες χρώματος μπλε αντιστοιχούν στον τύπο Iris-setosa. Είναι λοιπόν προφανές ότι μέσω των γραφικών παραστάσεων των καμπυλών μπορεί κανείς να ερευνήσει πιθανές ομαδοποιήσεις στα δεδομένα.

2.3 Ιδιότητες των Καμπυλών Andrews

Οι καμπύλες του Andrews παρουσιάζουν κάποιες πολύ σημαντικές μαθηματικές ιδιότητες που χρησιμοποιούνται για την εξαγωγή συμπερασμάτων αναφορικά με τα δεδομένα. Σε αυτή την παράγραφο δίνονται κάποιες από αυτές τις ιδιότητες.

1. Οι καμπύλες του Andrews διατηρούν το μέσο. Αν $\bar{\boldsymbol{x}}$ είναι ο μέσος των παρατηρήσεων, τότε αυτός αναπαρίσταται από την καμπύλη $f_{\bar{x}}(t)$, η οποία είναι ίση με το μέσο των καμπυλών όλων των παρατηρήσεων δηλαδή ισχύει η ισότητα $f_{\bar{x}}(t) = \bar{f}_{x}(t)$. Η παραπάνω ιδιότητα μπορεί εύκολα να διαπιστωθεί ως εξής:

$$f_{\bar{x}}(t) = \frac{\bar{x}_1}{\sqrt{2}} + \bar{x}_2 \cdot \sin(t) + \bar{x}_3 \cdot \cos(t) + \bar{x}_4 \cdot \sin(2t) + \bar{x}_5 \cdot \cos(2t) + \ldots \Rightarrow$$

$$f_{\bar{x}}(t) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{1i}}{\sqrt{2}} + \frac{1}{n}\sum_{i=1}^{n} x_{2i} \cdot \sin(t) + \frac{1}{n}\sum_{i=1}^{n} x_{3i} \cdot \cos(t) + \frac{1}{n}\sum_{i=1}^{n} x_{4i} \cdot \sin(2t) + \\ + \frac{1}{n}\sum_{i=1}^{n} x_{5i} \cdot \cos(2t) + \ldots \Rightarrow$$

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_{1i}}{\sqrt{2}} + x_{2i} \cdot \sin(t) + x_{3i} \cdot \cos(t) + x_{4i} \cdot \sin(2t) + x_{5i} \cdot \cos(2t) + \ldots\right) \Rightarrow$$

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t) \Rightarrow f_{\bar{x}}(t) = \bar{f}_{x}(t).$$

Για μία γραφική επίδειξη της συγκεκριμένης ιδιότητας θα χρησιμοποιηθούν οι δύο παρατηρήσεις από το σύνολο δεδομένων Iris που φαίνονται στον Πίνακα 2.1.

	Sepal Length	**Sepal Width**	**Petal Length**	**Petal Width**	**Class**
Παρατήρηση 1	5.1	3.5	1.4	0.2	Iris-setosa
Παρατήρηση 2	5.9	3.0	5.1	1.8	Iris-virginica

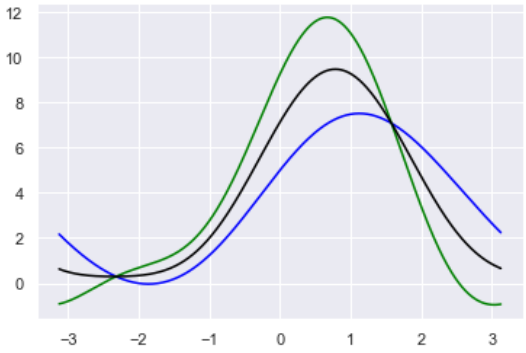
Πίνακας 2.1
Η πρώτη παρατήρηση του είδους Iris-setosa και η τελευταία του είδους Iris-virginica

Ο μέσος των δύο παρατηρήσεων για κάθε μία από τις τέσσερις μεταβλητές δίνεται στον Πίνακα 2.2.

	Sepal Length	**Sepal Width**	**Petal Length**	**Petal Width**
Μέσος Παρατηρήσεων 1 και 2	5.5	3.25	3,25	1,0

Πίνακας 2.2
Ο μέσος των παρατηρήσεων του Πίνακα 2.1

Οι καμπύλες Andrews για τις δύο παρατηρήσεις και για τον μέσο τους φαίνονται στο Γράφημα 2.2. Η καμπύλη με το μπλε χρώμα αντιστοιχεί στην Παρατήρηση 1, εκείνη με το πράσινο χρώμα στην Παρατήρηση 2 και εκείνη με το μαύρο χρώμα στο μέσο των δύο Παρατηρήσεων. Όπως αναμενόταν η καμπύλη του μέσου βρίσκεται για όλες τις τιμές του t, $-\pi < t < \pi$, μεταξύ των καμπυλών που αντιστοιχούν στις δύο παρατηρήσεις και πιο συγκεκριμένα σε ίση απόσταση από αυτές.



Γράφημα 2.2
Οι καμπύλες Andrews για τις παρατηρήσεις του
Πίνακα 2.1 καθώς και του μέσου τους

2. Οι καμπύλες του Andrews διατηρούν τις αποστάσεις. Σύμφωνα με την ιδιότητα αυτή, η απόσταση μεταξύ των καμπυλών που αντιστοιχούν σε δύο πολυδιάστατες παρατηρήσεις είναι ανάλογη της Ευκλείδειας απόστασης των παρατηρήσεων. Πιο συγκεκριμένα, αν $\boldsymbol{x}$ και $\boldsymbol{y}$ είναι δύο n-διάστατες παρατηρήσεις, τότε η απόσταση των καμπυλών συνδέεται με την Ευκλείδεια απόσταση των παρατηρήσεων μέσω της σχέσης:

$$\left\|f_{\boldsymbol{x}}(t) - f_{\boldsymbol{y}}(t)\right\|_{L_2}^2 = \int_{-\pi}^{\pi}\left[f_{\boldsymbol{x}}(t) - f_{\boldsymbol{y}}(t)\right]^2 dt = \pi \cdot \sum_{i=1}^{n}(x_i - y_i)^2 = \pi \cdot \|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Λόγω της παραπάνω ιδιότητας, είναι φανερό ότι σημεία που βρίσκονται κοντά μεταξύ τους στον Ευκλείδειο χώρο, θα απεικονίζονται σε κοντινές καμπύλες ενώ σημεία που βρίσκονται μακριά το ένα από το άλλο, θα απεικονίζονται σε μακρινές καμπύλες.

Για παράδειγμα, ας θεωρήσουμε τις δύο δισδιάστατες παρατηρήσεις $\boldsymbol{x} = (1,1)$ και $\boldsymbol{y} = (2,2)$. Η Ευκλείδεια απόσταση των δύο αυτών παρατηρήσεων υπολογίζεται ως εξής:

$$\|x - y\| = \sqrt{(2-1)^2 + (2-1)^2} = \sqrt{2}.$$

Οι συναρτήσεις του Andrews για τις παρατηρήσεις $\boldsymbol{x}$ και $\boldsymbol{y}$ είναι οι εξής

$$f_{\boldsymbol{x}}(t) = \frac{1}{\sqrt{2}} + sin(t),\ f_{\boldsymbol{y}}(t) = \frac{2}{\sqrt{2}} + 2 \cdot sin(t),\ -\pi < t < \pi,$$

οπότε είναι

$$\left\|f_{\boldsymbol{x}}(t) - f_{\boldsymbol{y}}(t)\right\|_{L_2}^2 = \int_{-\pi}^{\pi}\left(-\frac{1}{\sqrt{2}} - sin\, t\right)^2 dt = 2\pi$$

από όπου προκύπτει ότι

$$\left\|f_{\boldsymbol{x}}(t) - f_{\boldsymbol{y}}(t)\right\|_{L_2} = \sqrt{2\pi} = \sqrt{\pi}\,\|x - y\|\,.$$

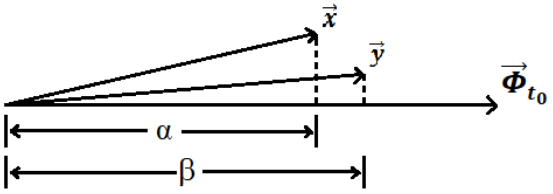
3. Η χρήση των καμπυλών Andrews αποδίδει μονοδιάστατες προβολές. Η ιδιότητα αυτή αναφέρεται στο γεγονός ότι αν $\boldsymbol{x}$ είναι μία πολυδιάστατη παρατήρηση, τότε για συγκεκριμένο

$t = t_0$, η τιμή της συνάρτησης $f_{\boldsymbol{x}}(t_0)$ είναι ανάλογη του μήκους της προβολής του διανύσματος $\boldsymbol{x}$ πάνω στο διάνυσμα $\boldsymbol{\Phi_{t_0}}$, το οποίο ορίζεται ως

$$\boldsymbol{\Phi_{t_0}} = (\tfrac{1}{\sqrt{2}}, \sin(t_0), \cos(t_0), \sin(2t_0), \cos(2t_0), \ldots).$$

Έστω τώρα ότι $\boldsymbol{x}$ και $\boldsymbol{y}$ είναι δύο πολυδιάστατες παρατηρήσεις και α και β είναι τα μήκη των προβολών τους στο διάνυσμα $\boldsymbol{\Phi_{t_0}}$ (βλ. Γράφημα 2.3). Αν τα $f_{\boldsymbol{x}}(t_0)$ και $f_{\boldsymbol{y}}(t_0)$ είναι κοντινές τιμές, τότε και τα α και β θα έχουν κοντινές τιμές, και αντίστοιχα αν τα $f_{\boldsymbol{x}}(t_0)$ και $f_{\boldsymbol{y}}(t_0)$ έχουν πολύ διαφορετικές τιμές, τότε και τα α και β θα είναι μακριά.

Η ιδιότητα αυτή είναι εξαιρετικά χρήσιμη γιατί μπορεί να αποκαλύψει μοτίβα ή ομαδοποιήσεις με παρατήρηση του μονοδιάστατου χώρου αντί του πολυδιάστατου που ανήκουν τα δεδομένα, για τον οποίο δε μπορούμε να έχουμε απεικόνιση.



Γράφημα 2.3
Γραφική απεικόνιση των διανυσμάτων $\boldsymbol{\Phi_{t_0}}$, $\boldsymbol{x}$ και $\boldsymbol{y}$

4. Οι καμπύλες του Andrews διατηρούν τις διασπορές (υπό προϋποθέσεις). Έστω $\boldsymbol{x}$ μία n-διάστατη παρατήρηση. Αν οι n μεταβλητές που υπάρχουν στα δεδομένα είναι ασυσχέτιστες και έχουν κοινή διασπορά, έστω σ^2, τότε η συνάρτηση του Andrews $f_{\boldsymbol{x}}(t)$ έχει διασπορά

$$Var\big(f_{\boldsymbol{x}}(t)\big) = \sigma^2(\tfrac{1}{2} + \sin^2(t) + \cos^2(t) + \sin^2(2t) + \cos^2(2t) + \ldots).$$

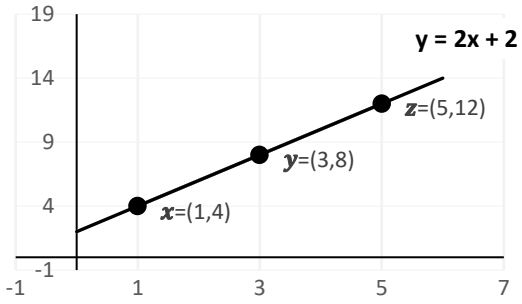
Αν ο αριθμός των μεταβλητών n είναι περιττός, τότε η παραπάνω διασπορά είναι ίση με $\frac{n\,\sigma^2}{2}$. Αντίστοιχα, αν ο αριθμός των μεταβλητών n είναι άρτιος, τότε η παραπάνω διασπορά είναι ίση με $\sigma^2(\frac{n-1}{2} + \sin^2\left(\frac{n}{2}\cdot t\right))$ και θα κινείται μεταξύ των αριθμών $\sigma^2\left(\frac{n-1}{2}\right)$ και $\sigma^2\left(\frac{n+1}{2}\right)$. Επομένως, η διασπορά της $f_{\boldsymbol{x}}(t)$ διατηρείται σταθερή για κάθε $-\pi < t < \pi$, όταν το n είναι περιττός και κινείται σε διάστημα πλάτους σ^2 όταν είναι άρτιος.

Η συγκεκριμένη ιδιότητα χρησιμοποιείται αρκετά σπάνια σε πραγματικές περιπτώσεις δεδομένων, αφού είναι εξαιρετικά σπάνιο φαινόμενο οι μεταβλητές να είναι πλήρως ασυσχέτιστες και να έχουν κοινή διασπορά.

5. Οι καμπύλες του Andrews διατηρούν τη γραμμικότητα μεταξύ των παρατηρήσεων. Έστω $\boldsymbol{x}$, $\boldsymbol{y}$ και $\boldsymbol{z}$ τρεις πολυδιάστατες παρατηρήσεις που βρίσκονται πάνω στην ίδια ευθεία. Αν η παρατήρηση $\boldsymbol{y}$ βρίσκεται μεταξύ των $\boldsymbol{x}$ και $\boldsymbol{z}$, δηλαδή $\boldsymbol{y} = \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{z}$, με $0 \le \lambda \le 1$, τότε οι καμπύλη $f_{\boldsymbol{y}}(t)$ που αντιστοιχεί στην παρατήρηση $\boldsymbol{y}$ θα βρίσκεται μεταξύ των καμπυλών

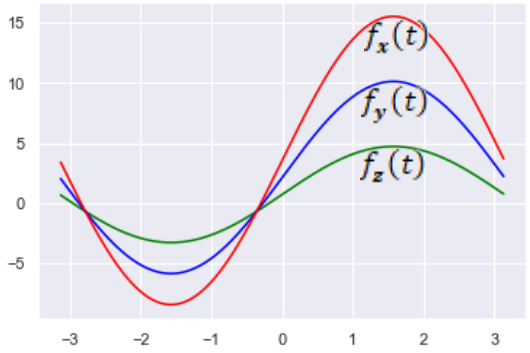
$f_x(t)$ και $f_z(t)$ που αντιστοιχούν στις παρατηρήσεις $\boldsymbol{x}$ και $\boldsymbol{z}$ αντίστοιχα και συγκεκριμένα θα ισχύει $f_y(t) = \lambda \cdot f_x(t) + (1-\lambda) \cdot f_z(t)$.

Για παράδειγμα, έστω οι τρεις δισδιάστατες παρατηρήσεις $\boldsymbol{x}$, $\boldsymbol{y}$ και $\boldsymbol{z}$, $\boldsymbol{x} = (1,4)$, $\boldsymbol{y} = (3,8)$ και $\boldsymbol{z} = (5,12)$ οι οποίες βρίσκονται επάνω στην ευθεία με εξίσωση $y = 2 \cdot x + 2$ όπως φαίνεται και στο Γράφημα 2.4.



Γράφημα 2.4
Γραφική απεικόνιση των σημείων $\boldsymbol{x}$, $\boldsymbol{y}$ και $\boldsymbol{z}$

Στο Γράφημα 2.5 φαίνονται οι καμπύλες Andrews για τις παρατηρήσεις $\boldsymbol{x}$, $\boldsymbol{y}$ και $\boldsymbol{z}$. Όπως γίνεται αντιληπτό, η καμπύλη $f_y(t)$ είναι μεταξύ των καμπυλών $f_x(t)$ και $f_z(t)$ σε ολόκληρο το διάστημα $(-\pi, \pi)$.



Γράφημα 2.5
Οι καμπύλες Andrews για τις παρατηρήσεις $\boldsymbol{x}$, $\boldsymbol{y}$ και $\boldsymbol{z}$

2.4 Κριτική για τις καμπύλες Andrews

Όπως συμβαίνει και με άλλες μεθόδους της πολυμεταβλητής ανάλυσης, έτσι και στην περίπτωση των καμπυλών του Andrews εμφανίζονται κάποια προβλήματα όταν χρησιμοποιούνται. Μερικά από αυτά θα αναφερθούν παρακάτω μαζί με πιθανές λύσεις ή προτάσεις.

α. Θα μπορούσε κανείς να πει πως η χρήση των καμπυλών είναι παρεμφερής με τη χρήση της μεθόδου Grand Tour, όπως αυτή περιγράφεται από τον D. Asimov (1985). Η μέθοδος Grand Tour είναι μία μέθοδος αναπαράστασης πολυδιάστατων δεδομένων μέσω ορθογώνιων προβολών σε μία ακολουθία από δισδιάστατους υποχώρους. Με την μέθοδο αυτή, μπορεί κανείς να δει τα δεδομένα από όλες τις οπτικές γωνίες αφού γίνεται προβολή πολυδιάστατων δεδομένων στους δύο άξονες. Η μέθοδος των καμπυλών του Andrews, δημιουργεί γραφήματα στους δύο άξονες, επομένως απεικονίζει με κάποιο τρόπο τα δεδομένα σε ένα επίπεδο, ωστόσο δεν δίνει τη δυνατότητα να τα δει κανείς από διάφορες οπτικές γωνίες.

Μία λύση, που προτείνεται για το παραπάνω πρόβλημα από τον Moustafa (2011), είναι να χρησιμοποιηθεί κατάλληλη περιστροφή είτε των δεδομένων είτε της βάσης που χρησιμοποιείται στη συνάρτηση του Andrews ώστε να μπορέσει κανείς να δει τα δεδομένα από εναλλακτικές οπτικές γωνίες. Οι τεχνικές λεπτομέρειες για την εφαρμογή της μεθόδου δίνονται στην εργασία του Moustafa (2011).

β. Μία από τις ιδιότητες των καμπυλών Andrews είναι ότι διατηρούν σταθερή την διασπορά με την προϋπόθεση ότι οι μεταβλητές που υπάρχουν στα δεδομένα είναι ασυσχέτιστες και με κοινή διασπορά σ^2. Η συγκεκριμένη υπόθεση δεν ισχύει σχεδόν πότε σε πραγματικά δεδομένα και επομένως είναι συχνό φαινόμενο να συμπεριλαμβάνονται σε ένα σύνολο δεδομένων είτε συσχετισμένες μεταβλητές είτε μεταβλητές με διαφορετική διασπορά είτε και τα δύο. Επομένως δεν είναι δυνατό να αξιοποιηθεί η ιδιότητα της διατήρησης της διασποράς, από τη στιγμή που δεν ισχύουν οι απαραίτητες προϋποθέσεις.

Μία πρόταση για να μπορεί κανείς να αξιοποιήσει την παραπάνω ιδιότητα είναι να κάνει μετατροπή των αρχικών μεταβλητών σε κύριες συνιστώσες και να χρησιμοποιήσει αυτές για τη σχεδίαση των καμπυλών, όπως αναφέρουν οι Khattree and Naik (2002). Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους, όμως δεν είναι βέβαιο ότι θα έχουν κοινή διασπορά. Για το πρόβλημα της μη σταθερής διασποράς ο Seber (1984) πρότεινε τη χρήση κανονικοποιημένων κύριων συνιστωσών. Οι κανονικοποιημένες κύριες συνιστώσες, έχουν εξ ορισμού κοινή διασπορά και είναι ασυσχέτιστες μεταξύ τους. Η ιδιότητα αυτή των κύριων συνιστωσών προέρχεται από την μέθοδο κατασκευής τους και χρησιμοποιώντας αυτές ως μεταβλητές στη χάραξη των καμπυλών επιτυγχάνεται η διατήρηση της διασποράς καθώς και η μη συσχέτιση.

γ. Όπως μπορεί κανείς να καταλάβει από τη μορφή της συνάρτησης $f_x(t)$, του τύπου (2.1), αν οι μεταβλητές εισαχθούν με δύο διαφορετικές σειρές, τότε οι δύο εικόνες που θα προκύψουν θα είναι διαφορετικές μεταξύ τους. Η διαφορά μεταξύ των δύο εικόνων μπορεί να είναι μεγάλη ή μικρή, όμως σίγουρα είναι υπαρκτή. Η διαφορετική σειρά, επομένως, δημιουργεί διαφορετικές καμπύλες και ως αποτέλεσμα θα οδηγήσει πιθανώς στην εξαγωγή διαφορετικών συμπερασμάτων. Η πρόταση που κάνει ο ίδιος ο Andrews, στην εργασία που παρουσίασε τη

χρήση των καμπυλών, είναι να γίνεται χρήση των μεταβλητών με βάση τη σημαντικότητα που έχουν για το συγκεκριμένο σύνολο δεδομένων. Άρα προτείνεται να εισάγονται πρώτες οι περισσότερο σημαντικές μεταβλητές δηλαδή εκείνες που περιέχουν περισσότερη πληροφορία για τα δεδομένα και στο τέλος οι λιγότερο σημαντικές, ώστε να επηρεάζεται η εικόνα κυρίως από αυτές που θεωρούνται σημαντικές.

Υπάρχουν σύνολα δεδομένων που περιέχουν μεγάλο αριθμό μεταβλητών, με τις περισσότερες από αυτές να είναι αρκετά σημαντικές. Σε τέτοιες περιπτώσεις, μπορεί να αποδειχθεί δύσκολη η κατασκευή των καμπυλών και η εξαγωγή συμπερασμάτων μιας και ο ερευνητής μπορεί να μην είναι σε θέση να αξιολογήσει πλήρως τη σημαντικότητα κάθε μεταβλητής ώστε να δημιουργήσει τη σειρά με την οποία θα εισάγει τις μεταβλητές στη συνάρτηση του Andrews.

Μία λύση για το παραπάνω πρόβλημα είναι, και πάλι, η χρήση των κύριων συνιστωσών και μάλιστα μόνο κάποιων από τις πρώτες κύριες συνιστώσες. Ο αριθμός των πρώτων κύριων συνιστωσών που θα χρησιμοποιηθούν αποφασίζεται από τον ερευνητή με γνώμονα το ποσοστό της πληροφορίας που θέλει να διατηρήσει. Ένα σύνηθες ποσοστό είναι της τάξεως του 80%. Η χρήση των κύριων συνιστωσών βοηθάει τόσο στο γεγονός ότι μειώνει τον αριθμό των μεταβλητών που θα χρησιμοποιηθούν στη χάραξη αλλά και στο γεγονός ότι δίνει τη σειρά με την οποία θα εισάγει κανείς τις τιμές στη συνάρτηση.

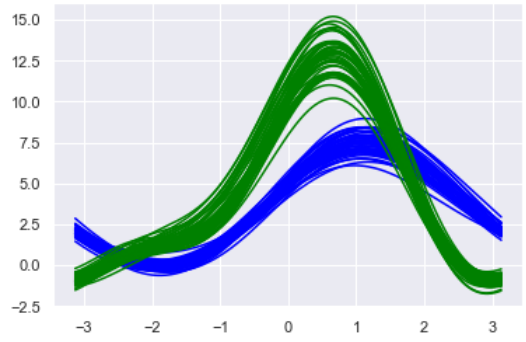
Η συγκεκριμένη πρόταση παρουσιάζεται αναλυτικότερα στην Παράγραφο 5.2.2 μέσω εφαρμογής των καμπυλών

δ. Έχει αναφερθεί, και ίσως είναι σχετικά ακριβές, ότι το ανθρώπινο μάτι παρατηρεί καλύτερα και με μεγαλύτερη ακρίβεια ότι συμβαίνει σε μία περιοχή γύρω από το μέσο ενός γραφήματος. Στην περίπτωση των καμπυλών Andrews, το ανθρώπινο μάτι θα αντιληφθεί καλύτερα ότι συμβαίνει γύρω από το $t = 0$. Όπως μπορεί να δει κανείς από τη μορφή της συνάρτησης του Andrews $f_x(t)$, στον τύπο (2.1), στο σημείο με $t = 0$ οι μισοί όροι του αθροίσματος γίνονται ίσοι με 0, αφού για το συγκεκριμένο σημείο ισχύει ότι $sin(0) = 0$. Αυτό σημαίνει ότι στο συγκεκριμένο σημείο οι μεταβλητές που έχουν χρησιμοποιηθεί σε άρτια σειρά και επομένως έχουν πολλαπλασιαστεί με τους ημιτονοειδείς όρους, δεν υπολογίζονται για την εύρεση των αντίστοιχων τιμών $f_x(0)$. Ως εκ τούτου τα όποια συμπεράσματα βγουν με βάση την εικόνα των καμπυλών γύρω από το 0, δεν θα τις έχουν λάβει υπόψη.

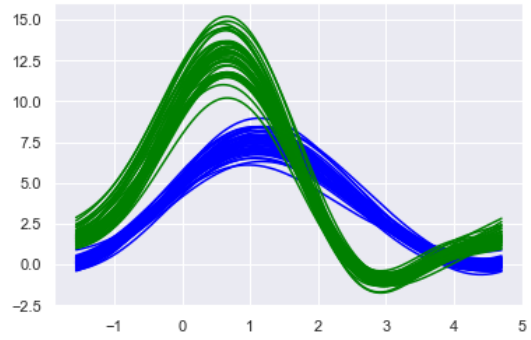
Μία απλή αλλά ταυτόχρονα ευφυής λύση είναι να γίνει χάραξη των καμπυλών Andrews και σε κάποιο διάστημα στο μέσο του οποίου να μην μηδενίζονται οι όροι που μηδενίζονται για $t = 0$. Το διάστημα που προτείνεται είναι το $\left[-\frac{\pi}{2}, \frac{3\pi}{2}\right]$ αντί του $[-\pi, \pi]$. Το μέσο του διαστήματος $\left[-\frac{\pi}{2}, \frac{3\pi}{2}\right]$ είναι το $t = \frac{\pi}{2}$ και σε αυτό το σημείο δε μηδενίζονται οι όροι που περιέχουν τα ημίτονα. Στο συγκεκριμένο σημείο, όμως, μηδενίζονται οι όροι που περιέχουν

συνημίτονο, αφού $cos\left(\frac{\pi}{2}\right) = 0$. Για το λόγο αυτό προτείνεται να χαράσσονται οι καμπύλες και στα δύο διαστήματα ώστε ο ερευνητής να μπορεί να βγάλει πιο ασφαλή συμπεράσματα για τα δεδομένα.

Στο Γράφημα 2.6 δίνονται οι καμπύλες Andrews για το σύνολο δεδομένων Iris στο διάστημα $[-\pi,\ \pi]$ και στο Γράφημα 2.7 δίνονται οι καμπύλες στο διάστημα $\left[-\frac{\pi}{2},\ \frac{3\pi}{2}\right]$. Στο μέσο του κάθε γραφήματος έχουμε μηδενισμό διαφορετικών όρων της συνάρτησης του Andrews. Λόγω αυτού στο μέσο του Γραφήματος 2.6 παρατηρείται η μέγιστη διαφοροποίηση μεταξύ των ομάδων ενώ στο μέσο του Γραφήματος 2.7 παρατηρείται οι δύο ομάδες να παίρνουν πολύ κοντινές τιμές.



Γράφημα 2.6
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris στο διάστημα $[-\pi,\ \pi]$



Γράφημα 2.7
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris στο διάστημα $\left[-\frac{\pi}{2},\ \frac{3\pi}{2}\right]$

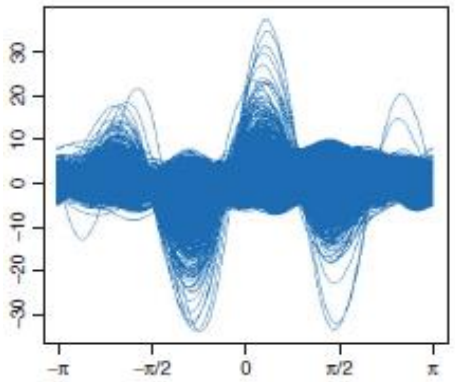
Αξιοποιώντας την εικόνα των καμπυλών στα παραπάνω δύο γραφήματα μπορεί κανείς να καταλήξει σε κάποια συμπεράσματα. Αρχικά με δεδομένο ότι γύρω από την περιοχή του 0 στο Γράφημα 2.6 υπάρχει προφανής διαχωρισμός των καμπυλών των δύο ομάδων, γίνεται φανερό πως οι μεταβλητές που πολλαπλασιάζονται με τους όρους των ημιτόνων δεν παίζουν σημαντικό ρόλο στην αναγνώριση των δύο ομάδων του συνόλου. Δεύτερον, στο Γράφημα 2.7, γύρω από την περιοχή του $\frac{\pi}{2}$ παρατηρείται ένας όχι και τόσο καλός διαχωρισμός μεταξύ των δύο ομάδων. Αυτό σημαίνει πως οι μεταβλητές που πολλαπλασιάζονται με τους όσους των συνημιτόνων δεν περιέχουν μεγάλο βαθμό πληροφορίας για το διαχωρισμό των δύο ομάδων.

ε. Πολλά από τα σύνολα δεδομένων που θα κληθεί να αναλύσει κανείς, περιέχουν ένα μεγάλο αριθμό από πολυδιάστατες παρατηρήσεις. Αυτό σημαίνει ότι μετά τη χάραξη των καμπυλών Andrews θα έχουμε ένα γράφημα με πάρα πολλές καμπύλες. Ένα τέτοιο γράφημα θα είναι εξαιρετικά δύσκολο ή και αδύνατο να αναλυθεί λόγω του μεγάλου πλήθους καμπυλών που εμφανίζονται και έτσι θα είναι ανέφικτο να εξαχθούν συμπεράσματα.

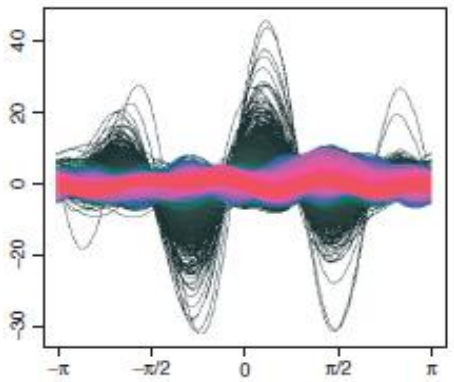
Για την αντιμετώπιση του παραπάνω προβλήματος ο Moustafa (2011) προτείνει μία τεχνική που χρησιμοποιεί το Quantized Generalized Parallel Coordinate Plot (QGPCP). Στην

περίπτωση αυτή χαράσσονται όλες οι καμπύλες και στη συνέχεια υπολογίζεται η πυκνότητα για όλα τα σημεία του γραφήματος. Ως πυκνότητα ενός σημείου θεωρείται ο αριθμός καμπυλών που περνούν από αυτό. Έτσι σημεία από τα οποία δεν περνούν καμπύλες αφήνονται λευκά, σημεία που περνά μόνο μία καμπύλη χρωματίζονται με μαύρο χρώμα και τα υπόλοιπα σημεία χρωματίζονται με θερμότερα χρώματα όσο μεγαλύτερη πυκνότητα παρουσιάζουν. Δηλαδή όσο μεγαλύτερος αριθμός καμπυλών περνάει από ένα σημείο, τόσο εντονότερο χρώμα έχει το σημείο. Με αυτόν τον τρόπο μπορούμε να βρούμε περιοχές του γραφήματος με μεγάλη πυκνότητα δηλαδή με μεγάλο αριθμό καμπυλών να περνούν από αυτά, δηλαδή μεγάλο αριθμό από παρατηρήσεις να είναι όμοιες μεταξύ τους.

Στα Γραφήματα 2.8 και 2.9 βλέπουμε την παραπάνω τεχνική για το σύνολο δεδομένων “California Housing Prices”, όπως αυτό περιγράφεται στο Παράρτημα 2. Συγκεκριμένα στο Γράφημα 2.8 έχουμε όλες τις καμπύλες και στο Γράφημα 2.9 έχουμε τις καμπύλες χρησιμοποιώντας την τεχνική του QGPCP. Στο Γράφημα 2.9 μπορεί κανείς εύκολα να διακρίνει μία ζώνη κόκκινου χρώματος. Το χρώμα της ζώνης αυτής δείχνει ότι από τη συγκεκριμένη περιοχή του επιπέδου περνούν πολλές καμπύλες και επομένως αυτές βρίσκονται πολύ κοντά μεταξύ τους. Μπορεί λοιπόν να βγει το συμπέρασμα ότι οι παρατηρήσεις που αντιστοιχούν σε αυτές τις καμπύλες είναι όμοιες παρατηρήσεις. Επίσης, παρατηρείται ότι η τιμή της συνάρτησης του Andrews για τη συγκεκριμένη ζώνη είναι αρκετά κοντά στο μηδέν σχεδόν για κάθε $-\pi < t < \pi$, συνεπώς οι τιμές που παίρνουν οι μεταβλητές του συνόλου για τις περισσότερες παρατηρήσεις, είναι μικρές.



Γράφημα 2.8
Οι καμπύλες Andrews για το σύνολο δεδομένων California Housing Prices



Γράφημα 2.9
Το QGPCP σε συνδυασμό των τις καμπύλες Andrews για το σύνολο δεδομένων California Housing Prices

Κ Ε Φ Α Λ Α Ι Ο 3

Παραλλαγές των Καμπυλών Andrews

3.1 Εισαγωγή

Σε αυτό το Κεφάλαιο παρουσιάζονται διάφορες παραλλαγές των καμπυλών Andrews που έχουν δοθεί στη βιβλιογραφία. Παρουσιάζονται αφ' ενός μεν παραλλαγές στις οποίες γίνεται χρήση της βάσης Fourier αλλά και αφ' ετέρου παραλλαγές στις οποίες δεν γίνεται χρήση της βάσης Fourier. Ακόμα δίνονται παραλλαγές στις τρεις διαστάσεις για ζευγαρωτές και μη παρατηρήσεις.

3.2 Παραλλαγές στις οποίες γίνεται χρήση της βάσης Fourier

Η συνάρτηση του Andrews είναι το εσωτερικό γινόμενο του διανύσματος $\boldsymbol{x}$ με το διάνυσμα της βάσης Fourier $\boldsymbol{\Phi_t}$. Ο Moustafa (2011) προτείνει αντί του διανύσματος $\boldsymbol{\Phi_t}$ να χρησιμοποιηθεί η παράγωγός του. Το διάνυσμα της κανονικοποιημένης παραγώγου του $\boldsymbol{\Phi_t}$ είναι το

$$\dot{\boldsymbol{\Phi}}_{\boldsymbol{t}} = (0, \cos(t), -\sin(t), \cos(2t), -\sin(2t), \dots).$$

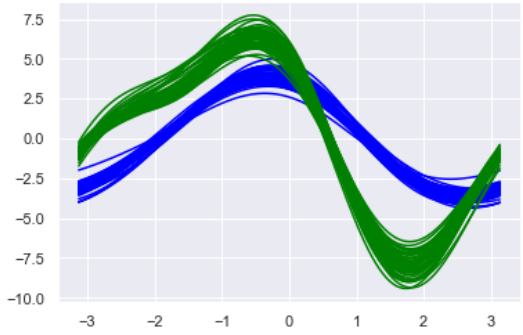
Οπότε η συνάρτηση που προτείνεται να χρησιμοποιηθεί παίρνει τη μορφή:

$$g_{\boldsymbol{x}}(t) = x_2 \cdot \cos(t) - x_3 \cdot \sin(t) + x_4 \cdot \cos(2t) - x_5 \cdot \sin(2t) + \dots, \qquad (3.1)$$

όπου $-\pi < t < \pi$.

Το διάνυσμα $\dot{\boldsymbol{\Phi}}_{\boldsymbol{t}}$ είναι και αυτό ορθοκανονική βάση του συνόλου των περιοδικών συναρτήσεων με περίοδο $T = 2$, επομένως το εσωτερικό γινόμενο $\langle \boldsymbol{x}, \dot{\boldsymbol{\Phi}}_{\boldsymbol{t}} \rangle$ θα διατηρεί όλες τις ιδιότητες του εσωτερικού γινομένου που βρίσκουμε στη συνάρτηση του Andrews. Συνεπώς μπορούμε να χαράξουμε τη συνάρτηση $g_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \dot{\boldsymbol{\Phi}}_{\boldsymbol{t}} \rangle$ και να χρησιμοποιήσουμε όλες τις χρήσιμες ιδιότητες των κλασικών καμπυλών Andrews.

Στο Γράφημα 3.1, μπορεί κανείς να δει τις καμπύλες Andrews για το σύνολο δεδομένων Iris, όπως αυτό περιγράφεται στο Παράρτημα 1. Αυτή τη φορά για τη χάραξη των καμπυλών έχει γίνει χρήση της συνάρτησης $g_{\boldsymbol{x}}(t)$ του τύπου (3.1). Οι καμπύλες πράσινου χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-virginica ενώ εκείνες μπλε χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-setosa. Φαίνεται ότι η συγκεκριμένη διαφοροποίηση των καμπυλών Andrews διατηρεί το χωρισμό των ομάδων που υπάρχουν στα δεδομένα, επομένως θα μπορούσε να χρησιμοποιηθεί για την εύρεση ομαδοποιήσεων σε μη ομαδοποιημένα δεδομένα.



Γράφημα 3.1
Οι γραφικές παραστάσεις της συνάρτησης (3.1) για το σύνολο δεδομένων Iris

Στην ίδια εργασία, αναφέρεται ότι μπορεί να χρησιμοποιηθεί στον υπολογισμό του εσωτερικού γινομένου είτε το διάνυσμα του αθροίσματος των δύο διανυσμάτων $\boldsymbol{\Phi_t}$ και $\dot{\boldsymbol{\Phi}}_{\boldsymbol{t}}$, δηλαδή το

$$\boldsymbol{\Phi_t} + \dot{\boldsymbol{\Phi}}_{\boldsymbol{t}} = (\frac{1}{\sqrt{2}}, \sin(t) + \cos(t), \cos(t) - \sin(t), \sin(2t) + \cos(2t), \cos(2t) - \sin(2t), \dots),$$

είτε το διάνυσμα της διαφοράς των $\boldsymbol{\Phi_t}$ και $\dot{\boldsymbol{\Phi}}_{\boldsymbol{t}}$, δηλαδή το

$$\boldsymbol{\Phi_t} - \dot{\boldsymbol{\Phi}}_{\boldsymbol{t}} = (\frac{1}{\sqrt{2}}, \sin(t) - \cos(t), \cos(t) + \sin(t), \sin(2t) - \cos(2t), \cos(2t) + \sin(2t), \dots).$$

Οι αντίστοιχες συναρτήσεις παίρνουν τις παρακάτω μορφές:

$$g_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \cdot (\sin(t) + \cos(t)) + x_3 \cdot (\cos(t) - \sin(t)) + x_4 \cdot (\sin(2t) + \cos(2t)) + \\ + x_5(\cos(2t) - \sin(2t)) + \dots$$

και

$$g_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \cdot (\sin(t) - \cos(t)) + x_3 \cdot (\cos(t) + \sin(t)) + x_4 \cdot (\sin(2t) - \cos(2t)) + \\ + x_5(\cos(2t) + \sin(2t)) + \dots,$$

όπου $-\pi < t < \pi$.

Το διάνυσμα $\boldsymbol{B'} = (\frac{1}{\sqrt{T}}, \sqrt{\frac{2}{T}} sin\left(\frac{2\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{2\pi}{T}x\right), \sqrt{\frac{2}{T}} sin\left(\frac{4\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{4\pi}{T}x\right), \dots,$ $\sqrt{\frac{2}{T}} sin\left(\frac{n\pi}{T}x\right), \sqrt{\frac{2}{T}} cos\left(\frac{n\pi}{T}x\right))$ αποδεικνύεται ότι αποτελεί μία ορθοκανονική βάση για το σύνολο των περιοδικών συναρτήσεων με περίοδο T. Συγκεκριμένα αποτελεί τη γενικευμένη βάση Fourier $\boldsymbol{\Psi_{\lambda t}} = (\frac{1}{\sqrt{T}}, \sqrt{\frac{2}{T}} sin(\lambda_1 t), \sqrt{\frac{2}{T}} cos(\lambda_2 t), \sqrt{\frac{2}{T}} sin(\lambda_3 t), \sqrt{\frac{2}{T}} cos(\lambda_4 t), \dots)$, με τα λ_i, $i = 1, 2, 3, 4, \dots$, να επιλέγονται κατάλληλα ώστε να είναι γραμμικά ανεξάρτητα στο σύνολο των πραγματικών αριθμών. Μία συνήθης επιλογή για τα $\lambda_i, i = 1, 2, 3, 4, \dots$, είναι $\lambda_i = 1, \sqrt{2}, \sqrt{3}, \sqrt{5}, \dots$ δηλαδή $\lambda_1 = 1$ και οι υπόλοιποι είναι οι τετραγωνικές ρίζες των πρώτων ακεραίων. Ο Moustafa (2011) επισημαίνει ότι αντί του εσωτερικού γινομένου $\langle \boldsymbol{x}, \boldsymbol{\Phi_t} \rangle$, θα μπορούσε κανείς να επιλέξει να χρησιμοποιήσει το $\langle \boldsymbol{x}, \boldsymbol{\Psi_{\lambda t}} \rangle$ αφού λόγω των περισσότερων παραμέτρων που περιέχει δίνει τη δυνατότητα πλουσιότερης απεικόνισης των δεδομένων στο επίπεδο μέσω της χάραξης της αντίστοιχης συνάρτησης. Η συνάρτηση που θα χαραχθεί παίρνει τη μορφή:

$$g_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{T}} + x_2 \cdot \sqrt{\frac{2}{T}} sin(\lambda_1 t) + x_3 \cdot \sqrt{\frac{2}{T}} cos(\lambda_2 t) + x_4 \cdot \sqrt{\frac{2}{T}} sin(\lambda_3 t) +$$

$$+ x_5 \cdot \sqrt{\frac{2}{T}} cos(\lambda_4 t) + \dots,$$

όπου $-\pi < t < \pi$.

Όμοια με παραπάνω, ο ίδιος συγγραφέας προτείνει και τη χρήση της παραγώγου της $\boldsymbol{\Psi_{\lambda t}}$ στο εσωτερικό γινόμενο που θα υπολογιστεί για να χαραχθούν οι καμπύλες. Προκύπτει λοιπόν,

$$\dot{\boldsymbol{\Psi}}_{\boldsymbol{\lambda t}} = (0, \sqrt{\frac{2}{T}} cos(\lambda_1 t), -\sqrt{\frac{2}{T}} sin(\lambda_2 t), \sqrt{\frac{2}{T}} cos(\lambda_3 t), -\sqrt{\frac{2}{T}} sin(\lambda_4 t), \dots)$$

και η συνάρτηση που θα χαραχθεί παίρνει τη μορφή

$$g_{\boldsymbol{x}}(t) = x_2 \cdot \sqrt{\frac{2}{T}} cos(\lambda_1 t) - x_3 \cdot \sqrt{\frac{2}{T}} sin(\lambda_2 t) + x_4 \cdot \sqrt{\frac{2}{T}} cos(\lambda_3 t) - x_5 \cdot \sqrt{\frac{2}{T}} sin(\lambda_4 t) + \dots,$$

όπου $-\pi < t < \pi$.

Τέλος, προτείνεται η χρήση του αθροίσματος και της διαφοράς των $\boldsymbol{\Psi_{\lambda t}}$ και $\dot{\boldsymbol{\Psi}}_{\boldsymbol{\lambda t}}$, δηλαδή η συνάρτηση

$$\boldsymbol{\Psi_{\lambda t}} + \dot{\boldsymbol{\Psi}}_{\boldsymbol{\lambda t}} = (\frac{1}{\sqrt{T}}, \sqrt{\frac{2}{T}} (sin(\lambda_1 t) + cos(\lambda_1 t)), \sqrt{\frac{2}{T}} (cos(\lambda_2 t) - sin(\lambda_2 t)),$$

$$\sqrt{\frac{2}{T}}(sin(\lambda_3 t) + cos(\lambda_3 t)) , \sqrt{\frac{2}{T}}(cos(\lambda_4 t) - sin(\lambda_4 t)), \dots)$$

και η

$$\boldsymbol{\Psi_{\lambda t}} - \dot{\boldsymbol{\Psi}}_{\boldsymbol{\lambda t}} = \frac{1}{\sqrt{T}}, \sqrt{\frac{2}{T}}(sin(\lambda_1 t) - cos(\lambda_1 t)) , \sqrt{\frac{2}{T}}(cos(\lambda_2 t) + sin(\lambda_2 t)),$$
$$\sqrt{\frac{2}{T}}(sin(\lambda_3 t) - cos(\lambda_3 t)) , \sqrt{\frac{2}{T}}(cos(\lambda_4 t) + sin(\lambda_4 t)), \dots) .$$

Οι αντίστοιχες συναρτήσεις προς χάραξη παίρνουν τη μορφή

$$g_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{T}} + x_2 \cdot \sqrt{\frac{2}{T}}(sin(\lambda_1 t) + cos(\lambda_1 t)) + x_3 \cdot \sqrt{\frac{2}{T}}(cos(\lambda_2 t) - sin(\lambda_2 t)) +$$
$$+ x_4 \cdot \sqrt{\frac{2}{T}}(sin(\lambda_3 t) + cos(\lambda_3 t)) + x_5 \cdot \sqrt{\frac{2}{T}}(cos(\lambda_4 t) - sin(\lambda_4 t)) + \dots$$

και

$$g_{\boldsymbol{x}}(t) = \frac{x_1}{\sqrt{T}} + x_2 \cdot \sqrt{\frac{2}{T}}(sin(\lambda_1 t) - cos(\lambda_1 t)) + x_3 \cdot \sqrt{\frac{2}{T}}(cos(\lambda_2 t) + sin(\lambda_2 t)) +$$
$$+ x_4 \cdot \sqrt{\frac{2}{T}}(sin(\lambda_3 t) - cos(\lambda_3 t)) + x_5 \cdot \sqrt{\frac{2}{T}}(cos(\lambda_4 t) + sin(\lambda_4 t)) + \dots ,$$

όπου $-\pi < t < \pi$.

3.3 Παραλλαγές στις οποίες δεν γίνεται χρήση της βάσης Fourier

Ο ίδιος ο Andrews προτείνει μία εναλλακτική για τις δικές του καμπύλες. Η εναλλακτική αυτή αφορά τη χάραξη της συνάρτησης με μορφή

$$g_{\boldsymbol{x}}(t) = x_1 \cdot sin(n_1 t) + x_2 \cdot cos(n_1 t) + x_3 \cdot sin(n_2 t) + x_4 \cdot cos(n_2 t) + \dots$$

όπου οι αριθμοί n_i, $i = 1, 2, 3, 4, \dots$, είναι ακέραιοι διαφορετικοί μεταξύ τους και $-\pi < t < \pi$.

Ο περιορισμός για τους αριθμούς n_i, $i = 1, 2, 3, 4, \dots$, να είναι ακέραιοι προτείνεται ώστε να εξακολουθεί να ισχύει η ιδιότητα της διατήρησης των αποστάσεων που περιγράφηκε στο Κεφάλαιο 2.

Σύμφωνα με τον συγγραφέα η χάραξη αυτής της συνάρτησης μπορεί να δίνει σημαντικά περισσότερη πληροφορία για τα δεδομένα σε σχέση με τις κλασικές καμπύλες. Ταυτόχρονα όμως, η συνάρτηση αυτή κάνει την ερμηνεία δυσκολότερη, όταν οι καμπύλες χρησιμοποιούνται για τη γραφική διερεύνηση των δεδομένων.

Μία υποπερίπτωση της παραπάνω συνάρτησης δίνεται επίσης από τον Andrews. Σε αυτήν, οι αριθμοί n_i, $i = 1, 2, 3, 4, \ldots$, είναι οι αντίστοιχες δυνάμεις του 2, δηλαδή $n_i = 2^i$, $i = 1, 2, 3, 4, \ldots$.

Η συνάρτηση που προκύπτει παίρνει τώρα τη μορφή

$$g_x(t) = x_1 \cdot sin(2t) + x_2 \cdot cos(2t) + x_3 \cdot sin(4t) + x_4 \cdot cos(4t) + x_5 \cdot sin(8t) + \ldots \quad (3.2)$$

όπου $-\pi < t < \pi$.

Η συνάρτηση αυτή έχει τα θετικά και τα αρνητικά της προηγούμενης συνάρτησης και απλά δίνει λύση στην επιλογή των συντελεστών n_i, $i = 1, 2, 3, 4, \ldots$.

Μία επιπλέον παραλλαγή της παραπάνω συνάρτησης δίνεται από τους Wegman and Shen (1993). Σε αυτή την περίπτωση η συνάρτηση παίρνει τη μορφή

$$g_x(t) = x_1 \cdot sin(w_1 t) + x_2 \cdot cos(w_2 t) + x_3 \cdot sin(w_3 t) + x_4 \cdot cos(w_4 t) + \ldots ,$$

όπου w_i, $i = 1, 2, 3, 4, \ldots$, άρρητοι αριθμοί μεταξύ του 0,5 και του 1 και $-\pi < t < \pi$.

Λόγω του ότι οι w_i, $i = 1, 2, 3, 4, \ldots$, είναι άρρητοι, η μόνη τριγωνομετρική ταυτότητα που ισχύει είναι η $\sin^2(w_i t) + cos^2(w_i t) = 1$.

Επίσης, επειδή οι w_i, $i = 1, 2, 3, 4, \ldots$, περιορίζονται μεταξύ των τιμών 0,5 και 1, καμία από τις μεταβλητές δεν υπερισχύει των υπολοίπων στη γραφική παράσταση και έτσι δίνεται η δυνατότητα σε όλες τις μεταβλητές να επηρεάσουν την εικόνα σχεδόν στον ίδιο βαθμό.

Το μειονέκτημα που έχει η συγκεκριμένη προσέγγιση αφορά το γεγονός πως δεν διατηρείται η ισχύς όλων των ιδιοτήτων που αποδείχθηκαν για τη συνάρτηση του Andrews στο Κεφάλαιο 2.

Μία πρόσθετη παραλλαγή που μοιάζει με την αμέσως προηγούμενη είναι μία συνάρτηση της μορφής

$$g_x(t) = x_1 \cdot sin(t) + x_2 \cdot cos\left(\sqrt{2}t\right) + x_3 \cdot sin\left(\sqrt{3}t\right) + x_4 \cdot cos\left(\sqrt{4}t\right) + \ldots ,$$

όπου τις θέσεις των συντελεστών w_i, $i = 1, 2, 3, 4, \ldots$, έχουν εδώ οι αριθμοί $\sqrt{1} = 1$, $\sqrt{2}$, $\sqrt{3}$, $\sqrt{4} = 2$, … . Είναι προφανές ότι κάποιοι από τους παραπάνω αριθμούς είναι άρρητοι ενώ κάποιοι άλλοι όχι. Επίσης, μπορεί κανείς να παρατηρήσει ότι οι συντελεστές σε αυτή την περίπτωση είναι μεγαλύτεροι ή ίσοι του 1 και όχι μεταξύ του 0,5 και του 1. Η τελευταία παραλλαγή αποδίδεται στον Tukey από τον Gnanadesikan (1997).

3.4 Παραλλαγές στις τρεις διαστάσεις

Με χρήση κατάλληλων προγραμμάτων σε ηλεκτρονικό υπολογιστή είναι δυνατό και αρκετά εύκολο να δημιουργήσει κανείς σχήματα στις τρεις διαστάσεις και επομένως είναι εφικτό να

δημιουργήσει και γραφικές παραστάσεις στις τρεις διαστάσεις. Στη συνέχεια, με κατάλληλη περιστροφή του παραχθέντος σχήματος δίνεται η δυνατότητα να το δει κανείς από όλες τις οπτικές γωνίες, από όλες τις «πλευρές», και να εξαχθούν χρήσιμα συμπεράσματα για τα στοιχεία που χρησιμοποιήθηκαν στην παραγωγή του σχήματος.

Χρησιμοποιώντας την παραπάνω λογική υπάρχει η δυνατότητα να σχεδιασθούν οι καμπύλες του Andrews στις τρεις διαστάσεις ώστε στη συνέχεια μέσω κατάλληλης περιστροφής του σχήματος να βγουν περισσότερα και ίσως πιο χρήσιμα συμπεράσματα σε σχέση με αυτά που βγαίνουν αν αποτυπωθούν οι καμπύλες στις δύο διαστάσεις.

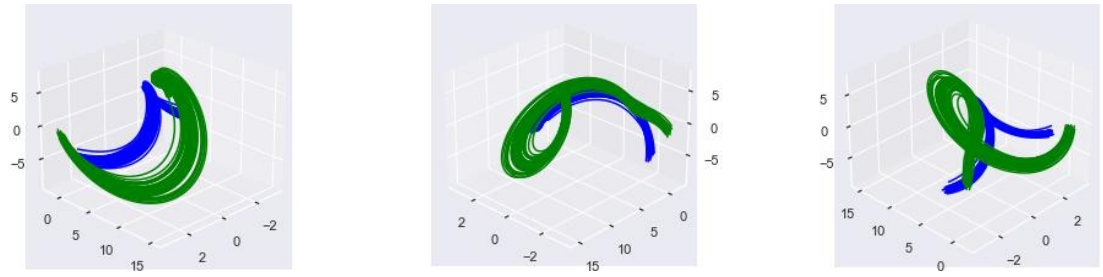
Ο Moustafa (2011) κάνει μία ενδιαφέρουσα πρόταση σχετικά με την χάραξη καμπυλών Andrews στις τρεις διαστάσεις. Πιο συγκεκριμένα προτείνεται να υπολογιστούν οι εξής δύο συναρτήσεις

$$f_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \boldsymbol{A_t} \rangle,\ g_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \boldsymbol{B_t} \rangle,$$

όπου τις θέσεις των $\boldsymbol{A_t}$ και $\boldsymbol{B_t}$ μπορούν να πάρουν τα διανύσματα $\boldsymbol{\Phi_t}$, $\boldsymbol{\dot{\Phi}_t}$, $\boldsymbol{\Psi_{\lambda t}}$ και $\boldsymbol{\dot{\Psi}_{\lambda t}}$ ή γραμμικοί συνδυασμοί τους, με κατάλληλο τρόπο ώστε $\boldsymbol{A_t} \neq \boldsymbol{B_t}$.

Στη συνέχεια γίνεται χάραξη των σημείων $\big(t, f_{\boldsymbol{x}}(t), g_{\boldsymbol{x}}(t)\big)$ στις τρεις διαστάσεις με $-\pi < t < \pi$. Έτσι σε κάθε παρατήρηση αντιστοιχεί μία καμπύλη στο χώρο. Αν τώρα, το σχήμα στραφεί ως προς οποιαδήποτε ευθεία του χώρου, προκύπτει και από μία προβολή των καμπυλών σε διαφορετικό επίπεδο. Με αυτόν τον τρόπο ανακαλύπτονται χαρακτηριστικά των καμπυλών και ως εκ τούτου και των παρατηρήσεων τα οποία δεν είναι ορατά στην απεικόνιση των δεδομένων με τις κλασικές καμπύλες του Andrews.

Για την επίδειξη της συγκεκριμένης διαφοροποίησης των καμπυλών Andrews θα χρησιμοποιηθεί ξανά το σύνολο δεδομένων Iris. Στο Γράφημα 3.2 δίνεται η γραφική αναπαράσταση σημείων της μορφής $\big(t, f_{\boldsymbol{x}}(t), g_{\boldsymbol{x}}(t)\big)$, όπου $-\pi < t < \pi$, $f_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \boldsymbol{\Phi_t} \rangle$ όπως στη συνάρτηση του τύπου (2.1) και $g_{\boldsymbol{x}}(t) = \langle \boldsymbol{x}, \boldsymbol{\dot{\Phi}_t} \rangle$ όπως στη συνάρτηση του τύπου (3.1). Στο γράφημα δίνονται επίσης τρεις διαφορετικές οπτικές γωνίες των καμπυλών.



Γράφημα 3.2
Η τρισδιάστατη παραλλαγή των καμπυλών Andrews για το σύνολο δεδομένων Iris από διαφορετικές γωνίες

Οι τρισδιάστατες, αυτή τη φορά, καμπύλες πράσινου χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-virginica ενώ εκείνες μπλε χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-setosa.

Με τον κώδικα που αναπτύχθηκε για την παρούσα εργασία υπάρχει η δυνατότητα περιστροφής του τρισδιάστατου γραφήματος ώστε να μπορεί ο ερευνητής να βγάλει περισσότερα χρήσιμα συμπεράσματα για το σύνολο των δεδομένων που μελετά.

Οι Koziol and Hacke (1991) προτείνουν μία παραλλαγή των καμπυλών Andrews ώστε αυτές να αποδοθούν στις τρεις διαστάσεις. Η προϋπόθεση για να χρησιμοποιηθεί η συγκεκριμένη μέθοδος είναι να έχουμε δύο διανύσματα, έστω $\boldsymbol{x}$ και $\boldsymbol{y}$, τα οποία να αποτελούν με φυσικό τρόπο ζεύγη παρατηρήσεων. Τέτοια ζεύγη παρατηρήσεων είναι για παράδειγμα μετρήσεις που γίνονται πριν και μετά την εφαρμογή μιας θεραπείας σε ασθενείς. Οι Koziol and Hacke προτείνουν την χάραξη των καμπυλών που διέρχονται από τα σημεία $\left(t, f_x(t), f_y(t)\right)$ για $-\pi < t < \pi$, όπου $f_x(t)$ είναι η συνάρτηση Andrews για το διάνυσμα $\boldsymbol{x}$ και $f_y(t)$ είναι η συνάρτηση Andrews για το διάνυσμα $\boldsymbol{y}$.

Με τη χρήση της μεθόδου τρισδιάστατης αναπαράστασης των δεδομένων, μπορεί κανείς να βρει συστάδες όμοιων παρατηρήσεων που σχετίζονται με μετρήσεις πριν και μετά από κάποιο γεγονός. Παραδείγματα τέτοιων γεγονότων θα μπορούσαν να είναι η εφαρμογή κάποιας θεραπείας σε ασθενείς ή γεωλογικές μετρήσεις πριν και μετά από κάποιο φαινόμενο. Με αυτόν τον τρόπο μπορεί κανείς να βρει παρατηρήσεις όπου η σχέση των μετρήσεων πριν και μετά να είναι όμοια.

Οι ιδιότητες που ισχύουν για τις συναρτήσεις $f_x(t)$ και $f_y(t)$ όταν αυτές χρησιμοποιούνται για τη χάραξη γραφημάτων στις δύο διαστάσεις, ισχύουν και στην περίπτωση που χρησιμοποιούνται για τη χάραξη γραφημάτων στις τρεις διαστάσεις. Επομένως, όλες οι ιδιότητες που περιγράφηκαν στο Κεφάλαιο 2 ισχύουν και στην περίπτωση των τρισδιάστατων καμπυλών Andrews και μπορούν να χρησιμοποιηθούν στην εξαγωγή συμπερασμάτων για τα δεδομένα.

3.5 Παραλλαγές στις οποίες δε γίνεται χρήση ημιτόνου και συνημιτόνου

Στις προηγούμενες ενότητες περιγράφηκαν διάφορες παραλλαγές των καμπυλών Andrews που χρησιμοποιούν τις τριγωνομετρικές συναρτήσεις του ημιτόνου και του συνημιτόνου. Σε αυτή την παράγραφο θα δοθούν τρεις παραλλαγές όπου δεν γίνεται χρήση τριγωνομετρικών συναρτήσεων.

Οι Embrechts, Hertzberg, Kalbfleisch, Traves and Whitla (1995) προτείνουν τη χρήση των wavelets στη χάραξη των καμπυλών Andrews.

Γενικά, τα wavelets κατασκευάζονται χρησιμοποιώντας μετατόπιση και αλλαγή του μεγέθους του σχήματος μίας συνάρτησης με επαρκή φθίνουσα πορεία τόσο στο χρόνο όσο και στη συχνότητα. Με τη φράση επαρκή φθίνουσα πορεία εννοούμε ότι μία συνάρτηση $\Psi(x)$ και ο μετασχηματισμός Fourier της $\widehat{\Psi}(f)$ φθίνουν ταχύτερα από τα $|x|^{-1}$ και $|f|^{-1}$ αντίστοιχα, δηλαδή ισχύουν οι σχέσεις

$$\int_{-\infty}^{+\infty}|x|^{-1}\cdot|\Psi(x)|\,dx<\infty,\ \int_{-\infty}^{+\infty}|f|^{-1}\cdot\left|\widehat{\Psi}(f)\right|df<\infty.$$

Μία συνάρτηση $\Psi(x)$ από την οποία προέρχεται ένα σύνολο από wavelets ονομάζεται βασικό wavelet ή μητρικό wavelet. Τα wavelets που προέρχονται από το ίδιο μητρικό wavelet μοιράζονται τόσο μαζί του όσο και μεταξύ τους διάφορες ιδιότητες του μητρικού wavelet όπως συνέχεια και διαφορισιμότητα. Μερικά παραδείγματα μητρικών wavelet είναι:

(α) Το wavelet του Haar:

$$\Psi(x)=\begin{cases}1, & 0\le x<\dfrac{1}{2}\\ -1, & \dfrac{1}{2}\le x<1\\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma\end{cases}$$

(β) Το wavelet του Franklin:

$$\Psi(x)=\begin{cases}0, & 0\le x<\frac{1}{2}\\ 2x-1, & \frac{1}{2}\le x<1\end{cases}$$

(γ) Το stetson hat wavelet:

$$\Psi(x)=\begin{cases}-2x-1, & -\frac{1}{2}\le x<0\\ 6x-1, & 0\le x<\frac{1}{2}\\ -6x+5, & \frac{1}{2}\le x<1\\ 2x-3, & 1\le x<\frac{3}{2}\end{cases}.$$

Το σύνολο των wavelets που παράγεται από κάποιο μητρικό wavelet, έστω $\Psi(x)$, δίνεται από την σχέση

$$\Psi_{\alpha,b}(x)=|a|^{-1/2}\Psi\left(\frac{x-b}{a}\right),$$

όπου a και b πραγματικοί αριθμοί με $a\neq 0$. Ο αριθμός a δίνει την αλλαγή κλίμακας στο wavelet και ο αριθμός b δίνει τη μετατόπιση. Μία οικογένεια wavelets είναι ένα σύνολο από wavelets όπου οι αριθμοί a και b μεταβάλλονται. Αν η μεταβολή των a και b γίνεται με συνεχή τρόπο τότε προκύπτει μία συνεχής οικογένεια wavelets, ενώ αν η μεταβολή των a και b γίνεται

με διακριτό τρόπο τότε προκύπτει μία διακριτή οικογένεια wavelets. Μία συνήθης επιλογή για τη δημιουργία διακριτών οικογενειών wavelets είναι να τεθεί το $a = 2^{-j}$ και το $b = k$, όπου j και k ακέραιοι αριθμοί με $0 \leq k \leq 2^j$. Θα πρέπει κανείς να προσέξει το γεγονός ότι τα στοιχεία μιας οικογένειας διακριτών wavelet δεν παίρνουν διακριτές τιμές, αλλά ονομάζονται έτσι λόγω των διακριτών τιμών των παραμέτρων a και b.

Επομένως, προτείνεται να χαραχθεί η συνάρτηση

$$g_x(t) = x_1 + x_2 \cdot \Psi_{0,0}(t) + x_3 \cdot \Psi_{1,0}(t) + x_4 \cdot \Psi_{1,1}(t) + x_5 \cdot \Psi_{2,1}(t) + \ldots,$$

όπου $0 \leq t \leq 1$.

Η παραπάνω συνάρτηση $g_x(t)$ διατηρεί κάποιες από τις ιδιότητες που περιγράφηκαν στο Κεφάλαιο 2, όμως όχι όλες. Για παράδειγμα δεν είναι βέβαιο ότι θα διατηρεί τη διασπορά όπως κάνει η συνάρτηση του Andrews.

Οι Embrechts and Herzberg (1991) πρότειναν μία ακόμα εναλλακτική των τριγωνομετρικών συναρτήσεων. Στη συγκεκριμένη εργασία προτείνεται η χρήση ορθογώνιων πολυωνύμων, όπως είναι τα πολυώνυμα Legendre και τα πολυώνυμα Chebyshev.

Τα πολυώνυμα Legendre $P_n(x)$ είναι οι λύσεις της διαφορικής εξίσωσης του Legendre

$$(1 - x^2)y''(x) - 2xy'(x) + n(n+1)y(x) = 0, -1 \leq x \leq 1$$

για $n = 0,1,2, \ldots$ και προκύπτουν από τις αναδρομικές σχέσεις

$$P_0(x) = 1,$$

$$P_1(x) = x,$$

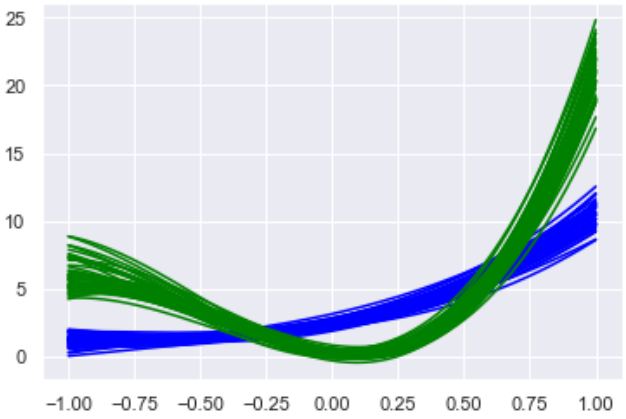
$$P_n(x) = \frac{2n-1}{n} x \cdot P_{n-1}(x) - \frac{n-1}{n} P_{n-2}(x), \; n \geq 2.$$

Η διαφοροποίηση που περιλαμβάνει τα πολυώνυμα Legendre αφορά στη χάραξη της συνάρτησης με μορφή

$$g_x(t) = \frac{1}{\sqrt{2}}\left\{x_1 + \sqrt{3}x_2 P_1(t) + \ldots + \sqrt{(2m-1)}x_m P_{m-1}(t)\right\}, \qquad (3.3)$$

όπου $-1 \leq t \leq 1$.

Στο Γράφημα 3.3, φαίνονται οι καμπύλες Andrews για το σύνολο δεδομένων Iris και εδώ η χάραξη των καμπυλών έγινε χρησιμοποιώντας τη συνάρτηση του τύπου (3.3). Οι καμπύλες πράσινου χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-virginica ενώ εκείνες μπλε χρώματος αντιστοιχούν στις παρατηρήσεις του είδους Iris-setosa. Η εικόνα είναι αρκετά διαφορετική από εκείνη των αρχικών καμπυλών Andrews, μιας και εδώ παρατηρείται ότι οι καμπύλες κάθε ομάδας βρίσκονται μέσα σε μία πιο στενή ζώνη και επομένως πιο κοντά η μία στην άλλη. Επίσης, παρατηρείται διατήρηση του διαχωρισμού των δύο ομάδων.



Γράφημα 3.3
Οι γραφικές παραστάσεις της συνάρτησης
(3.3) για το σύνολο δεδομένων Iris

Τα πολυώνυμα Chebyshev $T_n(x)$ ορίζονται από τις σχέσεις

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_n(x) = 2 \cdot x \cdot T_{n-1}(x) - T_{n-2}(x),\ n \geq 2\ ,$$

όπου $-1 \leq x \leq 1$.

Η πρόταση χρήσης των πολυωνύμων Chebyshev αφορά στη χάραξη της συνάρτησης με μορφή

$$g_x(t) = (1-t^2)^{-1/4}\left\{\frac{x_1}{2} \cdot T_0(t) + x_2 \cdot T_1(t) + x_3 \cdot T_2(t) + \ \ldots\right\},$$

όπου $-1 \leq t \leq 1$.

Κ Ε Φ Α Λ Α Ι Ο 4

Χρήσεις των Καμπυλών Andrews

4.1 Εισαγωγή

Σε αυτό το Κεφάλαιο περιγράφονται διάφορες χρήσεις τις οποίες έχουν οι καμπύλες Andrews. Παρουσιάζεται ο τρόπος που μπορούν να αξιοποιηθούν οι καμπύλες για Ομαδοποίηση και Ταξινόμηση παρατηρήσεων. Ακόμα παρουσιάζεται πως χρησιμοποιούνται οι καμπύλες για εντοπισμό ακραίων τιμών και τον εντοπισμό σημαντικών και μη μεταβλητών σε ένα σύνολο δεδομένων.

4.2 Ομαδοποίηση παρατηρήσεων

Όπως αναφέρθηκε και στο Κεφάλαιο 2, οι καμπύλες του Andrews έχουν κάποιες εξαιρετικά χρήσιμες ιδιότητες. Μία από αυτές τις ιδιότητες είναι εκείνη που αναφέρεται στη διατήρηση των αποστάσεων. Αυτό πρακτικά σημαίνει ότι παρατηρήσεις που βρίσκονται κοντά μεταξύ τους στον χώρο όπου ανήκουν, θα προβάλλονται μέσω της συνάρτησης του Andrews σε καμπύλες που βρίσκονται κοντά μεταξύ τους στις δύο διαστάσεις. Επομένως «κοντινές» παρατηρήσεις προβάλλονται σε «κοντινές» καμπύλες και αντιστρόφως, «κοντινές» καμπύλες αντιστοιχούν σε «κοντινές» παρατηρήσεις.

Αυτή η ιδιότητα είναι εξαιρετικά χρήσιμη διότι αξιοποιώντας την μπορεί κανείς να βρει ομαδοποιήσεις στα δεδομένα. Πιο συγκεκριμένα, αρχικά χαράζονται οι καμπύλες του Andrews για όλες τις παρατηρήσεις που βρίσκονται στα δεδομένα. Στη συνέχεια ερευνάται αν υπάρχουν καμπύλες οι οποίες να βρίσκονται κοντά μεταξύ τους και οι οποίες δημιουργούν συστάδες καμπυλών. Κάθε μία από αυτές τις συστάδες αποτελεί μία ομάδα καμπυλών και εφόσον αυτές οι καμπύλες έχουν μικρές αποστάσεις μεταξύ τους και οι παρατηρήσεις από τις οποίες παράχθηκαν οι συγκεκριμένες καμπύλες θα έχουν μικρές αποστάσεις μεταξύ τους. Επομένως,

μπορεί να κανείς να πει πως αυτές οι παρατηρήσεις αποτελούν μία ομάδα γιατί καθεμία τους απέχει μικρές αποστάσεις από τις υπόλοιπες.

Ακολουθώντας την ίδια λογική για κάθε συστάδα καμπυλών Andrews, μπορεί κανείς να δημιουργήσει ομάδες παρατηρήσεων εντός του συνόλου των δεδομένων. Αυτές οι ομάδες θα αποτελούνται από παρατηρήσεις με μικρές αποστάσεις μεταξύ τους που σημαίνει ότι κάθε ομάδα θα περιλαμβάνει όμοιες παρατηρήσεις. Συνεπώς, η χρήση των καμπυλών Andrews δίνει έναν αξιόλογο και οπτικό τρόπο για την εύρεση τυχόν ομάδων στα δεδομένα. Οι καμπύλες Andrews όπως φαίνεται δίνουν μία λύση σε προβλήματα συσταδοποίησης (clustering).

Η συγκεκριμένη χρησιμότητα των καμπυλών Andrews παρουσιάζεται σε πραγματικά δεδομένα στην Παράγραφο 5.2.3.

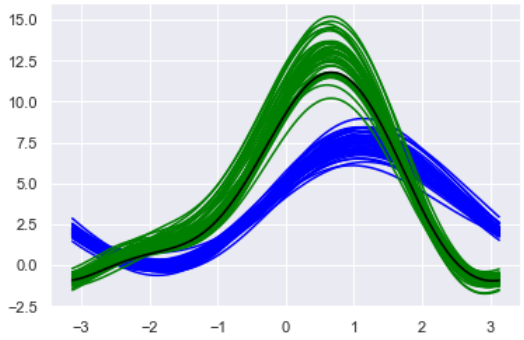
4.3 Ταξινόμηση παρατηρήσεων

Η ιδιότητα της διατήρησης των αποστάσεων μεταξύ των παρατηρήσεων μέσω της χάραξης των αντίστοιχων καμπυλών Andrews αποδεικνύεται σημαντική και στην περίπτωση που ο ερευνητής επιθυμεί να ταξινομήσει παρατηρήσεις σε υπάρχουσες ομάδες.

Πιο συγκεκριμένα ας υποτεθεί ένα σύνολο δεδομένων, για τις παρατηρήσεις του οποίου είναι γνωστό εκ των προτέρων σε ποιες ομάδες ανήκουν. Επίσης, θεωρούνται ακόμα μία ή περισσότερες παρατηρήσεις του ίδιου συνόλου δεδομένων για τις οποίες δεν είναι γνωστή η ομάδα στην οποία ανήκουν και ζητείται να ενταχθούν σε κάποια από τις ήδη γνωστές ομάδες που βρίσκονται στα δεδομένα. Ένα πρόβλημα τέτοιου τύπου, είναι ένα κλασικό πρόβλημα ταξινόμησης παρατηρήσεων (classification).

Για να αντιμετωπισθεί ένα τέτοιο πρόβλημα θα πρέπει αρχικά να χαραχθούν οι καμπύλες όλων των δεδομένων που ανήκουν στις γνωστές ομάδες, τονίζοντας με διαφορετικό χρώμα τις καμπύλες κάθε ομάδας. Έτσι προκύπτει ένα γράφημα από καμπύλες Andrews, στο οποίο εμφανίζονται οι ομάδες που υπάρχουν στα δεδομένα ως συστάδες καμπυλών με ίδιο χρώμα. Στη συνέχεια, αρκεί να χαραχθεί η καμπύλη Andrews, με διαφορετικό χρώμα από ότι τα υπάρχοντα στο γράφημα, για την παρατήρηση ή τις παρατηρήσεις για τις οποίες δεν είναι γνωστή η ομάδα στην οποία ανήκουν. Τέλος, ερευνάται αν κάποια από τις καμπύλες που χαράχθηκαν φαίνεται να απέχει μικρή απόσταση από το μέσο κάποιας από τις συστάδες καμπυλών που αντιστοιχούν στις ομάδες που βρίσκονται στα δεδομένα. Χρησιμοποιώντας την ιδιότητα πως «κοντινές» καμπύλες αντιστοιχούν σε «κοντινές» παρατηρήσεις, μπορεί κανείς να συμπεράνει πως οι παρατηρήσεις στις οποίες αντιστοιχούν οι καμπύλες που βρίσκονται κοντά σε κάποια συστάδα θα ανήκουν στην ομάδα στην οποία αντιστοιχεί η συγκεκριμένη συστάδα.

Για την επίδειξη της συγκεκριμένης χρησιμότητας των καμπυλών Andrews θα γίνει χρήση και πάλι του συνόλου δεδομένων Iris. Η διαφορά είναι ότι τώρα για την τελευταία παρατήρηση του είδους Iris-virginica θα γίνει η υπόθεση ότι δεν είναι γνωστή η κατηγορία στην οποία ανήκει. Στο Γράφημα 4.1 φαίνονται με πράσινο χρώμα οι καμπύλες Andrews για τις 49 παρατηρήσεις του είδους Iris-virginica, με μπλε χρώμα εκείνες του είδους Iris-setosa και με μαύρο χρώμα εκείνη που αντιστοιχεί στην παρατήρηση της οποίας έχει θεωρηθεί άγνωστο το είδος.



Γράφημα 4.1
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris με μία παρατήρηση άγνωστης κατηγορίας

Όπως φαίνεται και στο γράφημα, η μαύρη καμπύλη βρίσκεται, για κάθε $t \in (-\pi, \pi)$, πολύ πιο κοντά στις καμπύλες του είδους Iris-virginica από ότι σε εκείνες του είδους Iris-setosa καθώς και συμπεριφέρεται με τον ίδιο τρόπο. Συνεπώς, μπορεί κανείς να συμπεράνει ότι η παρατήρηση στην οποία αντιστοιχεί η καμπύλη μαύρου χρώματος ανήκει στο είδος Iris-virginica.

4.4 Εντοπισμός ακραίων παρατηρήσεων

Σε πολλά σύνολα δεδομένων και κυρίως σε σύνολα δεδομένων που περιέχουν πραγματικά δεδομένα συναντώνται ακραίες τιμές. Αυτές οι ακραίες τιμές μπορεί να αφορούν μία ή περισσότερες, ακόμα και όλες, τις μεταβλητές της παρατήρησης. Μία τέτοια παρατήρηση ονομάζεται ακραία παρατήρηση. Όταν κάποια από τις μεταβλητές μιας παρατήρησης παίρνει ακραίες τιμές, είτε μικρές είτε μεγάλες, τότε η συγκεκριμένη παρατήρηση θα είναι απομακρυσμένη από τις υπόλοιπες στο χώρο όπου ανήκουν. Τέτοιες παρατηρήσεις θα πρέπει να ανιχνεύονται πριν από οποιαδήποτε ανάλυση γίνεται στα δεδομένα γιατί έχουν τη δυνατότητα να οδηγήσουν σε λανθασμένα συμπεράσματα. Αν χρησιμοποιηθούν τα δεδομένα για τη δημιουργία μοντέλων πρόβλεψης, οι παρατηρήσεις αυτές μπορούν να επηρεάσουν σημαντικά το μοντέλο. Αφού ανιχνεύσει κανείς τις ακραίες παρατηρήσεις των δεδομένων, τις

αφαιρεί από αυτά και τις διαχειρίζεται με κάποιο διαφορετικό τρόπο ανάλογα με το ζητούμενο που έχει.

Για να λύσει κανείς το πρόβλημα της εύρεσης των ακραίων παρατηρήσεων μπορεί να χρησιμοποιήσει τις καμπύλες Andrews. Για ακόμα μία φορά θα χρησιμοποιηθεί η ιδιότητα της διατήρησης των αποστάσεων που έχουν οι συγκεκριμένες καμπύλες. Με βάση αυτή την ιδιότητα, παρατηρήσεις που βρίσκονται κοντά μεταξύ τους αντιστοιχούν σε καμπύλες που βρίσκονται κοντά μεταξύ τους. Κάνοντας μια «αντιστροφή» της συγκεκριμένης ιδιότητας, μπορεί κανείς να πει πως παρατηρήσεις που απέχουν πολύ μεταξύ τους, αντιστοιχούν σε καμπύλες που απέχουν πολύ μεταξύ τους.

Επομένως, αφού χαραχθούν όλες οι καμπύλες Andrews για τις παρατηρήσεις ενός συνόλου δεδομένων, θα πρέπει να αναζητηθούν αυτές που εμφανίζουν διαφορετική συμπεριφορά σε σχέση με τις υπόλοιπες. Πιο συγκεκριμένα θα πρέπει κανείς να αναζητήσει καμπύλες που διαφέρουν σημαντικά από τις υπόλοιπες, τόσο στην απόσταση από τον άξονα των x όσο και στη θέση των τοπικών και ολικών μεγίστων και ελαχίστων. Αυτές οι καμπύλες αντιστοιχούν στις παρατηρήσεις που θεωρούνται ακραίες και είναι εκείνες που θα πρέπει να εξαιρεθούν από οποιαδήποτε περαιτέρω ανάλυση των δεδομένων. Μετά την αφαίρεση των αντίστοιχων καμπυλών το γράφημα θα πρέπει να αποτελείται από καμπύλες που παρουσιάζουν παρόμοια συμπεριφορά, είτε όλες μαζί είτε σε ομάδες.

Ακραίες παρατηρήσεις μπορεί να βρεθούν και σε δεδομένα χρονοσειρών. Είναι πιθανό μερικές από τις τιμές μίας χρονοσειράς να μην βρίσκονται στο ίδιο επίπεδο με τις υπόλοιπες τιμές και επομένως θεωρούνται ακραίες τιμές. Προφανώς, μπορεί κανείς να βρει ακραίες τιμές αρκετά μεγαλύτερες από τις υπόλοιπες παρατηρήσεις αλλά και αρκετά μικρότερες από τις υπόλοιπες παρατηρήσεις.

Οι καμπύλες Andrews μπορούν να χρησιμοποιηθούν ώστε να βρεθούν αυτές οι ακραίες παρατηρήσεις στις τιμές μιας χρονοσειράς όπως επισημαίνουν οι Embrechts, Herzberg and Allen (1986). Αρχικά θα πρέπει να δημιουργηθεί το γράφημα των τιμών της χρονοσειράς ώστε να βρεθεί η περίοδος της. Έστω, λοιπόν, μια χρονοσειρά με αριθμό παρατηρήσεων s και περίοδο n. Στο επόμενο βήμα θα πρέπει να δημιουργηθούν n γραφήματα, όπου το πρώτο από αυτά θα περιέχει s/n καμπύλες Andrews και κάθε επόμενο θα περιέχει $(s/n) + 1$ καμπύλες. Κάθε μία από τις καμπύλες Andrews του πρώτου γραφήματος θα περιέχει n διαδοχικές παρατηρήσεις. Δηλαδή η πρώτη καμπύλη του πρώτου γραφήματος θα δημιουργείται από τις παρατηρήσεις $i, i + 1, i + 2, \ldots, i + n - 1$, η δεύτερη από τις παρατηρήσεις $i + n, i + n + 1, i + n + 2, \ldots, i + 2n - 1$ και ούτω καθεξής. Η πρώτη καμπύλη του δεύτερου γραφήματος θα δημιουργείται από τις παρατηρήσεις $i + 1,\ i + 2,\ i + 3,\ \ldots,\ i + n$, η δεύτερη από τις παρατηρήσεις $i + n + 1, i + n + 2, i + n + 3, \ldots, i + 2n$ και ούτε καθεξής. Με αυτή τη λογική δημιουργούνται τα n γραφήματα που περιέχουν s/n καμπύλες Andrews το πρώτο και $(s/n) + 1$ καμπύλες Andrews τα υπόλοιπα.

Τέλος, ελέγχεται η εικόνα κάθε γραφήματος και αναζητούνται εκείνα τα γραφήματα στα οποία δεν φαίνεται να υπάρχει μία ζώνη μέσα στην οποία να βρίσκονται όλες οι καμπύλες. Ουσιαστικά, τα γραφήματα όπου οι καμπύλες Andrews δεν είναι κοντά μεταξύ τους, για το μεγαλύτερο μέρος του διαστήματος $(-\pi, \pi)$, είναι αυτά που έχουν επηρεαστεί περισσότερο από την ύπαρξη των ακραίων τιμών. Σε ένα τέτοιο γράφημα, όπου οι καμπύλες δεν βρίσκονται κοντά μεταξύ τους, εντοπίζεται η καμπύλη ή οι καμπύλες που «συμπεριφέρονται» διαφορετικά και ελέγχονται οι τιμές των συντελεστών των όρων της συνάρτησης του Andrews. Αυτές οι τιμές αποτελούν διαδοχικές τιμές της χρονοσειράς και μέσα σε αυτές περιέχεται και η ακραία τιμή.

Αν η παραπάνω διαδικασία πραγματοποιηθεί για όλες τις καμπύλες Andrews που συμπεριφέρονται διαφορετικά από τις υπόλοιπες σε όλα τα γραφήματα που έχουν παραχθεί πιο πριν, τότε θα εντοπισθούν με αρκετή ευκολία όλες οι ακραίες τιμές που περιέχονται σε μία χρονοσειρά. Η εύρεση αυτών των ακραίων τιμών και η διαχείρισή τους με ιδιαίτερο τρόπο ανάλογα με την περίπτωση, μπορεί να οδηγήσει σε καλύτερα μοντέλα πρόβλεψης των μελλοντικών τιμών μίας χρονοσειράς.

Η παραπάνω διαδικασία εντοπισμού ακραίων τιμών παρουσιάζεται σε πραγματικά πολυδιάστατα δεδομένα στην Παράγραφο 5.2.4 και σε δεδομένα χρονοσειρών στην Παράγραφο 5.4.3.

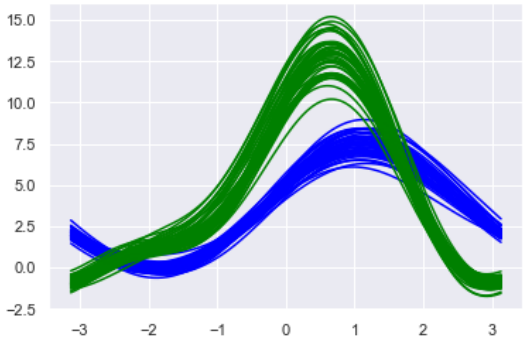
4.5 Εντοπισμός σημαντικών και μη μεταβλητών

Σε κάποια σύνολα δεδομένων, μπορεί κανείς να συναντήσει ένα μεγάλο αριθμό μεταβλητών. Σε αυτή την περίπτωση, αλλά και στην περίπτωση μικρού αριθμού μεταβλητών, είναι σημαντικό να γνωρίζει ο αναλυτής ποιες από τις μεταβλητές που έχει στη διάθεση του, είναι χρήσιμες ως προς το σύνολο των δεδομένων. Ο όρος χρήσιμες αναφέρεται στην ιδιότητα μιας μεταβλητής να περιέχει μεγάλο μέρος της πληροφορίας για τις παρατηρήσεις των δεδομένων. Οι καμπύλες Andrews μπορούν να χρησιμοποιηθούν ώστε να συμπεράνει κανείς αν μία μεταβλητή ή μία ομάδα από μεταβλητές περιέχει αρκετή πληροφορία για τα δεδομένα και επομένως αν αυτή η μεταβλητή ή η ομάδα μεταβλητών είναι σημαντική όπως αναφέρουν οι Embrechts and Herzberg (1991).

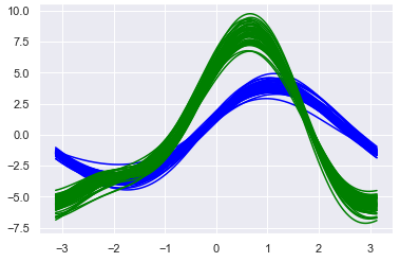
Για να γίνει ο παραπάνω έλεγχος, αρχικά πρέπει να χαραχθούν οι καμπύλες Andrews για όλες τις παρατηρήσεις των δεδομένων και αν υπάρχει εκ των προτέρων ομαδοποίηση να χρησιμοποιηθεί διαφορετικός χρωματισμός για τις καμπύλες κάθε ομάδας. Στη συνέχεια αντικαθίσταται στη συνάρτηση του Andrews ο συντελεστής του i-οστού όρου με 0 δηλαδή προκύπτει ότι $x_i = 0$. Ακολούθως, χαράσσονται οι καμπύλες που δημιουργήθηκαν από τις τιμές της συνάρτησης του Andrews με $x_i = 0$, σε ένα νέο γράφημα. Τέλος συγκρίνεται η εικόνα

που εμφανίζεται στα δύο γραφήματα, το αρχικό και το νέο, και ελέγχεται αν στο νέο γράφημα εξακολουθεί να υπάρχει παρόμοια εικόνα. Πιο συγκεκριμένα, ελέγχεται αν εξακολουθεί να υπάρχει η ομαδοποίηση που υπήρχε στο αρχικό γράφημα και αν οι καμπύλες συμπεριφέρονται με τον ίδιο τρόπο.

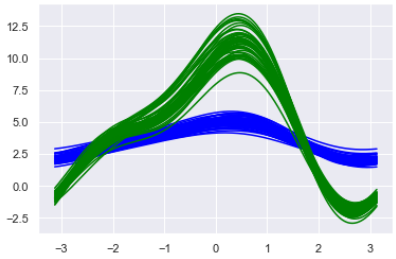
Για όποιο από τα νέα γραφήματα παρατηρηθεί παρόμοια εικόνα και συμπεριφορά των καμπυλών σε σχέση με το αρχικό, μπορεί να βγει το συμπέρασμα ότι η αντίστοιχη μεταβλητή δεν περιέχει μεγάλο μέρος από την πληροφορία των δεδομένων. Αυτό συμβαίνει γιατί η εξάλειψη αυτής της μεταβλητής δεν άλλαξε την εικόνα των καμπυλών Andrews. Αντίθετα, αν σε κάποιο γράφημα παρατηρηθεί αλλαγή της εικόνας και της συμπεριφοράς των καμπυλών, φαίνεται ότι η εξάλειψη αυτής της μεταβλητής οδηγεί σε απώλεια σημαντικού μέρους της πληροφορίας που περιέχεται στα δεδομένα.



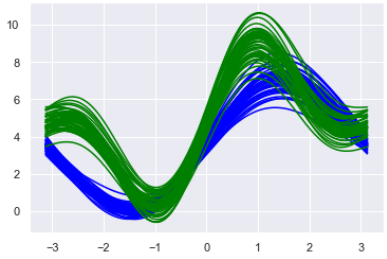
Γράφημα 4.2
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris



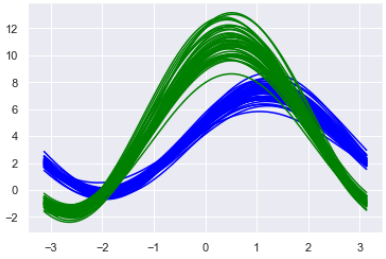
Γράφημα 4.3
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris μετά το μηδενισμό της μεταβλητής sepal_length



Γράφημα 4.4
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris μετά το μηδενισμό της μεταβλητής sepal_width



Γράφημα 4.5
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris μετά το μηδενισμό της μεταβλητής petal_length



Γράφημα 4.6
Οι καμπύλες Andrews για το σύνολο δεδομένων Iris μετά το μηδενισμό της μεταβλητής petal_width

Στο Γράφημα 4.2 φαίνονται οι καμπύλες Andrews για το σύνολο δεδομένων Iris. Αντίστοιχα, στα Γραφήματα 4.3, 4.4, 4.5 και 4.6 φαίνονται οι καμπύλες Andrews για το παραπάνω σύνολο, έχοντας κάθε φορά αντικαταστήσει διαδοχικά τις τέσσερις μεταβλητές του συνόλου με 0. Έτσι, για την παραγωγή του Γραφήματος 4.3 έχει γίνει αντικατάσταση της μεταβλητής sepal_length με 0, για την παραγωγή του Γραφήματος 4.4 έχει γίνει αντικατάσταση της μεταβλητής sepal_width με 0, για την παραγωγή του Γραφήματος 4.5 έχει γίνει αντικατάσταση της μεταβλητής petal_length με 0 και τέλος για την παραγωγή του Γραφήματος 4.6 έχει γίνει αντικατάσταση της μεταβλητής petal_width με 0.

Και πάλι, με πράσινο χρώμα φαίνονται οι καμπύλες του είδους Iris-virginica, ενώ με μπλε οι καμπύλες του είδους Iris-setosa. Είναι φανερό ότι ακόμα και μετά τον μηδενισμό των τεσσάρων μεταβλητών διαδοχικά, ο διαχωρισμός των δύο ομάδων παρατηρήσεων είναι εμφανής. Αυτό το γεγονός δείχνει ότι θα μπορούσαν να χρησιμοποιηθούν λιγότερες από τις διαθέσιμες μεταβλητές ενός συνόλου δεδομένων για τη χάραξη των καμπυλών, εφόσον αυτό δε δημιουργεί προβλήματα διαχωρισμού των ομάδων.

Για να αποφασίσει κανείς ποιες από τις τέσσερις διαθέσιμες μεταβλητές του συνόλου δεδομένων Iris περιέχουν μεγάλο ποσοστό πληροφορίας θα πρέπει να συγκρίνει διαδοχικά τα Γραφήματα 4.3, 4.4, 4.5 και 4.6 με το Γράφημα 4.2. Μετά από αυτή τη σύγκριση, γίνεται φανερό ότι η μεταβλητή petal_length περιέχει μεγάλο ποσοστό από την πληροφορία που υπάρχει στα δεδομένα και αυτό συμβαίνει γιατί το Γράφημα 4.5 εμφανίζει τη μεγαλύτερη διαφορά με το Γράφημα 4.2. Αντίθετα, το Γράφημα 4.3 εμφανίζει τις λιγότερες διαφορές με το Γράφημα 4.2 και επομένως γίνεται αντιληπτό ότι η μεταβλητή sepal_length περιέχει ένα πολύ μικρό ποσοστό πληροφορίας για τα δεδομένα.

Με παρόμοια διαδικασία με την παραπάνω μπορεί κανείς να ελέγξει και ομάδες μεταβλητών κάνοντας παράλληλες αντικαταστάσεις με 0, δηλαδή $x_i = 0$, $x_j = 0$, …, $x_k = 0$. Τα συμπεράσματα που θα προκύψουν είναι αντίστοιχα των παραπάνω.

4.6 Έλεγχος υποθέσεων και δημιουργία διαστημάτων εμπιστοσύνης

Βασιζόμενος στο γράφημα των καμπυλών Andrews για ένα σύνολο δεδομένων, ο ίδιος ο Andrews (1972) αναφέρει ότι μπορούν να πραγματοποιηθούν δύο ειδών στατιστικοί έλεγχοι υποθέσεων και να κατασκευασθούν τα αντίστοιχα διαστήματα εμπιστοσύνης. Η πρώτη μορφή ελέγχου αφορά συγκεκριμένες τιμές t, επιλεγμένες a priori από την εικόνα του γραφήματος, ενώ η δεύτερη αφορά το σύνολο των τιμών $-\pi < t < \pi$.

Πιο συγκεκριμένα για τον πρώτο τύπο, είναι φανερό από την εικόνα του γραφήματος ότι ενδέχεται να υπάρχουν συγκεκριμένες τιμές t για τις οποίες έχει ενδιαφέρον η πραγματοποίηση του ελέγχου

$$H_0: f_{\bar{\boldsymbol{x}}}(t) = f_{\boldsymbol{x_0}}(t)$$

$$H_1: f_{\bar{\boldsymbol{x}}}(t) \neq f_{\boldsymbol{x_0}}(t),$$

για κάποιο υποθετικό $\boldsymbol{x_0}$ και για συγκεκριμένο t. Κάνοντας χρήση του συγκεκριμένου ελέγχου μπορεί κανείς να αποφασίσει αν υπάρχουν επαρκείς στατιστικές ενδείξεις ότι η καμπύλη του μέσου των παρατηρήσεων $f_{\bar{\boldsymbol{x}}}(t)$ ταυτίζεται με την καμπύλη του διανύσματος $\boldsymbol{x_0}$, $f_{\boldsymbol{x_0}}(t)$, για κάποιο συγκεκριμένο t και επομένως αν υπάρχουν ενδείξεις ότι ο μέσος των πολυδιάστατων παρατηρήσεων $\bar{\boldsymbol{x}}$ είναι ίσος με το διάνυσμα $\boldsymbol{x_0}$.

Αν το n-διάστατο τυχαίο διάνυσμα $\boldsymbol{x}$ ακολουθεί n-διάστατη κανονική κατανομή, τότε κάθε $x_i,\ i = 1, 2, \dots, n$ ακολουθεί μονοδιάστατη κανονική κατανομή και αντίστοιχα κάθε $\bar{x}_i$, $i = 1, 2, \dots, n$ ακολουθεί μονοδιάστατη κανονική κατανομή. Εφόσον η συνάρτηση $f_{\bar{\boldsymbol{x}}}(t)$ αποτελεί γραμμικό συνδυασμό των n το πλήθος τυχαίων μεταβλητών $\bar{x}_i,\ i = 1, 2, \dots, n$, θα ακολουθεί μονοδιάστατη κανονική κατανομή. Επίσης, πρέπει να σημειωθεί ότι η διασπορά της συνάρτησης $f_{\bar{\boldsymbol{x}}}(t)$ είναι γνωστή και περίπου σταθερή όπως αποδείχθηκε στις ιδιότητες των καμπυλών Andrews.

Με βάση όλα τα παραπάνω μπορεί να κατασκευαστεί η στατιστική συνάρτηση

$$Z = \frac{f_{\bar{\boldsymbol{x}}}(t) - f_{\boldsymbol{x_0}}(t)}{\sqrt{Var(f_{\bar{\boldsymbol{x}}}(t))}},$$

η οποία κάτω από τη μηδενική υπόθεση ακολουθεί την τυποποιημένη κανονική κατανομή $N(0, 1)$. Επομένως, χρησιμοποιώντας τα ποσοστιαία σημεία της τυποποιημένης κανονικής κατανομής, μπορεί να πραγματοποιηθεί ο παραπάνω έλεγχος σε οποιοδήποτε επίπεδο εμπιστοσύνης α επιθυμεί ο ερευνητής. Ακόμα, μπορούν να κατασκευαστούν διαστήματα εμπιστοσύνης επιπέδου $(1 - \alpha) \cdot 100\%$, κάνοντας χρήση της κατανομής της στατιστικής συνάρτησης Z.

Για τον δεύτερο τύπο στατιστικού ελέγχου υποθέσεων, δεν θα χρησιμοποιηθούν μεμονωμένες τιμές t, αλλά πρόκειται για έναν έλεγχο για το σύνολο των t, όπου $-\pi < t < \pi$. Επομένως, ο έλεγχος εδώ είναι ο εξής:

$$H_0: f_{\bar{x}}(t) = f_{\boldsymbol{x_0}}(t) \text{ για κάθε } t \text{ με } -\pi < t < \pi$$

$$H_1: f_{\bar{x}}(t) \neq f_{\boldsymbol{x_0}}(t) \text{ για τουλάχιστον ένα } t \text{ με } -\pi < t < \pi,$$

για κάποιο υποθετικό $\boldsymbol{x_0}$.

Για αυτόν τον έλεγχο θα χρειαστούν τρία στοιχεία. Πρώτον, αν το n-διάστατο τυχαίο διάνυσμα $\bar{\boldsymbol{x}}$ αποτελείται από n ασυσχέτιστες τυχαίες μεταβλητές $\bar{x_i}$ και ακολουθεί n-διάστατη κανονική κατανομή με μέσο $\boldsymbol{\mu} = \{\mu_i, i = 1,2,\cdots,n\}$ και κοινή διασπορά σ^2, τότε η τυχαία μεταβλητή που δίνει το τετράγωνο του μέτρου της διαφοράς των διανυσμάτων $\bar{\boldsymbol{x}}$ και $\boldsymbol{x_0}$ προς τη διασπορά, δηλαδή η μεταβλητή $\frac{\|\bar{x}-\boldsymbol{x_0}\|^2}{\sigma^2}$, ακολουθεί την κατανομή χ^2 με n βαθμούς ελευθερίας. Άρα για τη συγκεκριμένη μεταβλητή ισχύει ότι

$$P\left(\frac{\|\bar{x}-\boldsymbol{x_0}\|^2}{\sigma^2} \leq \chi_n^2(a)\right) = 1 - a,$$

δηλαδή

$$\frac{\|\bar{x}-\boldsymbol{x_0}\|^2}{\sigma^2} \leq \chi_n^2(a) \text{ με πιθανότητα } 1 - a,$$

όπου $\chi_n^2(a)$ είναι το άνω a ποσοστιαίο σημείο της κατανομής χ^2 με n βαθμούς ελευθερίας.

Δεύτερον, είναι γνωστό ότι το τετράγωνο του μήκους της προβολής ενός διανύσματος $\boldsymbol{x}$ σε οποιοδήποτε άλλο διάνυσμα δεν είναι μεγαλύτερο από το τετράγωνο του μέτρου του διανύσματος $\boldsymbol{x}$.

Τρίτον, έστω το διάνυσμα $\boldsymbol{v} = \frac{\boldsymbol{\Phi_t}}{\|\boldsymbol{\Phi_t}\|}$ το οποίο έχει μέτρο ίσο με 1. Για το τετράγωνο του μήκους της προβολής του διανύσματος $\bar{\boldsymbol{x}} - \boldsymbol{x_0}$ πάνω στο διάνυσμα $\boldsymbol{v}$ ισχύει ότι

$$\pi\rho o\beta_{\boldsymbol{v}}(\bar{\boldsymbol{x}} - \boldsymbol{x_0})^2 = \|\langle(\bar{\boldsymbol{x}} - \boldsymbol{x_0}) \cdot \boldsymbol{v}\rangle\|^2 = \|\langle\bar{\boldsymbol{x}} \cdot \boldsymbol{v}\rangle - \langle\boldsymbol{x_0} \cdot \boldsymbol{v}\rangle\|^2 =$$

$$= \left\|\langle\bar{\boldsymbol{x}} \cdot \frac{\boldsymbol{\Phi_t}}{\|\boldsymbol{\Phi_t}\|}\rangle - \langle\boldsymbol{x_0} \cdot \frac{\boldsymbol{\Phi_t}}{\|\boldsymbol{\Phi_t}\|}\rangle\right\|^2 = \left\|\frac{\langle\bar{x}\cdot\boldsymbol{\Phi_t}\rangle}{\|\boldsymbol{\Phi_t}\|} - \frac{\langle\boldsymbol{x_0}\cdot\boldsymbol{\Phi_t}\rangle}{\|\boldsymbol{\Phi_t}\|}\right\|^2 = \left\|\frac{f_{\bar{x}}(t)}{\|\boldsymbol{\Phi_t}\|} - \frac{f_{\boldsymbol{x_0}}(t)}{\|\boldsymbol{\Phi_t}\|}\right\|^2 =$$

$$= \left\|\frac{f_{\bar{x}}(t) - f_{\boldsymbol{x_0}}(t)}{\|\boldsymbol{\Phi_t}\|}\right\|^2 = \frac{\left\|f_{\bar{x}}(t) - f_{\boldsymbol{x_0}}(t)\right\|^2}{\|\boldsymbol{\Phi_t}\|}.$$

Από τις τρεις παραπάνω παρατηρήσεις, προκύπτει διαδοχικά ότι ισχύουν τα εξής με πιθανότητα $1 - a$:

$$\frac{\|\bar{x}-\boldsymbol{x_0}\|^2}{\sigma^2} \leq \chi_n^2(a) \Rightarrow \|\bar{\boldsymbol{x}} - \boldsymbol{x_0}\|^2 \leq \sigma^2 \cdot \chi_n^2(a) \Rightarrow$$

$$\pi\rho o\beta_{\boldsymbol{v}}(\bar{\boldsymbol{x}} - \boldsymbol{x_0})^2 \leq \|\bar{\boldsymbol{x}} - \boldsymbol{x_0}\|^2 \leq \sigma^2 \cdot \chi_n^2(a) \Rightarrow$$

$$\pi\rho o\beta_{\boldsymbol{v}}(\bar{\boldsymbol{x}} - \boldsymbol{x_0})^2 \leq \sigma^2 \cdot \chi_n^2(a) \Rightarrow \frac{\left\|f_{\bar{x}}(t) - f_{\boldsymbol{x_0}}(t)\right\|^2}{\|\boldsymbol{\Phi_t}\|} \leq \sigma^2 \cdot \chi_n^2(a) \Rightarrow$$

$$\left\|f_{\bar{x}}(t) \; - \; f_{\boldsymbol{x_0}}(t)\right\|^2 \leq \|\boldsymbol{\Phi_t}\| \cdot \sigma^2 \cdot \chi_n^2(a) \text{ με πιθανότητα } 1 - a.$$

Τέλος, είναι γνωστό ότι το μέτρο του διανύσματος $\boldsymbol{\Phi_t}$ είναι μικρότερο από την ποσότητα $\frac{n+1}{2}$, επομένως η τελευταία σχέση γίνεται

$$\left\|f_{\bar{x}}(t) \; - \; f_{\boldsymbol{x_0}}(t)\right\|^2 \leq \frac{n+1}{2} \cdot \sigma^2 \cdot \chi_n^2(a) \text{ με πιθανότητα } 1 - a.$$

Τελικά, φαίνεται πως η καμπύλη $f_{\bar{x}}(t)$ βρίσκεται μέσα σε μία ζώνη σταθερού πλάτους γύρω από το $f_{\boldsymbol{x_0}}(t)$ και χρησιμοποιώντας αυτό το συμπέρασμα μπορεί να πραγματοποιηθεί ο έλεγχος

$$H_0: f_{\bar{x}}(t) = f_{\boldsymbol{x_0}}(t) \text{ για κάθε } t \text{ με } -\pi < t < \pi$$

$$H_1: f_{\bar{x}}(t) \neq f_{\boldsymbol{x_0}}(t) \text{ για τουλάχιστον ένα } t \text{ με } -\pi < t < \pi.$$

Προφανώς, αν το $\bar{\boldsymbol{x}}$ είναι ο μέσος των παρατηρήσεων, τότε οι πιθανές ακραίες παρατηρήσεις μπορούν να εντοπισθούν ευκολότερα μιας και θα βρίσκονται εκτός της ζώνης που περιγράφεται παραπάνω. Επίσης, αν το $\bar{\boldsymbol{x}}$ είναι ο μέσος των παρατηρήσεων μίας ομάδας, τότε μια ζώνη με κέντρο το $f_{\bar{x}}(t)$ αποτελεί ένα διάστημα εμπιστοσύνης επιπέδου $(1-\alpha)\cdot 100\%$ για το $f_{\boldsymbol{x_0}}(t)$, έτσι ώστε αν το $f_{\boldsymbol{x_0}}(t)$ βρίσκεται εκτός του συγκεκριμένου διαστήματος, τότε υπάρχει ισχυρή ένδειξη ότι η παρατήρηση $\boldsymbol{x_0}$ δεν ανήκει στην ομάδα.

4.7 Εντοπισμός περιόδου σε δεδομένα χρονοσειρών

Στην περίπτωση μιας χρονοσειράς, είναι εξαιρετικά σημαντικό να είναι γνωστή η περίοδός της. Δηλαδή, το σύνολο των διαδοχικών παρατηρήσεων όπου επανεμφανίζονται συγκεκριμένα μοτίβα. Οι καμπύλες Andrews μπορούν να βοηθήσουν στον προσδιορισμό της περιόδου μιας χρονοσειράς.

Πιο συγκεκριμένα, οι Embrechts, Herzberg and Allen (1986) πρότειναν να ακολουθηθούν τα παρακάτω βήματα. Αρχικά, από την γνώση και τις πληροφορίες σχετικά με τη χρονοσειρά θεωρείται ένα σύνολο φυσικών αριθμών που είναι πιθανές τιμές της περιόδου. Έτσι έστω $A = \{n_i \in \mathbb{N},\ i = 1,2, \dots\}$ ένα υποσύνολο των φυσικών αριθμών που περιέχει όλα τα n_i που αποτελούν πιθανές τιμές της περιόδου.

Στη συνέχεια, επιλέγονται διαδοχικά όλα τα στοιχεία του συνόλου A και χαράσσονται τα διαγράμματα με τις καμπύλες Andrews, όπως στη διαδικασία που περιγράφεται στην Παράγραφο 4.4 σχετικά με την εύρεση ακραίων τιμών σε δεδομένα χρονοσειρών. Κάθε φορά το επιλεγμένο στοιχείο του συνόλου A χρησιμοποιείται ως περίοδος της χρονοσειράς.

Ακολούθως, ελέγχονται όλα τα διαγράμματα που έχουν παραχθεί για κάθε ξεχωριστή τιμή $n_i,\ i = 1,2, \ldots$, και αναζητείται εκείνη η τιμή $n_i,\ i = 1,2, \ldots$, για την οποία όλα τα διαγράμματα εμφανίζουν τις καμπύλες Andrews μέσα σε μία στενή ζώνη. Αναζητείται, λοιπόν, εκείνη η τιμή $n_i,\ i = 1,2, \ldots$, που παράγει καμπύλες με πολύ μικρές αποστάσεις μεταξύ τους. Έστω ότι η τιμή n_k είναι η τιμή που έχει την παραπάνω ιδιότητα.

Η τιμή n_k δεν είναι απαραίτητα η πραγματική περίοδος των δεδομένων, αλλά είναι σίγουρα ένα πολλαπλάσιο της περιόδου. Για να βρεθεί ποιο υποπολλαπλάσιο του n_k είναι η πραγματική περίοδος, επαναλαμβάνεται η διαδικασία της δημιουργίας των γραφημάτων που περιέχουν καμπύλες Andrews, όπως αυτή περιγράφεται παραπάνω, διαδοχικά για τους διαιρέτες της τιμής n_k, έστω $d_1, d_2, \ldots$. Όταν βρεθεί μία τιμή, έστω d_l για την οποία οι καμπύλες στα παραγόμενα γραφήματα δεν είναι πια κοντά μεταξύ τους, σταματάει τη διαδικασία και προκύπτει το συμπέρασμα ότι η πραγματική περίοδος της χρονοσειράς είναι ο ακριβώς προηγούμενος διαιρέτης της τιμής n_k, δηλαδή ο φυσικός αριθμός d_{l-1}.

Στην Παράγραφο 5.4.2 εξηγείται περαιτέρω η εφαρμογή της παραπάνω μεθόδου σε παράδειγμα δεδομένων χρονοσειράς.

Κ Ε Φ Α Λ Α Ι Ο 5

Εφαρμογές των Καμπυλών Andrews

5.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζονται εφαρμογές της μεθόδου των καμπυλών Andrews. Πιο συγκεκριμένα δίνονται εφαρμογές των καμπυλών σε ομαδοποιημένα και μη ομαδοποιημένα δεδομένα καθώς και σε δεδομένα χρονοσειρών. Για κάθε μία από τις παραπάνω περιπτώσεις δεδομένων γίνεται αξιοποίηση του συνόλου των δυνατοτήτων των καμπυλών Andrews ώστε να εξαχθούν χρήσιμα συμπεράσματα για τα σύνολα δεδομένων και τις παρατηρήσεις που αυτά περιέχουν.

5.2 Μελέτη εφαρμογής των καμπυλών Andrews σε μη ομαδοποιημένα δεδομένα

5.2.1 Γενικά στοιχεία

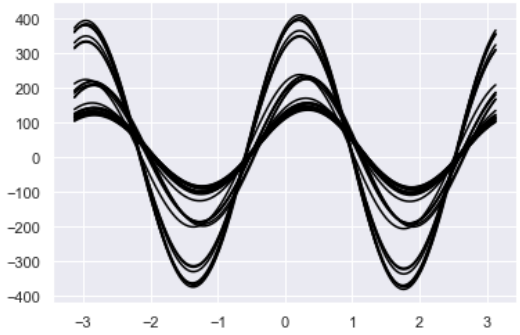
Για την εφαρμογή των καμπυλών Andrews σε μη ομαδοποιημένα δεδομένα θα χρησιμοποιηθεί το σύνολο δεδομένων που ονομάζεται Cars. Πλήρης περιγραφή του συγκεκριμένου συνόλου βρίσκεται στο Παράρτημα 3.

Για τους σκοπούς αυτής της εργασίας θα χρησιμοποιηθούν μόνο 25 από τις διαθέσιμες παρατηρήσεις και αντίστοιχα μόνο 6 από τις διαθέσιμες μεταβλητές. Οι μεταβλητές που θα χρησιμοποιηθούν είναι οι εξής:

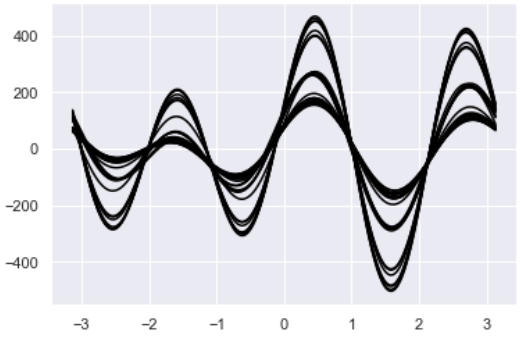
- MPG: απόσταση, σε μίλια, που μπορεί να διανύσει το όχημα με ένα λίτρο καυσίμου,
- Weight: βάρος του αυτοκινήτου,
- Drive Ratio: σχέση μετάδοσης της κίνησης από τον κινητήρα στους τροχούς,
- Horsepower: αριθμός ίππων του κινητήρα,
- Displacement: χωρητικότητα των κυλίνδρων του κινητήρα σε κυβικές ίντσες,
- Cylinders: αριθμός κυλίνδρων του κινητήρα.

5.2.2 Καμπύλες Andrews

Οι καμπύλες Andrews για το σύνολο δεδομένων Cars, δίνονται στο Γράφημα 5.1. Όπως μπορεί κανείς να παρατηρήσει στο γράφημα υπάρχουν 25 καμπύλες, όσες και οι παρατηρήσεις του συνόλου δεδομένων. Για την χάραξη των καμπυλών έχει χρησιμοποιηθεί η συνάρτηση του τύπου (2.1) και οι μεταβλητές των δεδομένων έχουν εισαχθεί στη συνάρτηση Andrews με τη σειρά που δίνονται στο Παράρτημα 3. Αντίστοιχα, στο Γράφημα 5.2 δίνονται οι καμπύλες με σειρά εισαγωγής των μεταβλητών αυτή που φαίνεται στον Πίνακα 5.1.



Γράφημα 5.1
Καμπύλες Andrews με σειρά εισαγωγής: MPG, Weight, Drive Ratio, Horsepower, Displacement, Cylinders



Γράφημα 5.2
Καμπύλες Andrews με σειρά εισαγωγής: Drive Ratio, Weight, Cylinders, MPG, Horsepower, Displacement

Όπως αναφέρθηκε παραπάνω και επιβεβαιώνεται και στη συγκεκριμένη εφαρμογή, η αλλαγή της σειράς εισαγωγής των μεταβλητών έχει ως αποτέλεσμα την αλλαγή της εικόνας των καμπυλών Andrews. Το γεγονός αυτό φαίνεται και στα Γραφήματα 5 - 1 και 5 - 2, όπου οι δύο εικόνες είναι σημαντικά διαφορετικές.

Η παραπάνω παρατήρηση δημιουργεί μια απορία σχετικά με το ποια είναι η καλύτερη σειρά εισαγωγής των μεταβλητών στη συνάρτηση του Andrews και η απάντηση που έχει δοθεί στη βιβλιογραφία αναφέρει πως οι μεταβλητές πρέπει να εισάγονται με βάση τη σημαντικότητά

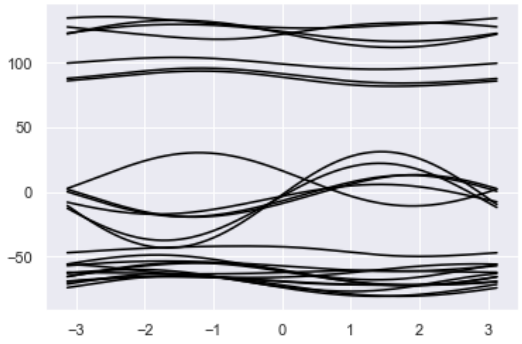
τους, δηλαδή πρώτη πρέπει να είναι η πιο σημαντική, στη συνέχεια η αμέσως λιγότερο σημαντική κλπ.

x_i	**Μεταβλητή**
x_1	Drive Ratio
x_2	Weight
x_3	Cylinders
x_4	MPG
x_5	Horsepower
x_6	Displacement

Πίνακας 5.1
Σειρά εισαγωγής μεταβλητών για την παραγωγή
του Γραφήματος 5.2

Όμως για το συγκεκριμένο σύνολο δεδομένων δεν είναι γνωστή κάποια πληροφορία σχετική με τη σημαντικότητα των μεταβλητών, επομένως η καλύτερη επιλογή είναι η χρήση των Κυρίων Συνιστωσών αντί των αρχικών μεταβλητών. Οι μέθοδος των Κύριων Συνιστωσών, χρησιμοποιεί τις υπάρχουσες μεταβλητές για να δημιουργήσει ένα σύνολο νέων μεταβλητών οι οποίες εξ ορισμού είναι ανεξάρτητες και κάθε μία περιλαμβάνει ένα μέρος της συνολικής πληροφορίας που περιέχεται στις αρχικές μεταβλητές. Λόγω της μεθόδου κατασκευής τους, η Πρώτη Κύρια Συνιστώσα περιλαμβάνει το μεγαλύτερο ποσοστό από την συνολική πληροφορία, ακολουθούμενη από τη Δεύτερη, την Τρίτη κλπ. Για την χάραξη των καμπυλών Andrews, μπορούν να χρησιμοποιηθούν είτε όλες είτε μερικές μόνο από τις πρώτες Κύριες Συνιστώσες, αρκεί να περιέχουν ικανοποιητικό ποσοστό από τη συνολική πληροφορία των δεδομένων. Ένα σύνηθες όριο που πρέπει να καλύπτεται είναι ένα ποσοστό της τάξεως του 80%.

Στο Γράφημα 5.3 δίνονται οι καμπύλες Andrews έχοντας χρησιμοποιήσει και τις 6 Κύριες Συνιστώσες που παρήχθησαν από τη μέθοδο των Κυρίων Συνιστωσών. Στο Γράφημα φαίνεται οι περισσότερες από τις καμπύλες να μοιάζουν με οριζόντιες ευθείες. Αυτό συμβαίνει γιατί η Πρώτη Κύρια Συνιστώσα περιέχει περίπου το 98% της πληροφορίας που περιέχεται στα δεδομένα, όπως φαίνεται και στον Πίνακα 5.2, και αυτή δεν πολλαπλασιάζεται με όρο που εξαρτάται από το t.



Γράφημα 5.3
Καμπύλες Andrews με χρήση των Κύριων Συνιστωσών

Κύριες Συνιστώσες	**Ποσοστό μεταβλητότητας**
Πρώτη	98,50%
Δεύτερη	1,42%
Τρίτη	0,09%
Τέταρτη	<0,01%
Πέμπτη	<0,01%
Έκτη	<0,01%

Πίνακας 5.2
Ποσοστό μεταβλητότητας που περιέχεται σε κάθε μία από τις έξι Κύριες Συνιστώσες που παράγονται

Αν κανείς συγκρίνει τα Γραφήματα 5.1, 5.2 και 5.3 θα παρατηρήσει ότι υπάρχουν τεράστιες διαφορές. Η μεγαλύτερη διαφοροποίηση εμφανίζεται στο Γράφημα 5.3 όπου έχουν χρησιμοποιηθεί οι Κύριες Συνιστώσες. Όπως αναφέρθηκε παραπάνω, η Πρώτη Κύρια Συνιστώσα που περιέχει το 98% της συνολικής πληροφορίας των δεδομένων, πολλαπλασιάζεται με τον πρώτο όρο της συνάρτησης που είναι το $\frac{1}{\sqrt{2}}$, επομένως δημιουργούνται καμπύλες που πλησιάζουν οπτικά τις οριζόντιες ευθείες γραμμές.

5.2.3 Ομαδοποίηση παρατηρήσεων

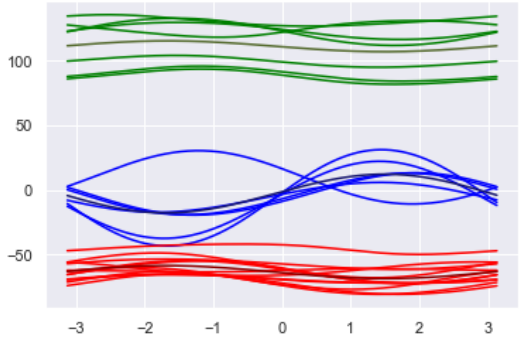
Οι παρατηρήσεις του συνόλου δεδομένων Cars δεν είναι ομαδοποιημένες με κάποιο τρόπο, δηλαδή δεν είναι γνωστό εκ των προτέρων αν υπάρχει κάποιος τρόπος ομαδοποίησης τους. Για να μπορέσει κανείς να βγάλει χρήσιμα συμπεράσματα για το συγκεκριμένο σύνολο θα ήταν σημαντική βοήθεια αν μπορούσε να καθορίσει μια ομαδοποίηση των παρατηρήσεων έτσι ώστε να μπορεί να αναλύσει κάθε ομάδα ξεχωριστά.

Οι καμπύλες Andrews δίνουν αυτή τη δυνατότητα στον αναλυτή, ώστε να μπορέσει μέσω της απεικόνισης κάθε παρατήρησης σε μία καμπύλη να βρει ομάδες καμπυλών, που όπως

αναφέρθηκε παραπάνω θα αντιστοιχούν σε ομάδες παρατηρήσεων. Για να εξαχθούν τα αντίστοιχα συμπεράσματα, θα χρησιμοποιηθεί το Γράφημα 5.3, στο οποίο έχει γίνει χρήση των Κύριων Συνιστωσών.

Όπως μπορεί κανείς να παρατηρήσει, δημιουργούνται τρεις καλά διαχωρισμένες ομάδες καμπυλών. Η πρώτη ομάδα αφορά τις καμπύλες που βρίσκονται ψηλότερα στο Γράφημα, η δεύτερη ομάδα αφορά εκείνες τις καμπύλες που βρίσκονται στη μέση του Γραφήματος και, τέλος, η τρίτη ομάδα αφορά τις παρατηρήσεις που βρίσκονται στο κάτω μέρος του Γραφήματος.

Στο Γράφημα 5.4 δίνονται οι καμπύλες του Γραφήματος 5.3 με τη διαφορά ότι αυτή τη φορά έχουν προστεθεί χρώματα για την διάκριση των καμπυλών κάθε ομάδας καθώς επίσης και οι καμπύλες των μέσων κάθε ομάδας με σκούρα χρώματα. Με βάση τα παραπάνω, η πρώτη ομάδα φαίνεται με πράσινο χρώμα, η δεύτερη με μπλε και η τρίτη με κόκκινο.



Γράφημα 5.4
Καμπύλες Andrews με χρήση των Κύριων Συνιστωσών και χρωματική διάκριση των ομάδων

Στον Πίνακα 5.3 δίνονται οι παρατηρήσεις μαζί με την ομάδα στην οποία ανήκουν καθώς και ο μέσος κάθε ομάδας. Όπως μπορεί κανείς να παρατηρήσει, οι τρεις ομάδες διαφέρουν σε όλες τις μεταβλητές, σε άλλες σε σημαντικό βαθμό και σε άλλες όχι. Οι μεγαλύτερες διαφορές φαίνεται να εμφανίζονται στις μεταβλητές Displacement και Cylinders όπου οι τιμές των παρατηρήσεων κάθε ομάδας είναι αρκετά κοντά στο μέσο της ομάδας και μακριά από τις τιμές των άλλων ομάδων.

Η ομαδοποίηση των παρατηρήσεων μπορεί να φανεί χρήσιμη για πολλούς λόγους. Ένας από τους λόγους αυτούς είναι η περαιτέρω ανάλυση των παρατηρήσεων κάθε ομάδας δεδομένου ότι τα μοντέλα αυτοκινήτων που ανήκουν σε αυτές είναι όμοια ως προς κάποιο ή κάποια χαρακτηριστικά. Ένας δεύτερος, είναι ότι στο εξής μπορεί να χρησιμοποιείται αυτή η ομαδοποίηση ώστε να γίνει ταξινόμηση μίας νέας παρατήρησης σε κάποια από τις τρεις υπάρχουσες ομάδες μέσω σύγκρισης της αντίστοιχης καμπύλης με τις καμπύλες των υπόλοιπων παρατηρήσεων όπως περιγράφεται στην Παράγραφο 4.3.

Ομάδα	Μοντέλο	MPG	Weight	Drive Ratio	Horsepower	Displacement	Cylinders
Ομάδα 3	Chevette	30.0	2.155	3.70	68	98	4
	Dodge Omni	30.9	2.230	3.37	75	105	4
	AMC Spirit	27.4	2.670	3.08	80	121	4
	Plymouth Horizon	34.2	2.200	3.37	70	105	4
	Mazda GLC	34.1	1.975	3.73	65	86	4
	Dodge Colt	35.1	1.915	2.97	80	98	4
	Honda Accord LX	29.5	2.135	3.05	68	98	4
	Datsun 210	31.8	2.020	3.70	65	85	4
	VW Scirocco	31.5	1.990	3.78	71	89	4
	VW Dasher	30.5	2.190	3.70	78	97	4
	VW Rabbit	31.9	1.925	3.78	71	89	4
	Fiat Strada	37.3	2.130	3.10	69	91	4
	Μέσος	32.0	2.128	3.44	72	97	4
Ομάδα 2	Mercury Zephyr	20.8	3.070	3.08	85	200	6
	Ford Mustang Ghia	21.9	2.910	3.08	109	171	6
	Chevy Citation	28.8	2.595	2.69	115	173	6
	Olds Omega	26.8	2.700	2.84	115	173	6
	Volvo 240 GL	17.0	3.140	3.50	125	163	6
	Peugeot 694 SL	16.2	3.410	3.58	133	163	6
	Μέσος	21.9	2.971	3.13	114	174	6
Ομάδα 1	Buick Estate Wagon	16.9	4.360	2.73	155	350	8
	Ford Country Squire Wagon	15.5	4.054	2.26	142	351	8
	Chrysler LeBaron Wagon	18.5	3.940	2.45	150	360	8
	Chevy Caprice Classic	17.0	3.840	2.41	130	305	8
	Ford LTD	17.6	3.725	2.26	129	302	8
	Mercury Grand Marquis	16.5	3.955	2.26	138	351	8
	Dodge St Regis	18.2	3.830	2.45	135	318	8
	Μέσος	17.2	3.958	2.4	140	334	8

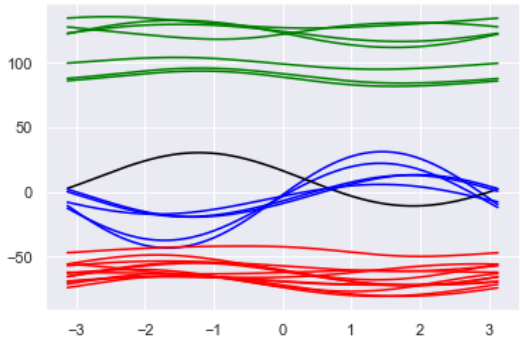
Πίνακας 5.3
Παρατηρήσεις και Μέσοι κάθε ομάδας

5.2.4 Εντοπισμός ακραίων τιμών

Η μέθοδος των καμπυλών Andrews μπορεί να χρησιμοποιηθεί αποτελεσματικά στον εντοπισμό ακραίων παρατηρήσεων. Ακραίες παρατηρήσεις είναι εκείνες που σε μία ή περισσότερες από τις μεταβλητές περιέχουν τιμές ακραία μεγαλύτερες ή μικρότερες από

εκείνες των υπόλοιπων παρατηρήσεων. Λόγω του παραπάνω γεγονότος, οι ακραίες παρατηρήσεις απεικονίζονται σε καμπύλες που συμπεριφέρονται διαφορετικά από τις υπόλοιπες καμπύλες. Μέσω αυτής της ιδιότητας, είναι αρκετά εύκολο να μπορέσει κανείς να εντοπίσει πιθανές ακραίες παρατηρήσεις εντοπίζοντας τις καμπύλες που συμπεριφέρονται διαφορετικά.

Για να πραγματοποιηθεί ο έλεγχος για ακραίες παρατηρήσεις, θα χρησιμοποιηθεί το Γράφημα 5.4 που δείχνει και την ομαδοποίηση των παρατηρήσεων του συνόλου δεδομένων Cars. Στο παραπάνω Γράφημα μπορεί κανείς να παρατηρήσει ότι μία από τις καμπύλες της ομάδας χρώματος μπλε συμπεριφέρεται διαφορετικά σε σχέση με τις υπόλοιπες καμπύλες της ομάδας της. Οι άλλες δύο ομάδες καμπυλών (κόκκινη, πράσινη) δεν φαίνεται να περιέχουν καμπύλες διαφορετικής συμπεριφοράς, επομένως δεν θα ελεγχθούν για τυχόν ακραίες τιμές.



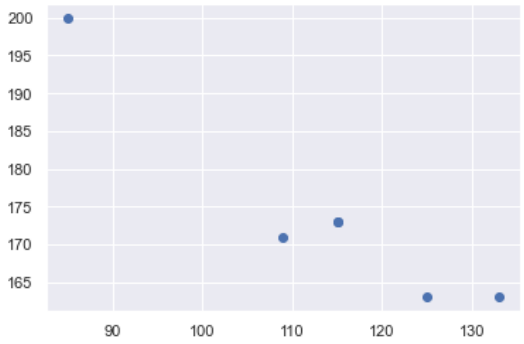
Γράφημα 5.5
Καμπύλες Andrews με χρήση των Κύριων Συνιστωσών, με χρωματική διάκριση των ομάδων και της πιθανής ακραίας παρατήρησης

Το Γράφημα 5.5 είναι όμοιο του Γραφήματος 5.4 με μοναδική διαφορά τον χρωματισμό, με μαύρο χρώμα, της καμπύλης που αντιστοιχεί σε πιθανή ακραία παρατήρηση. Όπως φαίνεται και στο σχήμα, οι καμπύλες της μπλε ομάδας είναι κυρτές στο διάστημα $(-\pi, 0)$ και κοίλες στο διάσημα $(0, \pi)$. Σε αντίθεση με τις υπόλοιπες καμπύλες της ομάδας της, η καμπύλη μαύρου χρώματος είναι κοίλη στο διάστημα $(-\pi, 0)$ και κυρτή στο διάστημα $(0, \pi)$ και αυτός ακριβώς είναι ο λόγος που αποτελεί την καμπύλη μίας πιθανής ακραίας παρατήρησης.

Μοντέλο	MPG	Weight	Drive Ratio	Horsepower	Displacement	Cylinders
Mercury Zephyr	20.8	3.070	3.08	85	200	6
Ford Mustang Ghia	21.9	2.910	3.08	109	171	6
Chevy Citation	28.8	2.595	2.69	115	173	6
Olds Omega	26.8	2.700	2.84	115	173	6
Volvo 240 GL	17.0	3.140	3.50	125	163	6
Peugeot 694 SL	16.2	3.410	3.58	133	163	6
Μέσος Ομάδας	**21.9**	**2.971**	**3.13**	**114**	**174**	**6**

Πίνακας 5.4
Παρατηρήσεις μπλε ομάδας και Μέσος ομάδας

Στον Πίνακα 5.4 δίνονται οι παρατηρήσεις που αντιστοιχούν στις καμπύλες της μπλε ομάδας καμπυλών. Η καμπύλη, του Γραφήματος 5.5, με μαύρο χρώμα αντιστοιχεί στο μοντέλο Mercury Zephyr. Οι τιμές των διάφορων μεταβλητών για τη συγκεκριμένη παρατήρηση είναι κοντά στο Μέσο της ομάδας και κοντά στις τιμές των υπόλοιπων παρατηρήσεων με εξαίρεση τις μεταβλητές Horsepower και Displacement. Στις τελευταίες δύο μεταβλητές, οι τιμές του μοντέλου Mercury Zephyr είναι ακραίες σε σχέση τόσο με το Μέσο της ομάδας όσο και με τις τιμές των υπόλοιπων παρατηρήσεων. Αυτό το γεγονός απεικονίζεται καλύτερα στο Γράφημα 5.6 που είναι ένα διάγραμμα διασποράς, όπου στον άξονα x βρίσκεται η μεταβλητή Horsepower και στον άξονα y βρίσκεται η μεταβλητή Displacement.



Γράφημα 5.6
Διάγραμμα διασποράς των μεταβλητών
Horsepower και Displacement

Το σημείο που διακρίνεται πάνω και αριστερά στο σχήμα αντιστοιχεί στις τιμές του μοντέλου Mercury Zephyr και είναι προφανές ότι είναι αρκετά μακριά σε σχέση με τα σημεία που αντιστοιχούν στις υπόλοιπες παρατηρήσεις.

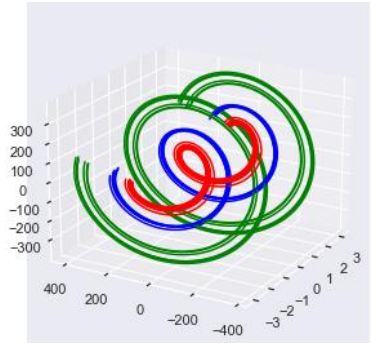
Από όλα τα παραπάνω μπορεί κανείς να κρίνει το μοντέλο Mercury Zephyr ως ακραία παρατήρηση. Η παραπάνω διαπίστωση βασίζεται στις μεταβλητές Horsepower και Displacement στις οποίες οι τιμές της παρατήρησης ήταν ακραίες. Ο ερευνητής, ανάλογα με την ανάλυση που θέλει να πραγματοποιήσει μπορεί να αποφασίσει τον τρόπο διαχείρισης της συγκεκριμένης παρατήρησης. Για τους σκοπούς της παρούσας εργασίας η συγκεκριμένη παρατήρηση θα εξακολουθεί να χρησιμοποιείται κανονικά και να συμπεριλαμβάνεται στα επόμενα.

5.2.5 Εντοπισμός σημαντικών και μη μεταβλητών

Σε πολλές περιπτώσεις ανάλυσης συνόλων δεδομένων είναι σημαντικό να γνωρίζει ο ερευνητής ποιες από τις μεταβλητές του συνόλου είναι πολύ σημαντικές και ποιες λιγότερο. Η μέθοδος των καμπυλών Andrews μπορεί να χρησιμοποιηθεί για να βρει κανείς αυτή την πληροφορία παρατηρώντας τα γραφήματα που παράγονται.

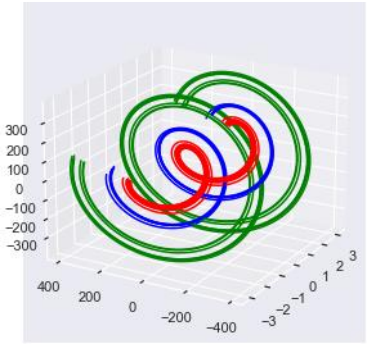
Η διαδικασία που πρέπει να ακολουθηθεί, όπως περιγράφεται και στην Παράγραφο 4.5, είναι να δημιουργηθούν γραφήματα που περιέχουν καμπύλες Andrews, όμως σε κάθε ένα από αυτά θα αντικαθίσταται μία από τις μεταβλητές με 0. Στη συνέχεια, αν το παραγόμενο γράφημα έχει σημαντικές διαφορές με το αρχικό, που περιέχει όλες τις μεταβλητές, τότε η μεταβλητή που έχει αντικατασταθεί με 0 περιέχει σημαντικό μέρος της πληροφορίας που υπάρχει στα δεδομένα. Αντίθετα, αν δεν υπάρχει μεγάλη διαφορά στα γραφήματα, τότε η πληροφορία που περιέχεται στη μεταβλητή που αντικαταστάθηκε με 0 είναι μικρή.

Στο σύνολο δεδομένων Cars περιέχονται έξι μεταβλητές οι οποίες και θα αντικατασταθούν με 0 διαδοχικά για τη χάραξη των γραφημάτων. Παρακάτω δίνονται τα γραφήματα τα οποία αναφέρθηκαν παραπάνω καθώς και το γράφημα που περιέχει όλες τις μεταβλητές. Αυτή τη φορά έχει γίνει χρήση των τρισδιάστατων καμπυλών Andrews, όπως αυτές περιγράφονται στην Παράγραφο 3.4, ενώ ως $f_x(t)$ και $g_x(t)$ έχουν χρησιμοποιηθεί αντίστοιχα οι συναρτήσεις των τύπων (2.1) και (3.1).

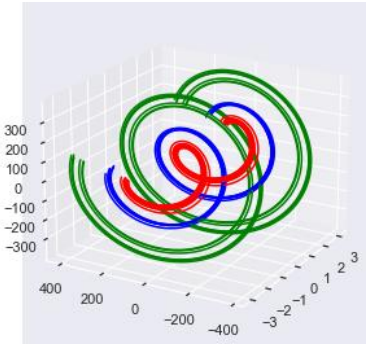


Γράφημα 5.7
Καμπύλες Andrews στις τρεις διαστάσεις με χρήση όλων των μεταβλητών

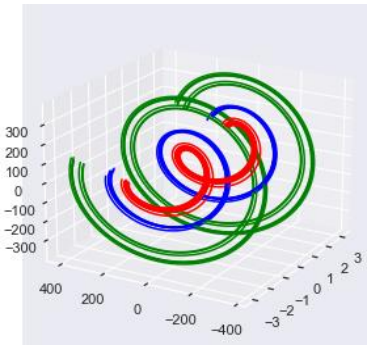
Στο Γράφημα 5.7 δίνονται οι καμπύλες Andrews στις τρεις διαστάσεις χωρίς να έχει γίνει αντικατάσταση με 0 σε καμία από αυτές. Στα Γραφήματα 5.8, 5.9, 5.10, 5.11, 5.12 και 5.13 δίνονται οι καμπύλες Andrews, στις οποίες έχουν αντικατασταθεί με 0 οι μεταβλητές MPG, Weight, Drive Ratio, Horsepower, Displacement και Cylinders, αντίστοιχα.



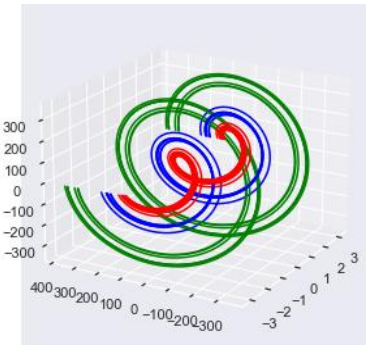
Γράφημα 5.8
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής MPG με 0



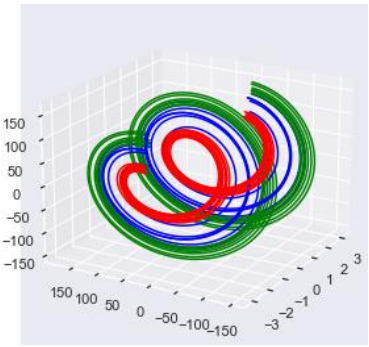
Γράφημα 5.9
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής Weight με 0



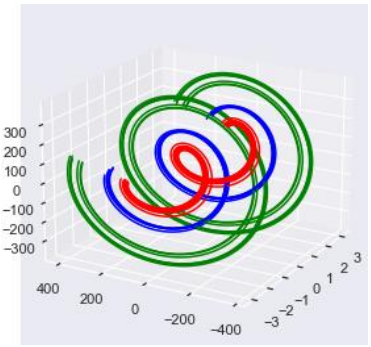
Γράφημα 5.10
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής Drive Ratio με 0



Γράφημα 5.11
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής Horsepower με 0



Γράφημα 5.12
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής Displacement με 0



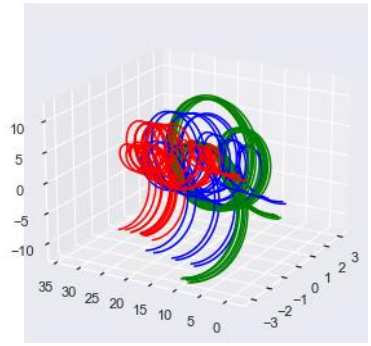
Γράφημα 5.13
Καμπύλες Andrews στις τρεις διαστάσεις μετά την αντικατάσταση της μεταβλητής Cylinders με 0

Όπως φαίνεται από τη σύγκριση του Γραφήματος 5.7 με το Γράφημα 5.8, η αντικατάσταση της μεταβλητής MPG με 0 δεν προκάλεσε σχεδόν καμία αλλαγή στην εικόνα των καμπυλών και επομένως μπορούμε να συμπεράνουμε ότι δεν περιέχει μεγάλο μέρος της πληροφορίας που υπάρχει στα δεδομένα. Αντίστοιχα, από τη σύγκριση των Γραφημάτων 5.7 και 5.9 παρατηρείται ότι ούτε η μεταβλητή Weight περιέχει μεγάλο μέρος πληροφορίας, ενώ το ίδιο παρατηρείται από τη σύγκριση των Γραφημάτων 5.7 και 5.10 για τη μεταβλητή Drive Ratio.

Αντίθετα με τα παραπάνω, υπάρχει διαφοροποίηση της εικόνας των καμπυλών μετά την αντικατάσταση με 0 των μεταβλητών Horsepower και Displacement. Στο Γράφημα 5.11 παρατηρείται ότι η αντικατάσταση της μεταβλητής Horsepower με 0 έφερε τις ομάδες κόκκινη και μπλε πιο κοντά κάνοντας την διάκριση μεταξύ τους δυσκολότερη, ενώ ταυτόχρονα κράτησε αρκετά μεγάλη της απόσταση μεταξύ της μπλε και πράσινης ομάδας. Το συμπέρασμα από τα παραπάνω είναι πως η συγκεκριμένη μεταβλητή περιέχει πληροφορία για τα δεδομένα σε σχετικά μεγάλο βαθμό. Ομοίως, στο Γράφημα 5.12 παρατηρείται μεγάλη διαφορά σε σχέση με το Γράφημα 5.7. Πιο συγκεκριμένα, φαίνεται οι τρεις ομάδες να έχουν έρθει πιο κοντά, με τη μεταξύ τους διάκριση να είναι πλέον εξαιρετικά δύσκολη ενώ παράλληλα οι καμπύλες της μίας ομάδας μπερδεύονται με καμπύλες άλλων ομάδων. Αυτό σημαίνει ότι η μεταβλητή Displacement περιέχει πολύ μεγάλο μέρος από την πληροφορία που υπάρχει στα δεδομένα, ενώ παίζει πολύ σημαντικό ρόλο και στον διαχωρισμό των ομάδων.

Τέλος, από την αντικατάσταση της μεταβλητής Cylinders με 0 δεν φαίνεται να υπάρχει μεγάλη αλλαγή στην εικόνα των καμπυλών στο Γράφημα 5.13.

Για φανεί λίγο πιο χαρακτηριστικά το παραπάνω αποτέλεσμα, στο επόμενο βήμα θα γίνει αντικατάσταση των μεταβλητών Horsepower και Displacement με 0 στο ίδιο γράφημα. Πιο συγκεκριμένα στο Γράφημα 5.14 βλέπουμε την εικόνα των καμπυλών αν γίνει ταυτόχρονη αντικατάσταση των δύο παραπάνω μεταβλητών με 0.



Γράφημα 5.14
Καμπύλες Andrews στις τρεις διαστάσεις μετά
την αντικατάσταση των μεταβλητών Horsepower
και Displacement με 0

Από τη σύγκριση των Γραφημάτων 5.7 και 5.14 βλέπουμε ότι οι δύο αυτές μεταβλητές περιέχουν πολύ μεγάλο μέρος της πληροφορίας των δεδομένων αφού υπάρχει μεγάλη διαφορά στις καμπύλες. Η διαφορά έγκειται τόσο στο σχήμα και τη θέση των καμπυλών όσο και στη διάκριση μεταξύ των ομάδων. Αυτό σημαίνει πως οι δύο αυτές μεταβλητές είναι εκείνες που διαχωρίζουν τις ομάδες μεταξύ τους και περιέχουν όλη τη σχετική πληροφορία.

5.2.6 Συμπεράσματα

Τα συμπεράσματα από την παραπάνω ανάλυση για το σύνολο δεδομένων Cars είναι αρκετά και περιγράφονται στη συγκεκριμένη παράγραφο.

Αρχικά, όπως αναφέρθηκε παραπάνω, στο σύνολο υπάρχουν τρεις ομάδες παρατηρήσεων που σημαίνει ότι θα μπορούσε κανείς να χωρίσει τα αντίστοιχα αυτοκίνητα σε τρεις κατηγορίες. Με δεδομένη αυτή την ομαδοποίηση των παρατηρήσεων θα μπορούσε κανείς να χρησιμοποιήσει τις καμπύλες Andrews για να ταξινομήσει νέες παρατηρήσεις σε μία από τις τρεις ομάδες. Επίσης, προέκυψε πως η πληροφορία για το διαχωρισμό των τριών ομάδων παρατηρήσεων βρίσκεται κυρίως στις μεταβλητές Horsepower και Displacement. Συμπεραίνει κανείς, λοιπόν, πως αυτές είναι οι μεταβλητές στις οποίες διαφέρουν σημαντικά οι τρεις ομάδες.

Τέλος, εντοπίστηκε μία παρατήρηση της οποίας η καμπύλη συμπεριφέρεται διαφορετικά και μετά από την κατάλληλη διερεύνηση των δεδομένων προέκυψε το συμπέρασμα ότι αποτελεί ακραία παρατήρηση μιας και σε δύο από τις μεταβλητές έπαιρνε ακραίες τιμές.

5.3 Μελέτη εφαρμογής των καμπυλών Andrews σε ομαδοποιημένα δεδομένα

5.3.1 Γενικά στοιχεία

Για την εφαρμογή των καμπυλών Andrews σε ομαδοποιημένα δεδομένα θα χρησιμοποιηθεί ένα σύνολο δεδομένων που περιλαμβάνει μερικές από τις εγγραφές της βάσης δεδομένων της εταιρίας Nimber Tech. Πληρέστερη περιγραφή της εταιρίας και του συνόλου δεδομένων δίνεται στο Παράρτημα 4.

Στο σύνολο δεδομένων υπάρχουν 24 παρατηρήσεις που αφορούν μεταφορές που ζητήθηκαν από χρήστες της πλατφόρμας. Η τελευταία παρατήρηση χρησιμοποιείται μόνο στην Παράγραφο 5.3.4 για την αξιοποίηση και αξιολόγηση της δυνατότητας ταξινόμησης που έχουν οι καμπύλες Andrews.

Για κάθε παρατήρηση έχουν συλλεχθεί οι παρακάτω μεταβλητές:

- Distance: συνολική απόσταση, σε χιλιόμετρα, που θα χρειαστεί να διανύσει ο μεταφορέας ώστε να ολοκληρώσει τη μεταφορά,
- Price: χρηματικό ποσό, σε Νορβηγικές Κορώνες, που θα πληρώσει ο αποστολέας με την ολοκλήρωση της μεταφοράς,
- Weight: συνολικό βάρος του μεταφερόμενου αντικειμένου, σε κιλά,
- Size: μέγεθος του αντικειμένου, σε μία κλίμακα από ένα έως και πέντε με το νούμερο ένα να αντιστοιχεί σε αντικείμενα που χωρούν σε μία τσέπη, το δύο σε αντικείμενα που χωρούν

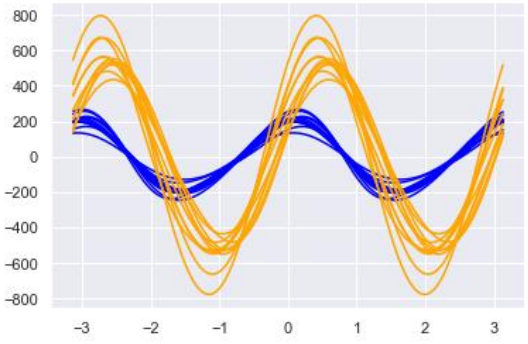
σε μία τσάντα, το τρία σε αντικείμενα που χωρούν σε ένα συμβατικό αυτοκίνητο, το τέσσερα σε αντικείμενα που απαιτούν μεγαλύτερα αυτοκίνητα και, τέλος, το πέντε σε αντικείμενα που απαιτούν φορτηγό για τη μεταφορά τους,

- State: αν το συγκεκριμένο αίτημα για μεταφορά έχει ολοκληρωθεί από κάποιον μεταφορέα ή αν δεν βρέθηκε κάποιος να το μεταφέρει.

Οι παρατηρήσεις χωρίζονται σε δύο ομάδες με βάση τη μεταβλητή State, τις επιτυχημένες και τις μη επιτυχημένες μεταφορές. Ως επιτυχημένες θεωρούνται εκείνες που πραγματοποιήθηκαν και ως μη επιτυχημένες εκείνες που δεν πραγματοποιήθηκαν αφού δε μπόρεσε να βρεθεί μεταφορέας διαθέσιμος να τις ολοκληρώσει.

5.3.2 Καμπύλες Andrews

Στο Γράφημα 5.15 φαίνονται οι καμπύλες Andrews για το σύνολο δεδομένων της εταιρίας Nimber Tech. Για τη χάραξη των καμπυλών έχει χρησιμοποιηθεί η συνάρτηση του τύπου (3.2) και η σειρά εισαγωγής των μεταβλητών είναι η εξής Distance, Price, Weight, Size. Η εισαγωγή των μεταβλητών με την παραπάνω σειρά έγινε λόγω της προτεραιοποίησης που χρησιμοποιείται από την εταιρία ως προς τη σημαντικότητα των μεταβλητών για τον τομέα δραστηριοποίησής της.



Γράφημα 5.15
Καμπύλες Andrews με την εξής σειρά εισαγωγής των μεταβλητών Distance, Price, Weight, Size

Στο γράφημα είναι ορατός ένας διαχωρισμός των καμπυλών σε δύο ομάδες. Με μπλε χρώμα έχουν χρωματισθεί οι καμπύλες που αφορούν επιτυχημένες μεταφορές ενώ με πορτοκαλί εκείνες που αφορούν μη επιτυχημένες. Ακόμα και αν δεν υπήρχε διαφορετικός χρωματισμός στις καμπύλες των δύο ομάδων, θα μπορούσε κανείς εύκολα να διακρίνει τον διαχωρισμό και τη διαφορετική συμπεριφορά των καμπυλών των δύο ομάδων. Στο Παράρτημα 4 δίνονται τα αριθμητικά περιγραφικά μέτρα ανά ομάδα καθώς και συνολικά για όλες τις παρατηρήσεις και έτσι γίνεται πιο ξεκάθαρος ο λόγος του διαχωρισμού των ομάδων.

Ακόμα, στο γράφημα δεν φαίνεται να υπάρχει κάποια καμπύλη που να συμπεριφέρεται με διαφορετικό τρόπο ή να βρίσκεται πολύ μακριά συγκριτικά με τις υπόλοιπες της ομάδας στην

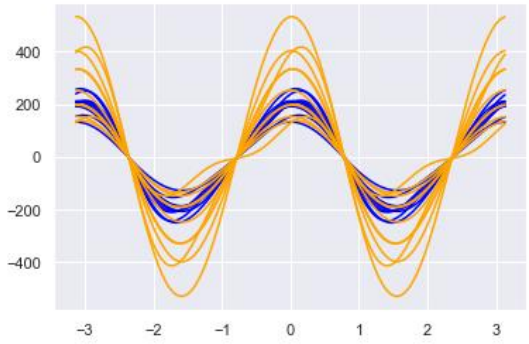
οποία ανήκει. Μια τέτοια καμπύλη θα υποδείκνυε την πιθανή ύπαρξη ακραίας παρατήρησης στα δεδομένα κάτι που θα έπρεπε να εξετασθεί περαιτέρω.

5.3.3 Εντοπισμός σημαντικών και μη μεταβλητών

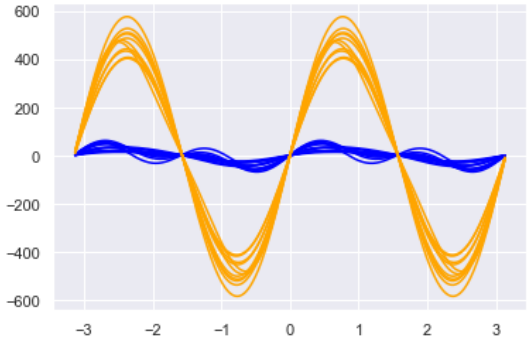
Μία από τις πλέον σημαντικές χρήσεις της μεθόδου των καμπυλών Andrews είναι αυτή της εύρεσης των σημαντικών και μη σημαντικών μεταβλητών ενός συνόλου δεδομένων. Στο σύνολο δεδομένων της εταιρίας Nimber Tech περιέχονται τέσσερις μεταβλητές οι οποίες δίνονται στο Παράρτημα 4 με σειρά σημαντικότητας για την δραστηριότητα της εταιρίας. Στην υποπαράγραφο αυτή θα εξεταστεί η σημαντικότητα των μεταβλητών αυτών και θα βρεθούν εκείνες οι μεταβλητές που περιέχουν το μεγαλύτερο ποσοστό από την πληροφορία που υπάρχει στα δεδομένα και θα αξιολογηθεί η σειρά σημαντικότητας που δόθηκε από την εταιρία.

Για να ελεγχθεί η σημαντικότητα των μεταβλητών θα πρέπει να αντικατασταθούν διαδοχικά όλες οι μεταβλητές με 0 και να χαραχθούν οι καμπύλες Andrews. Στη συνέχεια συγκρίνεται κάθε ένα από τα παραχθέντα γραφήματα με εκείνο που περιλαμβάνει και τις τέσσερις μεταβλητές. Όσο μεγαλύτερη είναι η διαφορά καθενός από τα γραφήματα με το αρχικό τόσο μεγαλύτερο μέρος της πληροφορίας των δεδομένων περιέχεται στη μεταβλητή που αντικαταστάθηκε κατά την κατασκευή του.

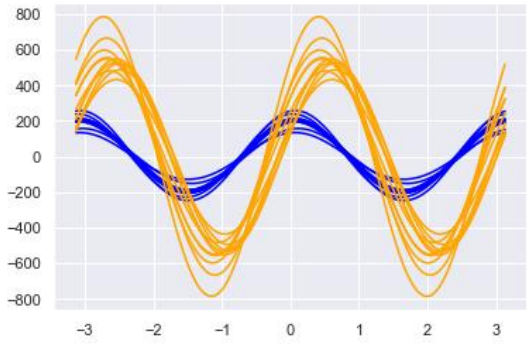
Αρχικά θα αντικατασταθεί με 0 η μεταβλητή Distance και το αποτέλεσμα φαίνεται στο Γράφημα 5.16. Όπως παρατηρεί κανείς συγκρίνοντας το συγκεκριμένο γράφημα με το Γράφημα 5.15, υπάρχει τεράστια διαφορά στην εικόνα αφού παρατηρείται πως πλέον δεν είναι εμφανής ο διαχωρισμός των δύο ομάδων παρατηρήσεων. Αυτό σημαίνει πως η μεταβλητή Distance περιέχει μεγάλο μέρος από την πληροφορία σχετικά με τον διαχωρισμό των ομάδων και κρίνεται ως μία από της σημαντικές μεταβλητές του συνόλου.



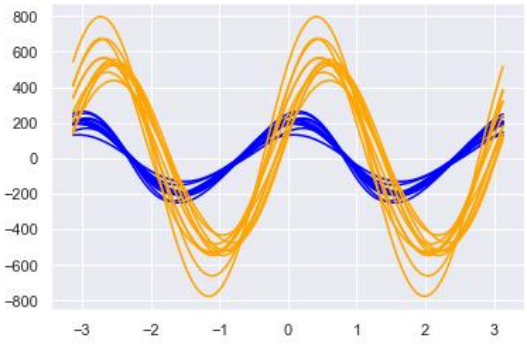
Γράφημα 5.16
Καμπύλες Andrews με την εξής σειρά εισαγωγής των μεταβλητών Distance, Price, Weight, Size και αντικατάσταση της μεταβλητής Distance με 0



Γράφημα 5.17
Καμπύλες Andrews με την εξής σειρά εισαγωγής των μεταβλητών Distance, Price, Weight, Size και αντικατάσταση της μεταβλητής Price με 0



Γράφημα 5.18
Καμπύλες Andrews με την εξής σειρά εισαγωγής των μεταβλητών Distance, Price, Weight, Size και αντικατάσταση της μεταβλητής Weight με 0



Γράφημα 5.19
Καμπύλες Andrews με την εξής σειρά εισαγωγής των μεταβλητών Distance, Price, Weight, Size και αντικατάσταση της μεταβλητής Size με 0

Στη συνέχεια, αντικαθίσταται με 0 η μεταβλητή Price. Το Γράφημα 5.17 δείχνει τις παραγόμενες καμπύλες, όπου μετά τη σύγκριση με το Γράφημα 5.15 είναι προφανές ότι η εικόνα των καμπυλών είναι κατά πολύ διαφορετική. Παρατηρείται ότι εξακολουθεί να διατηρείται ο διαχωρισμός των ομάδων όμως αυτή τη φορά υπάρχει μεγάλη διαφοροποίηση στη συμπεριφορά των καμπυλών. Αυτό οδηγεί στο συμπέρασμα ότι η μεταβλητή Price περιέχει μεγάλο ποσοστό από τη συνολική πληροφορία των δεδομένων και επομένως κρίνεται ως μία από τις σημαντικές μεταβλητές του συγκεκριμένου συνόλου.

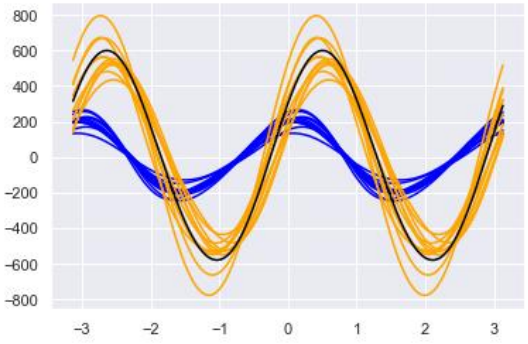
Στα Γραφήματα 5.18 και 5.19 φαίνονται οι καμπύλες Andrews μετά την αντικατάσταση με 0 των μεταβλητών Weight και Size, αντίστοιχα. Από τη σύγκριση και των δύο αυτών γραφημάτων με το Γράφημα 5.15 φαίνεται πως η διαφορά τους είναι μικρή και επομένως τόσο η μεταβλητή Weight όσο και η μεταβλητή Size δεν περιέχουν μεγάλο μέρος από την πληροφορία των δεδομένων. Για τον παραπάνω λόγο κρίνονται ως οι λιγότερο σημαντικές μεταβλητές του συνόλου.

5.3.4 Ταξινόμηση παρατήρησης

Μια από τις χρήσεις των καμπυλών Andrews είναι και η ταξινόμηση νέων παρατηρήσεων σε μία από τις υπάρχουσες ομάδες ενός συνόλου δεδομένων. Η διαδικασία απαιτεί την χάραξη της καμπύλης Andrews που αντιστοιχεί στη νέα παρατήρηση και στη συνέχεια σύγκριση της με τις καμπύλες των παρατηρήσεων που ανήκουν σε γνωστές ομάδες. Αν η νέα καμπύλη συμπεριφέρεται με τον ίδιο τρόπο όπως κάποια από τις ομάδες, τότε μπορεί κανείς να ισχυριστεί ότι η αντίστοιχη παρατήρηση ανήκει στη συγκεκριμένη ομάδα.

Για την αξιοποίηση της δυνατότητας αυτής των καμπυλών Andrews, θα χρησιμοποιηθεί η τελευταία παρατήρηση του συνόλου, η οποία δεν έχει χρησιμοποιηθεί παραπάνω. Όπως φαίνεται και στο Παράρτημα 4, η παρατήρηση αυτή ανήκει στην ομάδα των μη επιτυχημένων μεταφορών κάτι το οποίο θα αποδειχθεί και με τη χρήση των καμπυλών.

Στο Γράφημα 5.20 φαίνονται οι καμπύλες για όλες τις παρατηρήσεις του συνόλου δεδομένων, όπου με μπλε χρώμα είναι οι επιτυχημένες, με πορτοκαλί οι μη επιτυχημένες και με μαύρο αυτή της οποίας την ομάδα δεν είναι γνωστή.



Γράφημα 5.20
Καμπύλες Andrews όπου με μαύρο χρώμα
δίνεται η καμπύλη που πρέπει να ταξινομηθεί

Από την παρατήρηση του παραπάνω γραφήματος, φαίνεται πως η καμπύλη μαύρου χρώματος συμπεριφέρεται με πολύ παρόμοιο τρόπο με τις καμπύλες χρώματος πορτοκαλί. Επομένως, αυτό αποτελεί ένδειξη πως η αντίστοιχη παρατήρηση μπορεί να ταξινομηθεί στην ομάδα των μη επιτυχημένων παρατηρήσεων.

Με τον παραπάνω τρόπο είναι εφικτό να ταξινομηθεί οποιαδήποτε νέα παρατήρηση είτε σε μία από τις δύο υπάρχουσες ομάδες είτε να ελεγχθεί ως ακραία παρατήρηση.

5.3.5 Συμπεράσματα

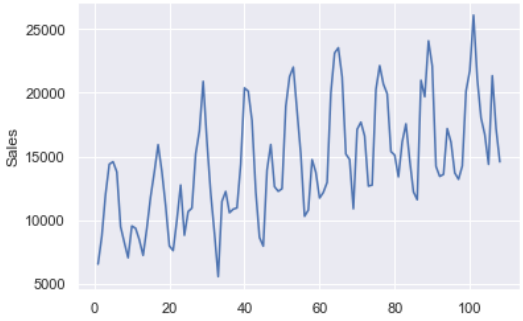
Από την ανάλυση που έγινε μέσω των καμπυλών Andrews στο σύνολο δεδομένων της εταιρίας Nimber, προέκυψαν κάποια συμπεράσματα. Αρχικά, παρατηρήθηκε πως η ομαδοποίηση που υπάρχει στα δεδομένα, και εμφανίζεται στη μεταβλητή State, αναδεικνύεται και με τη χρήση των γραφημάτων των καμπυλών Andrews. Επίσης, αποδείχθηκε ότι μπορεί να γίνει χρήση των καμπυλών για να ταξινομηθεί η νέα παρατήρηση, η οποία ταξινομήθηκε επιτυχώς στην κατηγορία των μη επιτυχημένων μεταφορών.

Ακόμα έγινε χρήση της κατάλληλης μεθόδου των καμπυλών Andrews ώστε να βρεθεί το ποιες από τις μεταβλητές είναι σημαντικές και ποιες όχι και δείχθηκε πως οι πλέον σημαντικές από τις μεταβλητές είναι η Distance και η Price. Τις ίδιες μεταβλητές θεωρεί ως σημαντικές και η ίδια η εταιρία.

5.4 Μελέτη εφαρμογής των καμπυλών Andrews σε δεδομένα χρονοσειρών

5.4.1 Γενικά στοιχεία

Για την εφαρμογή της μεθόδου των καμπυλών Andrews σε δεδομένα χρονοσειρών, θα χρησιμοποιηθεί η χρονοσειρά του συνόλου δεδομένων Monthly Car Sales. Το συγκεκριμένο σύνολο δεδομένων αφορά πωλήσεις αυτοκινήτων στην επαρχία Quebec του Καναδά. Η χρονοσειρά περιλαμβάνει τις μηνιαίες πωλήσεις από τον Ιανουάριο του 1960 μέχρι και τον Δεκέμβριο του 1968. Στο Γράφημα 5.21 δίνεται η εικόνα των τιμών της χρονοσειράς που φαίνονται στο Παράρτημα 5.



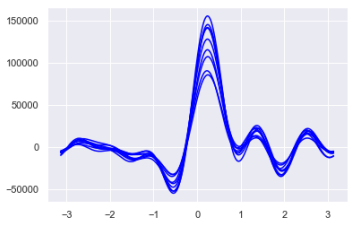
Γράφημα 5.21
Η καμπύλη της χρονοσειράς Monthly Car Sales

Σε αυτή την παράγραφο θα χρησιμοποιηθούν οι καμπύλες Andrews ώστε να βρεθεί η περίοδος της χρονοσειράς. Επίσης θα εντοπισθούν τυχόν ακραίες παρατηρήσεις ενώ αν δεν υπάρχουν τέτοιες θα προστεθούν ώστε να χρησιμοποιηθούν οι καμπύλες για τον εντοπισμό τους.

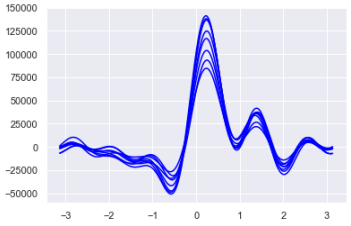
5.4.2 Εντοπισμός περιόδου χρονοσειράς

Σε αυτή την υποπαράγραφο θα πραγματοποιηθούν όλοι οι απαραίτητοι έλεγχοι ώστε να γίνει ο εντοπισμός της περιόδου της χρονοσειράς.

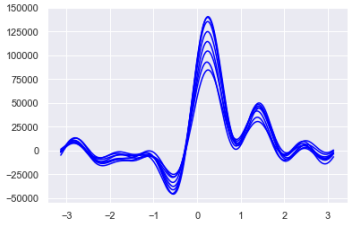
Αρχικά, μιας και τα δεδομένα είναι μηνιαίες μετρήσεις των πωλήσεων, θα πραγματοποιηθεί η διαδικασία της Παραγράφου 4.7 για περίοδο ίση με 12 παρατηρήσεις. Επομένως, θα πρέπει να παραχθούν δώδεκα γραφήματα, όπου το πρώτο θα περιλαμβάνει εννέα καμπύλες και όλα τα υπόλοιπα από οκτώ. Όλα τα γραφήματα θα παραχθούν κάνοντας χρήση της συνάρτησης του τύπου (2.1) και σε κάθε ένα θα αλλάζει διαδοχικά η σειρά με την οποία θα εισάγονται οι μήνες στη συνάρτηση.



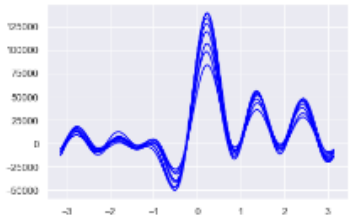
Γράφημα 5.22
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο



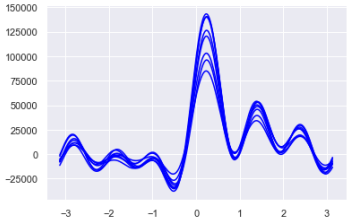
Γράφημα 5.23
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο



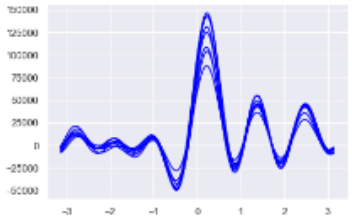
Γράφημα 5.24
Καμπύλες Andrews με αρχικό μήνα τον Μάρτιο



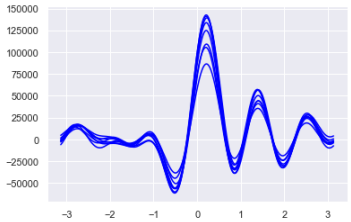
Γράφημα 5.25
Καμπύλες Andrews με αρχικό μήνα τον Απρίλιο



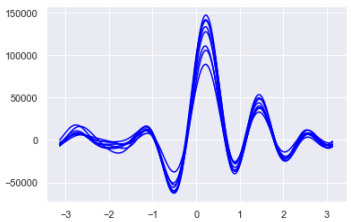
Γράφημα 5.26
Καμπύλες Andrews με αρχικό μήνα τον Μάιο



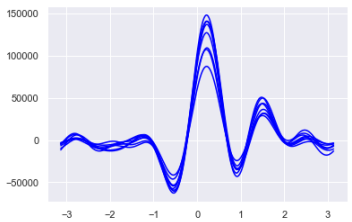
Γράφημα 5.27
Καμπύλες Andrews με αρχικό μήνα τον Ιούνιο



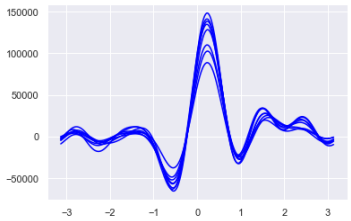
Γράφημα 5.28
Καμπύλες Andrews με αρχικό μήνα τον Ιούλιο



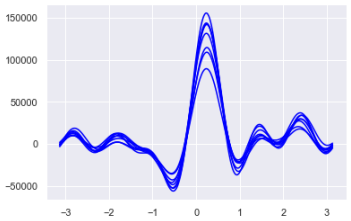
Γράφημα 5.29
Καμπύλες Andrews με αρχικό μήνα τον Αύγουστο



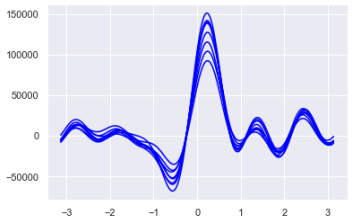
Γράφημα 5.30
Καμπύλες Andrews με αρχικό μήνα τον Σεπτέμβριο



Γράφημα 5.31
Καμπύλες Andrews με αρχικό μήνα τον Οκτώβριο



Γράφημα 5.32
Καμπύλες Andrews με αρχικό μήνα τον Νοέμβριο



Γράφημα 5.33
Καμπύλες Andrews με αρχικό μήνα τον Δεκέμβριο

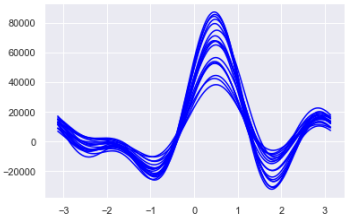
Στο Γράφημα 5.22 φαίνονται οι καμπύλες Andrews όπου η σειρά εισαγωγής των μηνών είναι Ιανουάριος, Φεβρουάριος, …, Δεκέμβριος. Αντίστοιχα στο Γράφημα 5.23 φαίνονται οι καμπύλες με την ακόλουθη σειρά εισαγωγής των μηνών Φεβρουάριος, Μάρτιος, …, Δεκέμβριος, Ιανουάριος. Με όμοια λογική δημιουργούνται και τα υπόλοιπα γραφήματα.

Σε όλα τα γραφήματα μπορεί κανείς να παρατηρήσει ότι οι καμπύλες βρίσκονται πολύ κοντά μεταξύ τους και συμπεριφέρονται με παρόμοιο τρόπο. Αυτό δημιουργεί σε κάθε γράφημα μία στενή ζώνη μέσα στην οποία περιέχονται όλες οι καμπύλες. Με βάση αυτή την παρατήρηση,

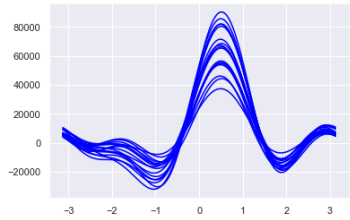
μπορείς κανείς να συμπεράνει ότι ο αριθμός 12 είναι πολλαπλάσιο της πραγματικής περιόδου της χρονοσειράς των πωλήσεων.

Η αναζήτηση της χρονοσειράς θα συνεχισθεί ερευνώντας αν κάποιο από τα υποπολλαπλάσια του 12 δημιουργεί σε όλα τα παραγόμενα γραφήματα στενές ζώνες στις οποίες να περιέχονται όλες οι καμπύλες Andrews.

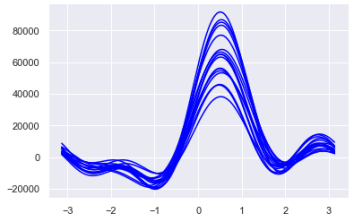
Αρχικά θα ελεγχθεί ως περίοδος ο αριθμός 6. Για τον έλεγχο αυτό θα δημιουργηθούν τα γραφήματα 5.34, 5.35, 5.36, 5.37, 5.38 και 5.39.



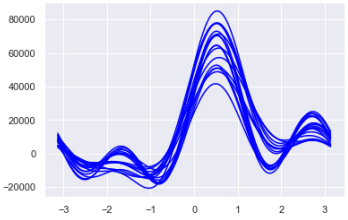
Γράφημα 5.34
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο και τον Ιούλιο



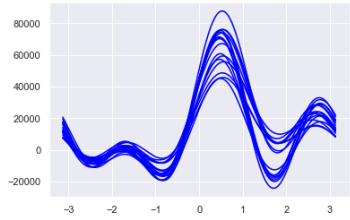
Γράφημα 5.35
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο και τον Αύγουστο



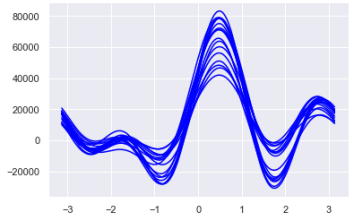
Γράφημα 5.36
Καμπύλες Andrews με αρχικό μήνα τον Μάρτιο και τον Σεπτέμβριο



Γράφημα 5.37
Καμπύλες Andrews με αρχικό μήνα τον Απρίλιο και τον Οκτώβριο



Γράφημα 5.38
Καμπύλες Andrews με αρχικό μήνα τον Μάιο και τον Νοέμβριο



Γράφημα 5.39
Καμπύλες Andrews με αρχικό μήνα τον Ιούνιο και τον Δεκέμβριο

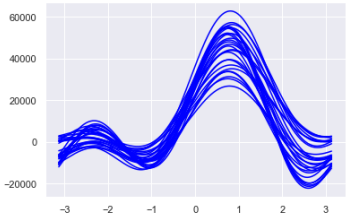
Στο Γράφημα 5.34 φαίνονται 18 καμπύλες για τη χάραξη των οποίων έχουν χρησιμοποιηθεί εξάδες μηνών με πρώτο μήνα τον Ιανουάριο και τον Ιούλιο. Αντίστοιχα στο Γράφημα 5.35 έχουν χρησιμοποιηθεί ως πρώτοι μήνες οι Φεβρουάριος και Αύγουστος. Με την ίδια λογική έχουν χαραχθεί και τα υπόλοιπα γραφήματα.

Σε όλα τα παραπάνω γραφήματα παρατηρεί κανείς ότι δεν εμφανίζεται μία αρκετά στενή ζώνη μέσα στην οποία να βρίσκονται όλες οι καμπύλες. Ειδικά στο Γράφημα 5.38 παρατηρείται η ύπαρξη τιμών του t για τις οποίες εμφανίζονται δύο ομάδες καμπυλών. Αυτό το γεγονός οδηγεί στο συμπέρασμα ότι ο αριθμός 6 δεν αποτελεί την πραγματική περίοδο της χρονοσειράς.

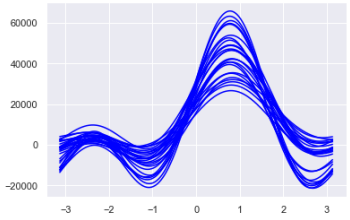
Χρησιμοποιώντας παρόμοια λογική με παραπάνω παράγονται τα Γραφήματα 5.40, 5.41, 5.42 και 5.43 όπου εξετάζεται ως περίοδος ο αριθμός 4, τα Γραφήματα 5.44, 5.45 και 5.46 όπου εξετάζεται ως περίοδος ο αριθμός 3 και τα Γραφήματα 5.47 και 5.48 όπου εξετάζεται ως περίοδος ο αριθμός 2. Και στις τρεις περιπτώσεις περιόδου φαίνεται η ίδια εικόνα μη ύπαρξης

στενής ζώνης μέσα στην οποία να εμπεριέχονται οι καμπύλες που χαράσσονται κάθε φορά. Επομένως, κανένας από τους αριθμούς 4, 3 και 2 δεν φαίνεται να είναι η πραγματική περίοδος της εξεταζόμενης χρονοσειράς.

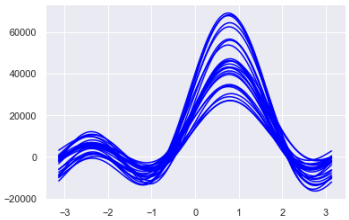
Το τελικό συμπέρασμα που προκύπτει από ολόκληρη την παραπάνω διαδικασία είναι ότι η περίοδος της χρονοσειράς φαίνεται να είναι ο αριθμός 12, που σημαίνει ότι η χρονοσειρά επαναλαμβάνει την «συμπεριφορά» της με ετήσια περίοδο.



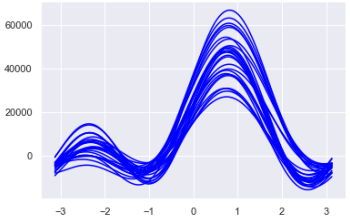
Γράφημα 5.40
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο, τον Μάιο και τον Σεπτέμβριο



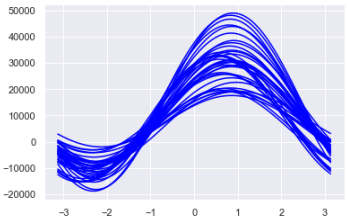
Γράφημα 5.41
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο, τον Ιούνιο και τον Οκτώβριο



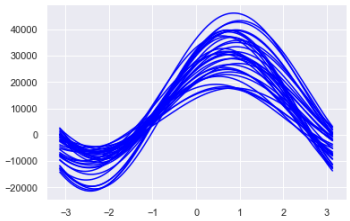
Γράφημα 5.42
Καμπύλες Andrews με αρχικό μήνα τον Μάρτιο, τον Ιούλιο και τον Νοέμβριο



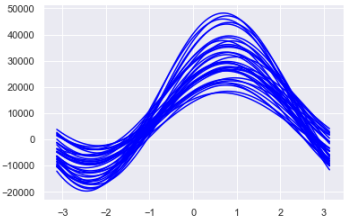
Γράφημα 5.43
Καμπύλες Andrews με αρχικό μήνα τον Απρίλιο, τον Αύγουστο και τον Δεκέμβριο



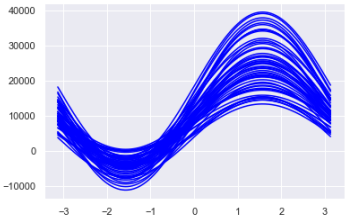
Γράφημα 5.44
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο, τον Απρίλιο, τον Ιούλιο και τον Οκτώβριο



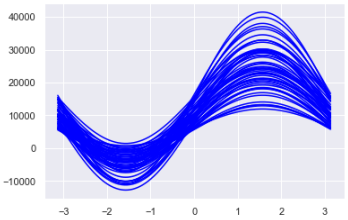
Γράφημα 5.45
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο, τον Μάιο, τον Αύγουστο και τον Νοέμβριο



Γράφημα 5.46
Καμπύλες Andrews με αρχικό μήνα τον Μάρτιο, τον Ιούνιο, τον Σεπτέμβριο και τον Δεκέμβριο



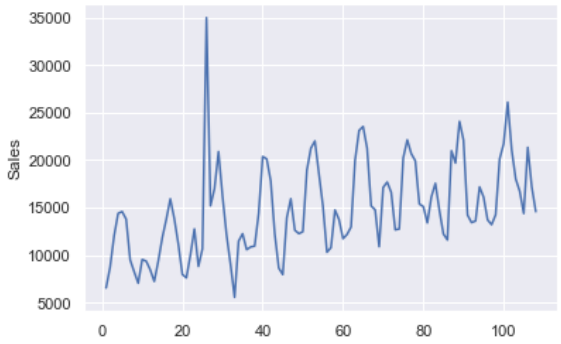
Γράφημα 5.47
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο, τον Μάρτιο, τον Μάιο, τον Ιούλιο, τον Σεπτέμβριο και τον Νοέμβριο



Γράφημα 5.48
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο, τον Απρίλιο, τον Ιούνιο, τον Αύγουστο, τον Οκτώβριο και τον Δεκέμβριο

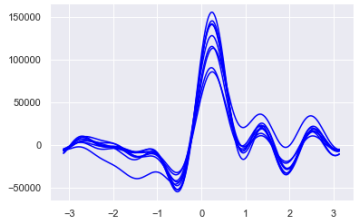
5.4.3 Εντοπισμός ακραίων τιμών

Όπως μπορεί κανείς να παρατηρήσει στο Γράφημα 5.21, στη χρονοσειρά δεν περιέχονται ακραίες τιμές. Για το λόγο αυτό, θα εισαχθεί μία νέα τιμή, ακραία μεγαλύτερη των υπολοίπων της χρονοσειράς, στη θέση μίας εκ των υπαρχόντων τιμών της. Η τιμή που θα αλλαχθεί είναι αυτή που αντιστοιχεί στο μήνα Φεβρουάριο του έτους 1962 και η νέα τιμή της τίθεται ίση με 35000, αντί 10947. Η νέα εικόνα της χρονοσειράς δίνεται στο Γράφημα 5.49.

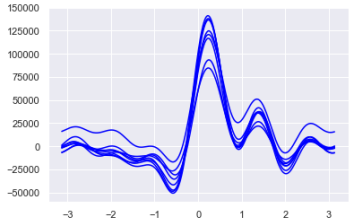


Γράφημα 5.49
Η καμπύλη της χρονοσειράς Monthly Car Sales με τον αντικατάσταση της τιμής του Φεβρουαρίου το 1962 με 35000

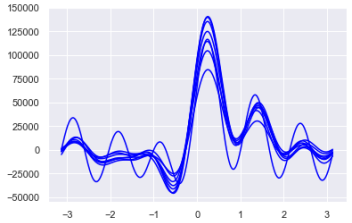
Στόχος σε αυτή την υποπαράγραφο είναι να χρησιμοποιηθούν οι καμπύλες Andrews για τον εντοπισμό της ακραίας τιμής που εισήχθη στη χρονοσειρά. Για το σκοπό αυτό θα χαραχθούν εκ νέου όλα τα επιμέρους γραφήματα των καμπυλών Andrews όπως περιγράφεται στην Παράγραφο 4.4.



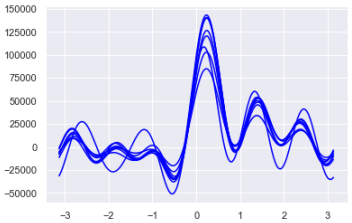
Γράφημα 5.50
Καμπύλες Andrews με αρχικό μήνα τον Ιανουάριο



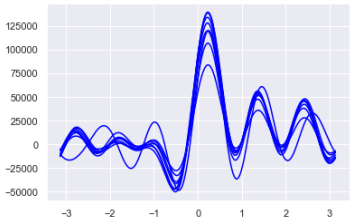
Γράφημα 5.51
Καμπύλες Andrews με αρχικό μήνα τον Φεβρουάριο



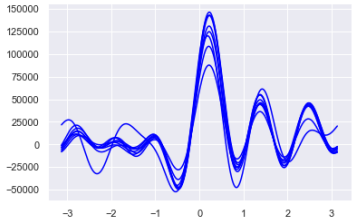
Γράφημα 5.52
Καμπύλες Andrews με αρχικό μήνα τον Μάρτιο



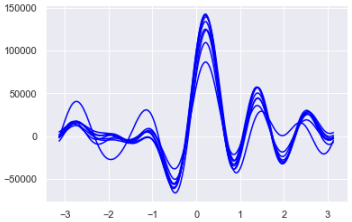
Γράφημα 5.53
Καμπύλες Andrews με αρχικό μήνα τον Απρίλιο



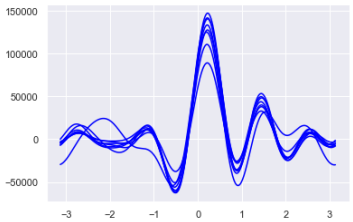
Γράφημα 5.54
Καμπύλες Andrews με αρχικό μήνα τον Μάιο



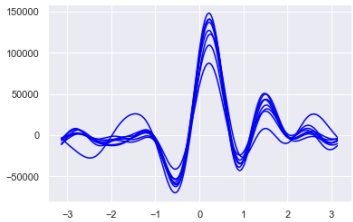
Γράφημα 5.55
Καμπύλες Andrews με αρχικό μήνα τον Ιούνιο



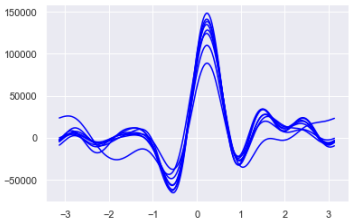
Γράφημα 5.56
Καμπύλες Andrews με αρχικό μήνα τον Ιούλιο



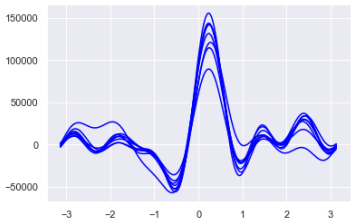
Γράφημα 5.57
Καμπύλες Andrews με αρχικό μήνα τον Αύγουστο



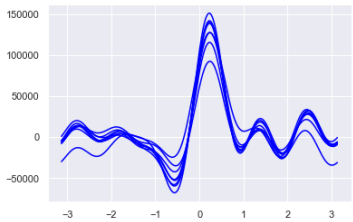
Γράφημα 5.58
Καμπύλες Andrews με αρχικό μήνα τον Σεπτέμβριο



Γράφημα 5.59
Καμπύλες Andrews με αρχικό μήνα τον Οκτώβριο



Γράφημα 5.60
Καμπύλες Andrews με αρχικό μήνα τον Νοέμβριο



Γράφημα 5.61
Καμπύλες Andrews με αρχικό μήνα τον Δεκέμβριο

Από τη μορφή της συνάρτησης του Andrews γίνεται φανερό πως ο τελευταίος όρος της έχει τη μικρότερη περίοδο σε σχέση με όλους τους υπόλοιπους. Αυτό είναι και το στοιχείο που χρησιμοποιείται για τον εντοπισμό ακραίων τιμών στη συγκεκριμένη περίπτωση.

Εύκολα παρατηρεί κανείς πως σε κανένα από τα Γραφήματα δεν υπάρχει μία στενή ζώνη μέσα στην οποία να βρίσκονται όλες οι καμπύλες. Αυτό υποδεικνύει την ύπαρξη κάποιας ακραίας τιμής. Αρχικά παρατηρείται πως στο Γράφημα 5.52 υπάρχει μία καμπύλη που φαίνεται να ταλαντώνεται ταχύτερα από τις υπόλοιπες. Αυτό σημαίνει πως η μεταβλητή που έχει μπει

στην τελευταία θέση της συνάρτησης παίρνει ακραία τιμή σε σχέση με τις υπόλοιπες. Επομένως, η ακραία τιμή βρίσκεται στον μήνα Φεβρουάριο. Αυτό επιβεβαιώνεται και από το Γράφημα 5.51, όπου αυτή η ταχύτερη ταλάντωση δεν υφίσταται αφού ο μήνας Φεβρουάριος είναι τώρα στην πρώτη θέση της συνάρτησης και πολλαπλασιάζεται με σταθερό όρο.

Στη συνέχεια, ελέγχοντας όλες τις μετρήσεις που αφορούν το μήνα Φεβρουάριο, φαίνεται πως η ακραία τιμή εμφανίζεται τον Φεβρουάριο του έτους 1962, ακριβώς όπως αναμενόταν αφού εισήχθη για την απόδειξη της ιδιότητας αυτής

Με χρήση της παραπάνω μεθόδου μπορεί κανείς να αναζητήσει τυχόν ακραίες τιμές σε δεδομένα χρονοσειρών και στη συνέχεια να προβεί σε όλες τις ενέργειες που απαιτούνται για την αντιμετώπισή τους ανάλογα και με το πεδίο έρευνας.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα 1

Σύνολο δεδομένων Iris

Το σύνολο δεδομένων Iris είναι ένα σύνολο πολυμεταβλητών παρατηρήσεων που έχει παραχθεί από τον Βρετανό Στατιστικό Ronald Fisher. Έχει χρησιμοποιηθεί για μία πληθώρα εφαρμογών στον τομέα της Στατιστικής και είναι από τα γνωστότερα σύνολα δεδομένων. Το συγκεκριμένο σύνολο δεδομένων μπορεί κανείς να το βρει στην ηλεκτρονική διεύθυνση

https://archive.ics.uci.edu/ml/datasets/iris

καθώς και σε πολλές ακόμα στο διαδίκτυο.

Αποτελείται από 150 παρατηρήσεις που διαχωρίζονται ισόποσα σε τρία διαφορετικά είδη του λουλουδιού ίριδα. Για κάθε μία από τις παρατηρήσεις δίνονται οι μετρήσεις τεσσάρων μεταβλητών που αφορούν το μήκος και το πλάτος τόσο του σέπαλου όσο και του πετάλου.

Για τους σκοπούς αυτής της εργασία θα χρησιμοποιηθεί ένα μέρος του συγκεκριμένου συνόλου δεδομένων. Πιο συγκεκριμένα, θα χρησιμοποιηθούν μόνο οι παρατηρήσεις που αφορούν τους τύπους λουλουδιών Iris-setosa και Iris-virginica. Για κάθε μία από τις παρατηρήσεις θα χρησιμοποιηθούν όλες οι διαθέσιμες μεταβλητές. Στον παρακάτω πίνακα δίνονται οι πρώτες πέντε και οι τελευταίες πέντε παρατηρήσεις κάθε μίας από τις δύο ομάδες.

sepal length	sepal width	petal length	petal width	class
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
4,6	3,1	1,5	0,2	Iris-setosa
5,0	3,6	1,4	0,2	Iris-setosa
:	:	:	:	:
4,8	3,0	1,4	0,3	Iris-setosa
5,1	3,8	1,6	0,2	Iris-setosa
4,6	3,2	1,4	0,2	Iris-setosa
5,3	3,7	1,5	0,2	Iris-setosa
5,0	3,3	1,4	0,2	Iris-setosa
6,3	3,3	6,0	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica
7,1	3,0	5,9	2,1	Iris-virginica
6,3	2,9	5,6	1,8	Iris-virginica
6,5	3,0	5,8	2,2	Iris-virginica
:	:	:	:	:
6,7	3,0	5,2	2,3	Iris-virginica
6,3	2,5	5,0	1,9	Iris-virginica
6,5	3,0	5,2	2,0	Iris-virginica
6,2	3,4	5,4	2,3	Iris-virginica
5,9	3,0	5,1	1,8	Iris-virginica

Παράρτημα 2

Σύνολο δεδομένων California Housing Prices

Το σύνολο δεδομένων California Housing Prices είναι ένα σύνολο πολυδιάστατων δεδομένων. Περιέχει δεδομένα που αφορούν τα σπίτια μίας συγκεκριμένης περιοχής στην Καλιφόρνια και διάφορα στατιστικά στοιχεία από την απογραφή του έτους 1990 και μπορεί κανείς να το βρει στην ηλεκτρονική διεύθυνση

https://www.kaggle.com/camnugent/california-housing-prices

καθώς και σε πολλές ακόμα στο διαδίκτυο.

Στο σύνολο περιέχονται 20640 παρατηρήσεις, ενώ για κάθε παρατήρηση έχει γίνει συλλογή 10 μεταβλητών. Οι δέκα μεταβλητές που βρίσκονται στο σύνολο αφορούν τις συντεταγμένες των σπιτιών, τη διάμεση ηλικία των σπιτιών του οικοδομικού τετραγώνου, το συνολικό αριθμό δωματίων στο οικοδομικό τετράγωνο, το συνολικό αριθμό υπνοδωματίων στο οικοδομικό τετράγωνο, το συνολικό αριθμό κατοίκων στο οικοδομικό τετράγωνο, τον αριθμό των νοικοκυριών στο οικοδομικό τετράγωνο, το διάμεσο εισόδημα των νοικοκυριών του οικοδομικού τετραγώνου και τέλος, μια κατηγορική μεταβλητή σχετικά με την απόσταση του σπιτιού από τη θάλασσα.

Για τους σκοπούς αυτής της εργασίας έχει γίνει χρήση ολόκληρου του συνόλου δεδομένων. Στους παρακάτω πίνακες δίνονται οι τρεις πρώτες και οι τρεις τελευταίες παρατηρήσεις του συνόλου California Housing Prices.

id	longitude	latitude	housing_median_age	total_rooms	total_bedrooms
1	-122.23	37.88	41.0	880.0	129.0
2	-122.22	37.86	21.0	7099.0	1106.0
3	-122.24	37.85	52.0	1467.0	190.0
:	:	:	:	:	:
20638	-121.22	39.43	17.0	2254.0	485.0
20639	-121.32	39.43	18.0	1860.0	409.0
20640	-121.24	39.37	16.0	2785.0	616.0

id	population	households	median_income	median_house_value	ocean_proximity
1	322.0	126.0	8.3252	452600.0	NEAR BAY
2	2401.0	1138.0	8.3014	358500.0	NEAR BAY
3	496.0	177.0	7.2574	352100.0	NEAR BAY
:	:	:	:	:	:
20638	1007.0	433.0	1.7000	92300.0	INLAND
20639	741.0	349.0	1.8672	84700.0	INLAND
20640	1387.0	530.0	2.3886	89400.0	INLAND

Παράρτημα 3

Σύνολο δεδομένων Cars

Το σύνολο δεδομένων Cars προέρχεται από δεδομένα του περιοδικού "Motor Trend Magazine" του έτους 1974 και περιέχει 32 μοντέλα αυτοκινήτων διάφορων εταιρειών και διάφορων χωρών προέλευσης και μπορεί να βρεθεί στην ηλεκτρονική διεύθυνση

https://dasl.datadescription.com/datafile/cars/.

Για κάθε όχημα έχουν συλλεγεί 11 μεταβλητές, μερικές από τις οποίες είναι η κατανάλωση, ο αριθμός των κυλίνδρων της μηχανής, το αν έχει αυτόματο ή μηχανικό κιβώτιο ταχυτήτων.

Σε αυτή την εργασία θα γίνει χρήση μόνο 25 από τα διαθέσιμα αυτοκίνητα και αντίστοιχα μόνο 6 από τις διαθέσιμες μεταβλητές όπως φαίνεται και στον πίνακα παρακάτω. Οι μεταβλητές που θα χρησιμοποιηθούν είναι οι εξής:

- MPG: απόσταση, σε μίλια, που μπορεί να διανύσει το όχημα με ένα λίτρο καυσίμου,
- Weight: βάρος του αυτοκινήτου,
- Drive Ratio: σχέση μετάδοσης της κίνησης από τον κινητήρα στους τροχούς,
- Horsepower: αριθμός ίππων του κινητήρα,
- Displacement: χωρητικότητα των κυλίνδρων του κινητήρα σε κυβικές ίντσες,
- Cylinders: αριθμός κυλίνδρων του κινητήρα.

Μοντέλο	MPG	Weight	Drive Ratio	Horsepower	Displacement	Cylinders
Chevette	30.0	2.155	3.70	68	98	4
Dodge Omni	30.9	2.230	3.37	75	105	4
AMC Spirit	27.4	2.670	3.08	80	121	4
Plymouth Horizon	34.2	2.200	3.37	70	105	4
Mazda GLC	34.1	1.975	3.73	65	86	4
Dodge Colt	35.1	1.915	2.97	80	98	4
Honda Accord LX	29.5	2.135	3.05	68	98	4
Datsun 210	31.8	2.020	3.70	65	85	4
VW Scirocco	31.5	1.990	3.78	71	89	4
VW Dasher	30.5	2.190	3.70	78	97	4
VW Rabbit	31.9	1.925	3.78	71	89	4
Fiat Strada	37.3	2.130	3.10	69	91	4
Mercury Zephyr	20.8	3.070	3.08	85	200	6
Ford Mustang Ghia	21.9	2.910	3.08	109	171	6
Chevy Citation	28.8	2.595	2.69	115	173	6
Olds Omega	26.8	2.700	2.84	115	173	6
Volvo 240 GL	17.0	3.140	3.50	125	163	6
Peugeot 694 SL	16.2	3.410	3.58	133	163	6

Μοντέλο	MPG	Weight	Drive Ratio	Horsepower	Displacement	Cylinders
Buick Estate Wagon	16.9	4.360	2.73	155	350	8
Ford Country Squire Wagon	15.5	4.054	2.26	142	351	8
Chrysler LeBaron Wagon	18.5	3.940	2.45	150	360	8
Chevy Caprice Classic	17.0	3.840	2.41	130	305	8
Ford LTD	17.6	3.725	2.26	129	302	8
Mercury Grand Marquis	16.5	3.955	2.26	138	351	8
Dodge St Regis	18.2	3.830	2.45	135	318	8

Παράρτημα 4

Σύνολο δεδομένων εταιρίας Nimber Tech

Η εταιρία Nimber Tech δραστηριοποιείται στο χώρο των μεταφορών με κυριότερες αγορές αυτές της Νορβηγίας και της Μεγάλης Βρετανίας. Η Nimber Tech διαθέτει στους πελάτες της μία διαδικτυακή πλατφόρμα μέσω της οποίας συμπληρώνουν τα στοιχεία της μεταφοράς που θέλουν να πραγματοποιήσουν. Στη συνέχεια η εταιρία αναλαμβάνει να βρει τον μεταφορέα που θα την πραγματοποιήσει και φέρνει τα δύο μέρη σε επαφή ώστε να πραγματοποιηθεί η μεταφορά.

Τα στοιχεία που συμπληρώνουν οι χρήστες της πλατφόρμας αποτελούνται από τον τίτλο του αντικειμένου που θα ήθελαν να μεταφερθεί, το μέγεθος, το βάρος, τη διεύθυνση του αποστολέα, τη διεύθυνση του παραλήπτη καθώς και το χρηματικό ποσό που θα ήταν διατεθειμένος ο αποστολέας να διαθέσει για τη μεταφορά. Επομένως, στη βάση δεδομένων της εταιρίας υπάρχουν όλα τα παραπάνω δεδομένα.

Το σύνολο δεδομένων που θα χρησιμοποιηθεί στη συγκεκριμένη εργασία είναι ένα μικρό υποσύνολο της βάσης δεδομένων της εταιρίας, η οποία για λόγους εμπιστευτικότητας δεν είναι διαθέσιμη ελεύθερα. Το σύνολο αποτελείται από 24 παρατηρήσεις που αφορούν μεταφορές που ζητήθηκαν από χρήστες της πλατφόρμας και τέσσερις μεταβλητές.

Size	Weight	Price	Distance	State
3	2	209	25.542	Successful
3	37	200	36.805	Successful
3	22	150	36.805	Successful
3	6	190	36.805	Successful
5	43	233	18.359	Successful
5	23	250	38.013	Successful
3	4	192	20.339	Successful
3	1	130	18.981	Successful
3	29	200	36.806	Successful
5	15	205	28.123	Successful
3	3	189	39.823	Successful
3	5	192	21.758	Successful
3	1	250	412.612	Unsuccessful
3	6	194	515.192	Unsuccessful
3	75	390	453.561	Unsuccessful
3	4	150	407.514	Unsuccessful
2	10	135	508.787	Unsuccessful
5	10	194	493.720	Unsuccessful
3	15	330	437.851	Unsuccessful
2	10	530	580.062	Unsuccessful
3	10	330	447.327	Unsuccessful
3	5	400	532.426	Unsuccessful
5	50	123	475.001	Unsuccessful
3	14	300	506.446	Unsuccessful

Στον παρακάτω πίνακα δίνονται μερικά από τα αριθμητικά περιγραφικά μέτρα για τις αριθμητικές μεταβλητές του συνόλου δεδομένων.

State	Μέσος	Τυπική Απόκλιση	1ο τεταρτημόριο	2ο τεταρτημόριο	3ο τεταρτημόριο
Distance (Km)	255	233	35	224	480
Price (Nok)	236	99	190	200	263
Weight (Kg)	17	18	5	10	22

Στον παρακάτω πίνακα δίνονται μερικά από τα αριθμητικά περιγραφικά μέτρα για τις αριθμητικές μεταβλητές της ομάδας των επιτυχημένων μεταφορών του συνόλου.

State	Μέσος	Τυπική Απόκλιση	1ο τεταρτημόριο	2ο τεταρτημόριο	3ο τεταρτημόριο
Distance (Km)	30	8	21	32	37
Price (Nok)	195	32	190	196	206
Weight (Kg)	16	15	4	11	25

Στον παρακάτω πίνακα δίνονται μερικά από τα αριθμητικά περιγραφικά μέτρα για τις αριθμητικές μεταβλητές της ομάδας των μη επιτυχημένων μεταφορών του συνόλου.

State	Μέσος	Τυπική Απόκλιση	1ο τεταρτημόριο	2ο τεταρτημόριο	3ο τεταρτημόριο
Distance (Km)	481	51	445	484	510
Price (Nok)	277	125	183	275	345
Weight (Kg)	18	22	6	10	14

Παράρτημα 5

Σύνολο δεδομένων Monthly Car Sales

Το σύνολο δεδομένων Monthly Car Sales αφορά μηνιαίες πωλήσεις αυτοκινήτων στην επαρχία Quebec του Καναδά. Η περίοδος που καλύπτεται από τη χρονοσειρά ξεκινάει τον Ιανουάριο του 1960 και ολοκληρώνεται τον Δεκέμβριο του 1968.

Το συγκεκριμένο σύνολο δεδομένων μπορεί να βρεθεί στην ηλεκτρονική διεύθυνση:

https://www.kaggle.com/dinirimameev/monthly-car-sales-in-quebec-1960.

Date	Sales
01/1960	6550
02/1960	8728
03/1960	12026
04/1960	14395
05/1960	14587
06/1960	13791
07/1960	9498
08/1960	8251
09/1960	7049
10/1960	9545
11/1960	9364
12/1960	8456
01/1961	7237
02/1961	9374
03/1961	11837
04/1961	13784
05/1961	15926
06/1961	13821
07/1961	11143
08/1961	7975
09/1961	7610
10/1961	10015
11/1961	12759
12/1961	8816
01/1962	10677
02/1962	10947
03/1962	15200
04/1962	17010
05/1962	20900
06/1962	16205
07/1962	12143
08/1962	8997
09/1962	5568
10/1962	11474
11/1962	12256
12/1962	10583

Date	Sales
01/1963	10862
02/1963	10965
03/1963	14405
04/1963	20379
05/1963	20128
06/1963	17816
07/1963	12268
08/1963	8642
09/1963	7962
10/1963	13932
11/1963	15936
12/1963	12628
01/1964	12267
02/1964	12470
03/1964	18944
04/1964	21259
05/1964	22015
06/1964	18581
07/1964	15175
08/1964	10306
09/1964	10792
10/1964	14752
11/1964	13754
12/1964	11738
01/1965	12181
02/1965	12965
03/1965	19990
04/1965	23125
05/1965	23541
06/1965	21247
07/1965	15189
08/1965	14767
09/1965	10895
10/1965	17130
11/1965	17697
12/1965	16611

Date	Sales
01/1966	12674
02/1966	12760
03/1966	20249
04/1966	22135
05/1966	20677
06/1966	19933
07/1966	15388
08/1966	15113
09/1966	13401
10/1966	16135
11/1966	17562
12/1966	14720
01/1967	12225
02/1967	11608
03/1967	20985
04/1967	19692
05/1967	24081
06/1967	22114
07/1967	14220
08/1967	13434
09/1967	13598
10/1967	17187
11/1967	16119
12/1967	13713
01/1968	13210
02/1968	14251
03/1968	20139
04/1968	21725
05/1968	26099
06/1968	21084
07/1968	18024
08/1968	16722
09/1968	14385
10/1968	21342
11/1968	17180
12/1968	14577

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Andrews, D. (1972). Plots of high-dimensional data, *Biometrics*, **28**, 125-136.

[2] Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data, *SIAM Journal on Scientific and Statistical Computing*, **6**, 128-143.

[3] Embrechts, P. and Herzberg, A. M. (1991). Variations of Andrews' plots, *International Statistical Review*, **59**, 175-194.

[4] Embrechts, P., Herzberg, A. M. and Allen, C. K. (1986). An investigation of Andrews' plots to detect period and outliers in time series data, *Communications in Statistics - Simulation and Computation*, **15**, 1027-1051.

[5] Embrechts, P., Herzberg, A. M., Kalbfleisch, H. K., Traves, W. N. and Whitla, J. R. (1995). An introduction to wavelets with applications to Andrews' plots, *Journal of Computational and Applied Mathematics*, **64**, 41-56.

[6] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd Edition, Wiley, New York

[7] Khattree, R. and Naik, D. (2002). Andrews' plots for multivariate data: Some new suggestions and applications, *Journal of Statistical Planning and Inference*, **100**, 411-425

[8] Koziol, J. A. and Hacke, W. (1991). A bivariate version of Andrews' plots, *IEEE Transactions on Biomedical Engineering*, **38**, 1271-1274

[9] Moustafa, R. (2011). Andrews curves, *WIREs Computational Statistics*, **3**, 373-382

[10] Seber, G. A. F. (1984). *Multivariate Observations*, Wiley, New York

[11] Wegman, E. J. and Shen, J. (1993). Three-dimensional Andrews plots and the grand tour, *Computing Science and Statistics*, **25**, 284-288