Πανεπιστημιο Κρητης - Τμημα Εφαρμοσμενων Μαθηματικων
University of Crete - Department of Applied Mathematics

Ανάλυση Κυρίων Συνιστωσών

Επιβλέπων: Μιχάλης Πλεξουσάκης

Γεωργία Σφακιανάκη
AM: 1098
gsfakian@tem.uoc.gr

Τριμελής επιτροπή:

Θεόδωρος Κατσαούνης

Μιχάλης Πλεξουσάκης

Βαγγέλης Χαρμανδάρης

Ακαδημαϊκό Έτος 2013 - 2014

Περιεχόμενα

1 Εισαγωγή

Στην παρούσα εργασία θα μελετήσουμε την Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis, PCA). Το αντικείμενο της PCA είναι (α) η μείωση των διαστάσεων των δεδομένων και (β) η ανάλυσή (ερμηνεία) τους. Ειδικότερα, σε πολυδιάστατα δεδομένα όπου η γραφική τους αναπαράσταση δεν είναι εφικτή, η ανάλυση κυρίων συνιστωσών είναι ένα πολύ ισχυρό εργαλείο. Οι εφαρμογές της μεθόδου είναι ποικίλες, από τη στατιστική ανάλυση δεδομένων, τη μετεωρολογία και τις γεωφυσικές επιστήμες μέχρι την συμπίεση εικόνας, όπως θα δούμε στα παραδείγματά μας.

Η Ανάλυση Κυρίων Συνιστωσών είναι από τις παλαιότερες και πιθανώς, από τις πιο διαδεδομένες τεχνικές της πολυμεταβλητής ανάλυσης δεδομένων. Εισήχθη για πρώτη φορά από τον Karl Pearson το 1901, κι έπειτα αναπτύχθηκε από πολλούς ακόμη. Παρότι είναι συχνή η χρήση του όρου "Ανάλυση Κυρίων Συνιστωσών", πολλές φορές συναντάται με διαφορετική ορολογία, ανάλογα με τον τομέα στον οποίο εφαρμόζεται. Η μέθοδος αυτή έγινε ευρέως γνωστή στις ατμοσφαιρικές επιστήμες, όταν παρουσιάστηκε σε μία δημοσίεση του Lorenz (1956) ο οποίος χρησιμοποίησε τον όρο "Empirical Orthogonal Function" (EOF). Ακόμη και σήμερα, τα δύο αυτά ονόματα παραπέμπουν στην ίδια τεχνική και μάλιστα, τα περισσότερα εγχειρίδια στατιστικής ανάλυσης περιλαμβάνουν κεφάλαια που αφορούν την Ανάλυση Κυρίων Συνιστωσών.

Η ανάλυση πολυδιάστατων δεδομένων καθίσταται ιδιαίτερα δύσκολη όταν το πλήθος των μεταβλητών n είναι μεγάλο. Επίσης, υπάρχει δυσκολία στην ανάλυση όταν οι μεταβλητές είναι υψηλά συσχετισμένες μεταξύ τους. Παρόλο που απαιτούνται n μεταβλητές για να ερμηνευτεί η συνολική μεταβλητότητα του δείγματος, συχνά το μεγαλύτερο ποσοστό της μεταβλητότητας αυτής, μπορεί να ερμηνευτεί από έναν (αρκετά) μικρότερο αριθμό k συνιστωσών. Αν πράγματι συμβεί αυτό, τότε υπάρχει (σχεδόν) τόση πληροφορία στις k συνιστώσες, όση υπάρχει και στις n αρχικές μεταβλητές. Οι k αυτές συνιστώσες, ονομάζονται κύριες συνιστώσες και μπορούν να αντικαταστήσουν τις αρχικές n μεταβλητές, απλοποιώντας κατά πολύ τις διαστάσεις. Οι κύριες συνιστώσες είναι γραμμικός συνδυασμός των n αρχικών μεταβλητών, και μάλιστα είναι ασυσχέτιστες μεταξύ τους. Έτσι, οδηγούμαστε από ένα σύνολο n συσχετισμένων μεταβλητών, σ' ένα μικρότερο σύνολο k ασυσχέτιστων μεταβλητών. Σε ορισμένες περιπτώσεις που το k, η νέα διάσταση, είναι 2 ή 3 τότε μπορούμε να έχουμε μια οπτική ιδέα, μια εικόνα των δεδομένων.

Στην ενότητα 2 εισάγουμε τους στατιστικούς όρους που θα χρησιμοποιήσουμε, καθώς και κάποιους βασικούς ορισμούς από την γραμμική άλγεβρα πινάκων. Έπειτα, στην ενότητα 3 παρουσιάζουμε τη μέθοδο αναλυτικά και παραθέτουμε παραδείγματα για την κατανόησή της. Στην ενότητα 4 εξετάζουμε πώς μπορούμε να εφαρμόσουμε την ανάλυση κυρίων συνιστωσών σε δεδομένα που εξαρτώνται από το χρόνο και παρουσιάζουμε ένα τελευταίο παράδειγμα. Τέλος, έχουμε την ενότητα 5 στην οποία παρουσιάζουμε τα δεδομένα που χρησιμοποιούμε καθώς και τους κώδικες των προγραμμάτων που υλοποιούν τα παραδείγματα, οι οποίοι είναι γραμμένοι σε Matlab.

2 Μαθηματικό υπόβαθρο

Η Ανάλυση Κυρίων Συνιστωσών αποτελεί μία σημαντική μέθοδο ανάλυσης δεδομένων. Πριν προχωρήσουμε στην παρουσίαση της, πρέπει πρώτα να εισάγουμε ορισμένες πολύ βασικές έννοιες από τη Στατιστική και από την Άλγεβρα Πινάκων.

2.1 Εμπειρική Στατιστική

Έστω ότι έχουμε ένα σύνολο δεδομένων $x = [x_1, x_2, \ldots, x_n]$. Αυτό που μας ενδιαφέρει είναι να υπολογίσουμε κάποια στατιστικά μέτρα, από τα οποία θα μπορέσουμε να βγάλουμε συμπέρασμα για τις σχέσεις μεταξύ των επιμέρους σημείων του δείγματος.

- Η μέση τιμή (mean) του δείγματος ορίζεται ως:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Η διακύμανση (variance) υποδηλώνει πόσο συγκεντρωμένες γύρω από τη μέση τιμή είναι οι τιμές του δείγματος και ορίζεται ως:

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Η τυπική απόκλιση (standard deviation) είναι η θετική τετραγωνική ρίζα της διακύμανσης και ορίζεται ως:

$$\sigma_x = \sqrt{Var(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Εάν έχουμε δύο μεταβλητές, έστω $x = [x_1, x_2, \ldots, x_n]$ και $y = [y_1, y_2, \ldots, y_n]$ τότε μπορούμε να ορίσουμε τη συνδιακύμανση τους (covariance) ως:

$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

  Προφανώς ισχύει $Cov(x, y) = Cov(y, x)$.
  Ενδιαφέρον παρουσιάζει η συνδιακύμανση όταν έχουμε πάνω από δύο διαστάστεις. Μπορούμε να υπολογίσουμε τις τιμές της συνδιακύμανσης μεταξύ όλων των διαφορετικών διαστάσεων που έχουμε και να τις βάλουμε σε έναν πίνακα.
  Έστω ότι έχουμε τρεις μεταβλητές x, y και z. Ο πίνακας συνδιακύμανσης (covariance

matrix), είναι συμμετρικός και ορίζεται ως:

$$C = \begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}.$$

Παρατηρούμε ότι στην κύρια διαγώνιο του πίνακα βρίσκονται οι τιμές των διακυμάνσεων των x, y, z. Δηλαδή

$$Cov(x,x) = Var(x), Cov(y,y) = Var(y), Cov(z,z) = Var(z).$$

Στα επόμενα κεφάλαια της εργασίας θα έχουμε πολυδιάστατα δεδομένα, έστω $X \in \mathbb{R}^{m\times n}$. Τα στοιχεία του πίνακα συνδιακύμανσης θα δίνονται από τον τύπο

$$s_{i,j} = \frac{1}{m-1}\sum_{k=1}^{m}\left(X_{k,i} - \overline{X}_i\right)\left(X_{k,j} - \overline{X}_j\right).$$

Ο πίνακας θα έχει τη μορφή

$$C_X = \begin{bmatrix} s_{1,1} & s_{1,2} & s_{1,3} & \cdots & s_{1,j} \\ s_{2,1} & s_{2,2} & s_{2,3} & \cdots & s_{2,j} \\ s_{3,1} & s_{3,2} & s_{3,3} & \cdots & s_{3,j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{i,1} & s_{i,2} & s_{i,3} & \cdots & s_{i,j} \end{bmatrix}, \quad \text{με } C_X \in \mathbb{R}^{m\times m}.$$

Στη συνέχεια, θα αναφερόμαστε στον πίνακα συνδιακύμανσης ως $C_X = \frac{1}{n-1}XX^T$, όπου X είναι ο πίνακας δεδομένων από τον οποίο έχουν αφαιρεθεί οι μέσες τιμές.

- Ο πίνακας συσχέτισης (correlation matrix) είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης για κάθε ζευγάρι μεταβλητών και ορίζεται ως:

$$R = \begin{bmatrix} 1 & \rho_{x,y} & \rho_{x,z} \\ \rho_{y,x} & 1 & \rho_{y,z} \\ \rho_{z,x} & \rho_{z,y} & 1 \end{bmatrix}, \quad \text{όπου}$$

$$\rho_{u,v} = \frac{Cov(u,v)}{\sigma_u\sigma_v} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(u_i-\overline{u})(v_i-\overline{v})}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(u_i-\overline{u})^2\right]^{\frac{1}{2}}\left[\frac{1}{n-1}\sum_{i=1}^{n}(v_i-\overline{v})^2\right]^{\frac{1}{2}}}$$

$$= \frac{\sum_{i=1}^{n}(u_i-\overline{u})(v_i-\overline{v})}{\left[\sum_{i=1}^{n}(u_i-\overline{u})^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{n}(v_i-\overline{v})^2\right]^{\frac{1}{2}}}.$$

2.2 Άλγεβρα Πινάκων

2.2.1 Ιδιοτιμές και Ιδιοδιανύσματα

Έστω A ένας τετραγωνικός πίνακας που ανήκει στο διανυσματικό χώρο $\mathbb{R}^{n\times n}$. Οι **ιδιοτιμές** $\lambda_i(A), i = 1, 2, \ldots, n$ του πίνακα A, είναι οι ρίζες του χαρακτηριστικού πολυωνύμου

$$p_A(\lambda) = det(A - \lambda I).$$

Σε κάθε ιδιοτιμή λ αντιστοιχεί τουλάχιστον ένα μη-μηδενικό **ιδιοδιάνυσμα** v έτσι ώστε να ισχύει

$$Av = \lambda v.$$

Βασική παρατήρηση είναι ότι τα ιδιοδιανύσματα που αντιστοιχούν σε διαφορετικές ιδιοτιμές είναι γραμμικά ανεξάρτητα μεταξύ τους. Αν επιπλέον, ο A είναι συμμετρικός, τότε έχει ακριβώς n ιδιοτιμές, όχι κατ'ανάγκη διαφορετικές μεταξύ τους. Δύο ιδιοδιανύσματα που αντιστοιχούν σε διαφορετικές ιδιοτιμές του πίνακα A είναι ορθοκανονικά (κάθετα) μεταξύ τους.

Πρόταση 2.2.1 *Έστω ότι ο A έχει διακριτές ιδιοτιμές. Αν επιπλέον, είναι συμμετρικός (άρα οι ιδιοτιμές του είναι πραγματικές) τότε υπάρχει πραγματικός ορθογώνιος πίνακας Q τέτοιος ώστε*

$$Q^T AQ = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

2.2.2 Ανάλυση Ιδιαζουσών Τιμών (SVD)

Θεώρημα 2.2.2 *Έστω A ένας πραγματικός πίνακας $m \times n$ διαστάσεων. Τότε, υπάρχουν δύο ορθογώνιοι πίνακες*

$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}, \quad V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$$

τέτοιοι ώστε

$$U^T\ A\ V = \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p), \quad p = \min(m, n)$$

και $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$, όπου τα σ_i είναι οι ιδιάζουσες τιμές του πίνακα A.

Απόδειξη. Έστω $x \in \mathbb{R}^n$ και $y \in \mathbb{R}^m$ τέτοια ώστε $||x||_2 = ||y||_2 = 1$ και $Ax = \sigma y$ με $\sigma = ||A||_2$. Έστω $V = [x, V_1] \in \mathbb{R}^{n\times n}$ και $U = [y, U_1] \in \mathbb{R}^{m\times m}$ ορθογώνιοι πίνακες. Τότε,

$$U^T AV = \begin{bmatrix} y^T \\ U_1^T \end{bmatrix} A \begin{bmatrix} x & V_1 \end{bmatrix} = \begin{bmatrix} y^T A \\ U_1^T A \end{bmatrix} \begin{bmatrix} x & V_1 \end{bmatrix} = \begin{bmatrix} y^T Ax & y^T AV_1 \\ U_1^T Ax & U_1^T AV_1 \end{bmatrix}$$
$$\equiv A_1 = \begin{bmatrix} y^T \sigma y & w^T \\ U_1^T \sigma y & B \end{bmatrix} = \begin{bmatrix} \sigma \|y\|_2^2 & w^T \\ 0 & B \end{bmatrix} = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix}.$$

$$\text{Άρα,} \quad A_1 = U^T A V = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix}.$$

Όμως,

$$\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \left\| U^T A V \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \left\| U^T A V \begin{bmatrix} \sigma \\ \left(y^T A V_1\right)^T \end{bmatrix} \right\|_2^2$$
$$= \begin{bmatrix} y^T \\ U_1^T \end{bmatrix} A \left[x, V_1\right] \begin{bmatrix} \sigma \\ V_1^T A^T y \end{bmatrix} \quad \text{και}$$

$$A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix} \begin{pmatrix} \sigma \\ w \end{pmatrix} = \begin{bmatrix} \sigma^2 + w^T w \\ Bw \end{bmatrix}.$$

Συνεπώς, καταλήγουμε ότι

$$\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 \geq \left(\sigma^2 + w^T w\right)^2,$$

άρα $\|A_1\|_2^2 \geq \sigma^2 + w^T w$. Όμως, $\sigma^2 = \|A\|_2^2 = \|A_1\|_2^2 \geq \sigma^2 + w^T w$. Επομένως, για να ισχύει η προηγούμενη σχέση, πρέπει $w^T w = 0$. Άρα $w = 0$.

Έχουμε λοιπόν

$$A_1 = U^T A V = \begin{bmatrix} \sigma & 0 \\ 0 & B \end{bmatrix}, \quad \text{όπου } B \in \mathbb{R}^{(m-1)\times(n-1)}.$$

Επαναλαμβάνουμε το ίδιο επιχείρημα για τον πίνακα B και ολοκληρώνουμε την απόδειξη. ■

Οι ιδιάζουσες τιμές του πίνακα A σχετίζονται με τις ιδιοτιμές του, καθώς ισχύει:

$$\sigma_i^2 = \lambda_i \implies \sigma_i = \sqrt{\lambda_i}.$$

Στην Ανάλυση Κυρίων Συνιστωσών το θεώρημα της ανάλυσης ιδιαζουσών τιμών που μόλις αποδείχθηκε, θα φανεί ιδιαίτερα χρήσιμο, καθώς είναι μία μέθοδος αριθμητικά πιο ευσταθής από αυτήν της ανάλυσης ιδιοτιμών ενός πίνακα (π.χ. QR).

3 Ανάλυση Κυρίων Συνιστωσών (PCA)

3.1 Η μέθοδος Ανάλυσης Κυρίων Συνιστωσών

Πολλές φορές οι αναλυτές δεδομένων καλούνται να μελετήσουν δείγματα μεγάλων διαστάσεων και η εξαγωγή συμπερασμάτων από αυτά είναι πάρα πολύ δύσκολη. Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis) είναι μία μέθοδος μείωσης των διαστάσεων ενός συνόλου δεδομένων, χωρίς μεγάλη απώλεια πληροφοριών.
Κύριο χαρακτηριστικό της είναι ότι χρησιμοποιεί έναν ορθογώνιο μετασχηματισμό (αλλαγή βάσης) και καταφέρνει να μετατρέψει το αρχικό σύνολο δεδομένων (το οποίο περιέχει μεγάλο αριθμό μεταβλητών, πιθανόν συσχετιζόμενες μεταξύ τους), σε ένα σύνολο με αρκετά λιγότερες μεταβλητές, οι οποίες είναι ασυσχέτιστες μεταξύ τους. Αυτές οι καινούριες μεταβλητές ονομάζονται **κύριες συνιστώσες** και είναι γραμμικοί συνδυασμοί των μεταβλητών του αρχικού δείγματος. Ο μετασχηματισμός ορίζεται με τέτοιο τρόπο ώστε η πρώτη κύρια συνιστώσα να έχει τη μεγαλύτερη δυνατή διακύμανση (δηλαδή να αντιπροσωπεύει το μεγαλύτερο μέρος της μεταβλητότητας των δεδομένων), και κάθε επόμενη συνιστώσα με τη σειρά της να αντιστοιχεί στην υψηλότερη δυνατή διακύμανση με τον περιορισμό ότι είναι ορθογώνια προς τις προηγούμενες κύριες συνιστώσες (το οποίο σημαίνει ότι οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους).
Έστω $X \in \mathbb{R}^{m\times n}$ ένα σύνολο δεδομένων. Οι στήλες του X είναι τα n δείγματα και οι γραμμές του είναι οι m διαφορετικές ποσότητες που μελετάμε. Θέλουμε να βρούμε έναν γραμμικό μετασχηματισμό του πίνακα X σε έναν πίνακα Y έτσι ώστε

$$\mathbf{Y} = \mathbf{PX}, \qquad \text{όπου } P \in \mathbb{R}^{m\times m} \text{ κατάλληλος πίνακας.}$$

Θέλουμε να βρούμε τον μετασχηματισμό με τον οποίο θα εκφράσουμε τα δεδομένα, έτσι ώστε οι κύριες συνιστώσες της καινούριας βάσης να είναι ανεξάρτητες μεταξύ τους. Η ανάλυση κυρίων συνιστωσών εξασφαλίζει αυτή την ανεξαρτησία εντοπίζοντας τις κατευθύνσεις στις οποίες η διακύμανση των δεδομένων στην αρχική βάση είναι η μέγιστη δυνατή, κι έπειτα χρησιμοποιεί αυτές τις κατευθύνσεις για να ορίσει την καινούρια βάση.
Για να βρούμε λοιπόν τη διακύμανση των δεδομένων, χρησιμοποιούμε τον πίνακα συνδιακύμανσης[1], τον οποίο ορίζουμε

$$C_X = \frac{1}{n-1} XX^T,$$

όπου X είναι ο πίνακας των δεδομένων από τον οποίο έχουμε αφαιρέσει τις μέσες τιμές (κανονικοποίηση των δεδομένων). Ο C_X περιέχει στην κύρια διαγώνιο τις διακυμάνσεις, ενώ στα εκτός διαγωνίου στοιχεία περιέχει τις συνδιακυμάνσεις των μεταβλητών. Η συνδιακύμανση είναι το μέτρο το οποίο δείχνει πόσο συσχετίζονται δύο μεταβλητές.
Οι συνιστώσες στην καινούρια βάση πρέπει να είναι όσο γίνεται πιο ασυσχέτιστες μεταξύ τους. Επομένως, θέλουμε ο πίνακας συνδιακύμανσης του Y να έχει πραγματικές τιμές στη διαγώνιο, και στα εκτός διαγωνίου στοιχεία (που αντιπροσωπεύουν τις τιμές της συνδιακύμανσης) να

[1] Εναλλακτικά χρησιμοποιούμε τον πίνακα συσχέτισης. Οι ορισμοί δίνονται στην ενότητα 2.1

έχει μηδενικές (ή πολύ κοντά στο μηδέν) τιμές. Έτσι πρέπει να διαλέξουμε τον πίνακα P τέτοιο ώστε να διαγωνιοποιεί τον C_Y.
Ορίζουμε λοιπόν τον πίνακα συνδιακύμανσης του Y ως εξής

$$\begin{aligned} C_Y &= \frac{1}{n-1}\, Y\, Y^T \\ &= \frac{1}{n-1}\, (PX)\,(PX)^T \\ &= \frac{1}{n-1}\, P\left(XX^T\right) P^T. \end{aligned}$$

Παρατηρούμε ότι ο πίνακας $\frac{1}{n-1}XX^T$ είναι συμμετρικός (δηλαδή έχει πραγματικές ιδιοτιμές) κι έτσι, υπάρχει πραγματικός ορθογώνιος πίνακας Q τέτοιος ώστε

$$Q^T \left(\frac{1}{n-1}XX^T\right) Q = D,$$

όπου $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Άρα πρέπει να διαλέξουμε $P = Q^T$.
Βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης C_X. Ταξινομούμε τις ιδιοτιμές σε φθίνουσα σειρά και στη συνέχεια κατασκευάζουμε τον ορθοκανονικό πίνακα Q, ταξινομώντας τα ιδιοδιανύσματά του στις αντίστοιχες θέσεις των ιδιοτιμών τους. Έτσι πετυχαίνουμε να διαγωνιοποιήσουμε τον πίνακα συνδιακύμανσης των μετασχηματισμένων δεδομένων Y. Οι κύριες συνιστώσες είναι τα ιδιοδιανύσματα του C_X και είναι ταξινομημένα έτσι ώστε η πρώτη κύρια συνιστώσα να είναι η πιο σημαντική, η δεύτερη κύρια συνιστώσα να είναι η δεύτερη σημαντική, κ.ο.κ.

3.1.1 Ανάλυση Κυρίων Συνιστωσών με χρήση της SVD

Στην ενότητα αυτή θα μελετήσουμε κατά πόσο μπορούμε να χρησιμοποιήσουμε το θεώρημα της ανάλυσης ιδιαζουσών τιμών[2] ώστε να κάνουμε την ανάλυση κυρίων συνιστωσών. Έστω ότι έχουμε πάλι τον πίνακα δεδομένων $X \in \mathbb{R}^{m\times n}$ από τον οποίο έχουμε αφαιρέσει τις μέσες τιμές. Ορίζουμε $Z = \frac{1}{\sqrt{n-1}}X^T$. Τότε, ο πίνακας συνδιακύμανσης είναι

$$Z^T Z = \frac{1}{n-1}XX^T = C_X.$$

Οι κύριες συνιστώσες του πίνακα X, όπως είδαμε νωρίτερα, είναι τα ιδιοδιανύσματα του C_X. Έστω λοιπόν, ότι έχουμε την ανάλυση ιδιαζουσών τιμών (SVD) του πίνακα Z.

$$\begin{aligned} U^T Z V = \Sigma \qquad &(\text{Πολλαπλασιάζουμε αριστερά με } U) \implies \\ ZV = U\Sigma \qquad &(\text{Πολλαπλασιάζουμε δεξιά με } V^T) \implies \\ \mathbf{Z = U\Sigma V^T}, \quad &\text{όπου } \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p). \end{aligned}$$

[2] Απόδειξη στην ενότητα 2.2.2

Επομένως, ο πίνακας συνδιακύμανσης γίνεται

$$\begin{aligned} C_X = Z^T Z &= \left(U\Sigma V^T\right)^T \left(U\Sigma V^T\right) \\ &= V\Sigma^T \underbrace{U^T U}_{\text{ορθογώνιος}} \Sigma V^T \\ &= V\Sigma^T\Sigma V^T. \end{aligned}$$

Αν κάνουμε κατάλληλους πολλαπλασιασμούς πινάκων, θα έχουμε

$$\begin{aligned} C_X &= V\Sigma^T\Sigma V^T && (\text{Πολλαπλασιάζουμε αριστερά με } V^T) \implies \\ V^T C_X &= \Sigma^T\Sigma V^T && (\text{Πολλαπλασιάζουμε δεξιά με } V) \implies \\ V^T C_X V &= \underbrace{\Sigma^T\Sigma}_{\text{διαγώνιος}}. \end{aligned}$$

Παρατηρούμε ότι οι μη μηδενικές ιδιάζουσες τιμές του πίνακα Z είναι οι τετραγωνικές ρίζες των μη μηδενικών ιδιοτιμών του πίνακα ZZ^T. Ισχύει δηλαδή ότι

$$\sigma_i^2 = \lambda_i \implies \sigma_i = \sqrt{\lambda_i}.$$

Ο πίνακας C_X είναι συμμετρικός (πραγματικές ιδιοτιμές) και υπάρχει ο ορθογώνιος πίνακας V^T ο οποίος τον διαγωνιοποιεί. Στο σημείο αυτό βλέπουμε την άμεση σύνδεση της SVD με τα προηγούμενα. Έτσι λοιπόν, ο μετασχηματισμός που ψάχνουμε είναι

$$\mathbf{Y} = \mathbf{V^T X}.$$

Τέλος, εάν επιθυμούμε να προβάλουμε τα δεδομένα στην αρχική τους βάση, τότε εφαρμόζουμε τον μετασχηματισμό

$$\mathbf{X} = \mathbf{V}\ \mathbf{Y},$$

ο οποίος ισχύει επειδή ο V είναι ορθογώνιος πίνακας.

3.2 Παραδείγματα

3.2.1 Παράδειγμα "Διατροφικές Συνήθειες"

Για την κατανόηση της ανάλυσης κυρίων συνιστωσών, θα παρουσιάσουμε ένα πρώτο παράδειγμα. Έχουμε δεδομένα για 17 τροφές που καταναλώθηκαν από 4 διαφορετικές περιοχές του Ηνωμένου Βασιλείου το 1998 [3]. Θέλουμε να εξετάσουμε αν τα δεδομένα αυτά σχετίζονται με κάποιο τρόπο μεταξύ τους, δηλαδή αν υπάρχουν κοινά χαρακτηριστικά στην κατανάλωση των τροφών στις τέσσερις αυτές περιοχές. Στον πίνακα που ακολουθεί δεν μπορούμε να βγάλουμε εύκολα συμπέρασμα κοιτώντας τα δεδομένα, καταλαβαίνουμε λοιπόν πόσο δύσκολο θα είναι αυτό όταν έχουμε δείγματα μεγαλύτερου μεγέθους.

Consumption	**England**	**Wales**	**Scotland**	**N. Ireland**
Cheese	105	103	103	66
Carcase meat	245	227	242	267
Meat products	685	803	750	586
Fish	147	160	122	93
Fats and oils	193	235	184	209
Sugar	156	175	147	139
Fresh potatoes	720	874	566	1033
Fresh green vegetables	253	265	171	143
Other fresh vegetables	488	570	418	355
Processed potatoes	198	203	220	187
Processed vegetables	360	365	337	334
Fruit	1102	1137	957	674
Cereals	1472	1582	1462	1494
Beverages	57	73	53	47
Soft drinks	1374	1256	1572	1506
Alcoholic drinks	375	475	458	135
Confectionery	54	64	62	41

Βάζουμε τα δεδομένα στον πίνακα X, του οποίου οι γραμμές αντιστοιχούν στις 17 διαφορετικές τροφές και οι στήλες στις 4 περιοχές. Στο διάγραμμα πλαισίων και απολήξεων (Box plot) που ακολουθεί, έχουμε την αναπαράσταση των δεδομένων του προηγούμενου πίνακα. Κάθε πλαίσιο (box) αντιστοιχεί σε μία από τις 17 τροφές που έχουμε. Στο box plot προσδιορίζονται με άμεσο τρόπο:

- η θέση των δεδομένων με τη διάμεσο (κόκκινη οριζόντια γραμμή μέσα σε κάθε πλαίσιο),
- η διασπορά τους με το μήκος του πλαισίου και των καθέτων διακεκομμένων γραμμών (whiskers) που εκτείνονται έξω από αυτό,
- η ύπαρξη ακραίων τιμών (outliers) με μεμονωμένα σημεία στα δεδομένα.

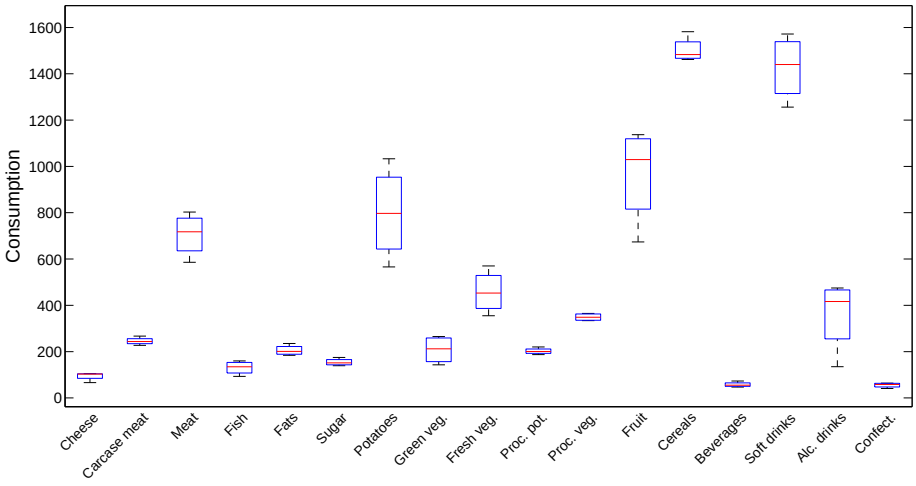

Figure 3.1: Box plot αρχικών δεδομένων.

Η διαδικασία που ακολουθούμε είναι η εξής:

- Για κάθε γραμμή x_i, $\quad 1 \leq i \leq 17$ του πίνακα υπολογίζουμε τη μέση τιμή $\overline{x}_i$ της γραμμής.
- Αφαιρούμε από κάθε γραμμή τη μέση τιμή που υπολογίσαμε και καταλήγουμε στον νέο πίνακα X με γραμμές $x_i = x_i - \overline{x}_i$.
- Έπειτα, υπολογίζουμε τον πίνακα συνδιακύμανσης $C_X = \frac{1}{n-1} X X^T$ για $n = 4$.
- Βρίσκουμε τις ιδιοτιμές $Q \in \mathbb{R}^{17\times 1}$ και τα ιδιοδιανύσματα $V \in \mathbb{R}^{17\times 17}$ του πίνακα C_X. Ταξινομούμε τις ιδιοτιμές και τα ιδιοδιανύσματά τους σε φθίνουσα σειρά (εφόσον γνωρίζουμε ότι στο ιδιοδιάνυσμα με την μέγιστη ιδιοτιμή αντιστοιχεί η πρώτη κύρια συνιστώσα, στο αμέσως επόμενο η δεύτερη, και ούτω καθ'εξής).
- Μετασχηματίζουμε τα δεδομένα μας χρησιμοποιώντας τον ορθογώνιο πίνακα των ιδιοδιανυσμάτων: ορίζουμε $Y = V^T X$.

Σημειώνουμε πως μπορούμε να ανακτήσουμε τα αρχικά δεδομένα σε μόλις δύο βήματα: αρκεί να πολλαπλασιάσουμε από αριστερά τον πίνακα Y με τον πίνακα ιδιοδιανυσμάτων V, και στο αποτέλεσμα να προσθέσουμε (κατά γραμμές) τις μέσες τιμές που υπολογίσαμε αρχικά.

Ο πίνακας Y που προκύπτει από την PCA είναι η προβολή των αρχικών δεδομένων στο ορθοκανονικό σύστημα αξόνων που ορίζουν οι κύριες συνιστώσες. Προβάλοντας τις 17 τροφές πάνω στις κύριες συνιστώσες μπορούμε να καταλάβουμε περισσότερα για τα δεδομένα μας. Συγκεκριμένα, στα γραφήματα που ακολουθούν, παρατηρούμε μία ομαδοποίηση των δεδομένων. Οι περιοχές England, Wales και Scotland συσχετίζονται, ενώ η N. Ireland δε φαίνεται να έχει κοινά χαρακτηριστικά με τις προηγούμενες.

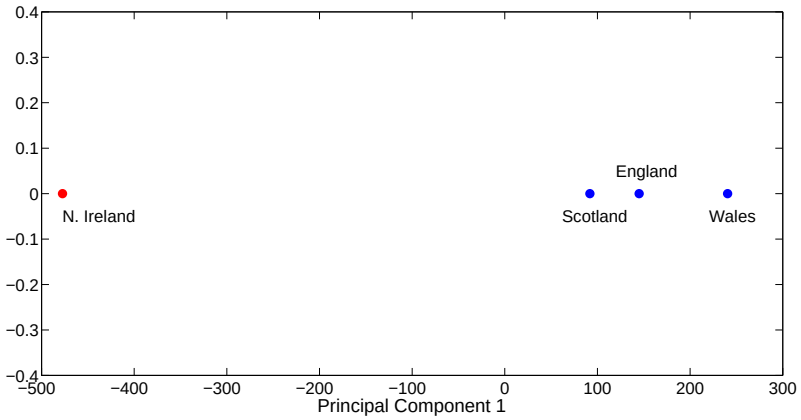

Figure 3.2: Προβολή πάνω στην πρώτη κύρια συνιστώσα.

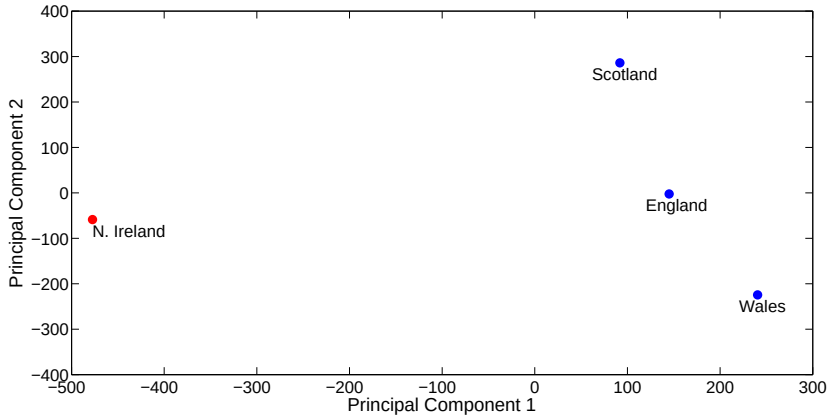

Figure 3.3: Προβολή στις διαστάσεις που ορίζουν οι 2 πρώτες κύριες συνιστώσες.

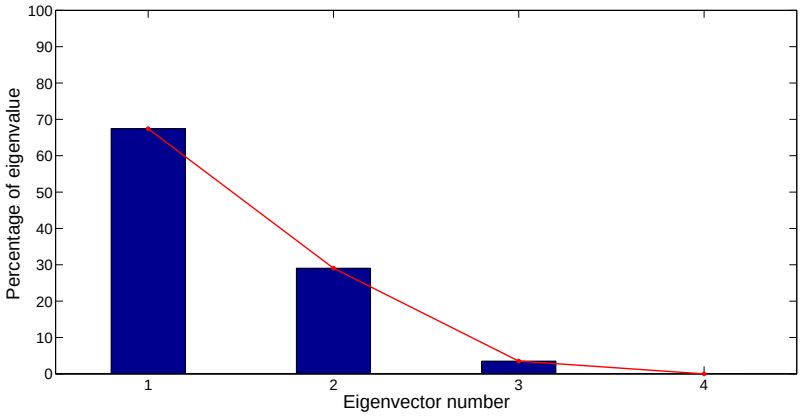

Figure 3.4: Φάσμα ιδιοτιμών.

Η ομαδοποίηση που παρατηρούμε στα προηγούμενα γραφήματα, σαφώς δεν είναι τυχαία καθώς, στο τελευταίο διάγραμμα βλέπουμε το φάσμα των ιδιοτιμών (eigenspectrum) για τα τέσσερα πρώτα ιδιοδιανύσματα. Παρατηρούμε πως ήδη από τα δύο πρώτα ιδιοδιανύσματα παίρνουμε το μεγαλύτερο ποσοστό της διακύμανσης, πράγμα που σημαίνει ότι η μεγαλύτερη πληροφορία για τα δεδομένα δίνεται κυρίως από τις δύο πρώτες κύριες συνιστώσες.

3.2.2 Παράδειγμα "Iris Flower Data set"

Στο παράδειγμα που ακολουθεί, χρησιμοποιούμε το Iris Flower data set [5]. Ο Ronald A. Fisher το 1936, συνέλεξε 50 δείγματα για κάθε ένα από τα τρία είδη (οικογένειες) του λουλουδιού Ίριδα (Iris setosa, Iris virginica και Iris versicolor). Τέσσερα χαρακτηριστικά μετρήθηκαν από κάθε δείγμα: το μήκος και το πλάτος των σεπάλων και των πετάλων, σε εκατοστά.
Το σύνολο αυτών των δεδομένων (επειδή είναι ένα χαρακτηριστικό πολυδιάστατο dataset) υπάρχει ήδη στη Matlab σε ένα mat-file και περιέχει έναν πίνακα και ένα διάνυσμα. Ο πίνακας περιέχει τις μετρήσεις των τεσσάρων χαρακτηριστικών που προαναφέρθηκαν για κάθε δείγμα και έχει διαστάσεις 150×4, ενώ το διάνυσμα περιέχει την πληροφορία για το είδος του εκάστοτε λουλουδιού και είναι διάστασης 150×1.
Θα κάνουμε ανάλυση κυρίων συνιστωσών προκειμένου να δούμε αν τα δείγματα του κάθε είδους έχουν όμοια χαρακτηριστικά μεταξύ τους. Τα βήματα που ακολουθούμε είναι ανάλογα με αυτά του προηγούμενου παραδείγματος όμως, εδώ χρησιμοποιούμε την ανάλυση ιδιαζουσών τιμών (SVD) του πίνακα συνδιακύμανσης των δεδομένων.

- Θεωρούμε τον πίνακα X ως τον ανάστροφο των αρχικών δεδομένων (για πιο εύκολο χειρισμό). Βρίσκουμε τις μέσες τιμές του X κατά γραμμές και τις αφαιρούμε από τα στοιχεία του πίνακα.

- Υπολογίζουμε τον πίνακα συνδιακύμανσης $C_X = \frac{1}{n-1} X X^T$ για $n = 150$ (αριθμός δειγμάτων).

- Κάνουμε την ανάλυση ιδιαζουσών τιμών του πίνακα C_X και βρίσκουμε το διαγώνιο πίνακα $S \in \mathbb{R}^{4\times4}$ με τις ιδιάζουσες τιμές στη διαγώνιο, και τον πίνακα $V \in \mathbb{R}^{4\times4}$ με στήλες τα ιδιοδιανύσματα που τους αντιστοιχούν.

- Προβάλουμε τα δεδομένα στους άξονες που ορίζουν οι κύριες συνιστώσες, ορίζουμε $Y = V^T X$.

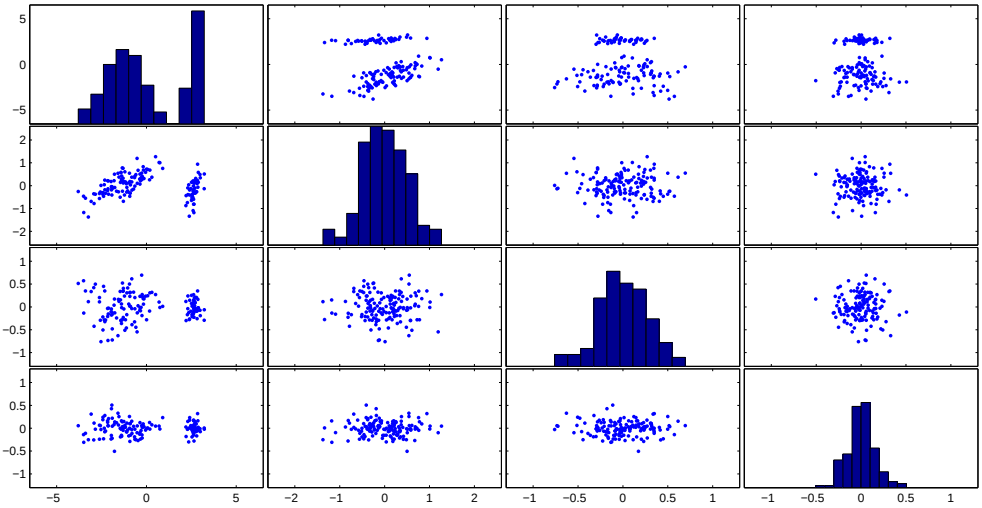

Figure 3.5: Στο διάγραμμα έχουμε τις προβολές των κυρίων συνιστωσών ανά δύο. Στη διαγώνιο βλέπουμε το ιστόγραμμα της κάθε κύριας συνιστώσας.

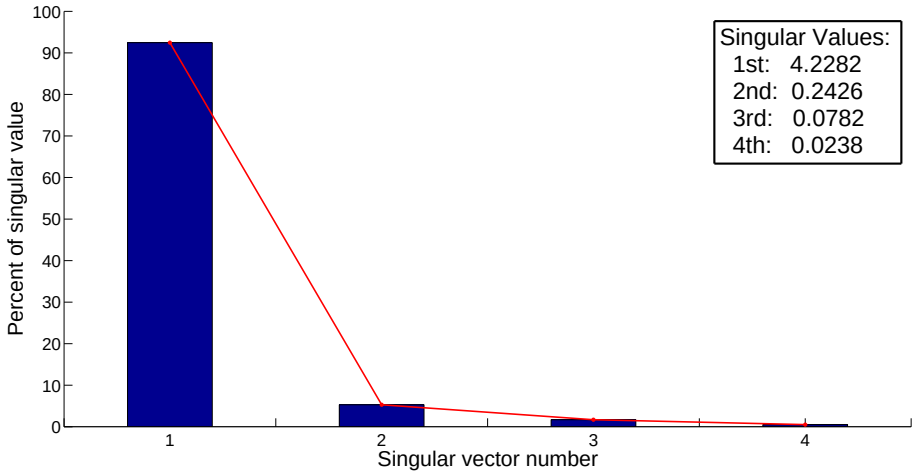

Figure 3.6: Στο διάγραμμα έχουμε το φάσμα των ιδιαζουσών τιμών σε κλίμακα του εκατό. Παρατηρούμε ότι η πρώτη ιδιάζουσα τιμή, η οποία είναι και η μεγαλύτερη, έχει σημαντική διαφορά από τις υπόλοιπες, καθώς καλύπτει ποσοστό άνω του 90% της συνολικής διακύμανσης.

Στα διαγράμματα που ακολουθούν, προβάλουμε τα δεδομένα ως προς την πρώτη κύρια συνιστώσα (η οποία αντιστοιχεί στην μεγαλύτερη ιδιάζουσα τιμή) σε συνδυασμό με κάθε μία από τις υπόλοιπες κύριες συνιστώσες. Σε κάθε ένα από αυτά, παρατηρείται εύκολα η ομαδοποίηση των δεδομένων ως προς τις τρεις οικογένειες της ίριδας.

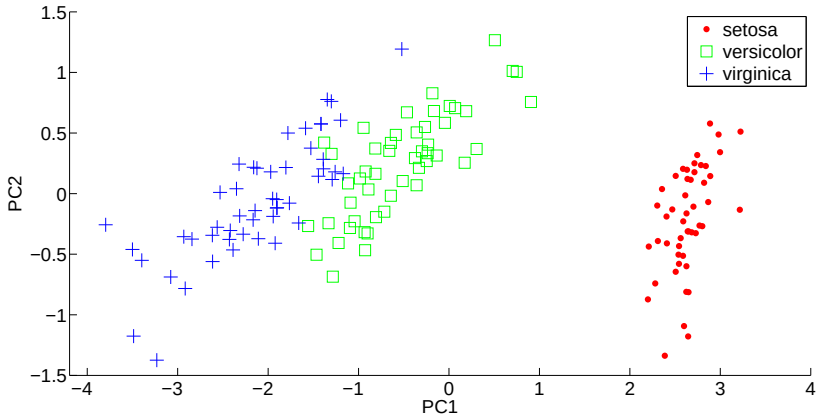

Figure 3.7: Προβολή δεδομένων στους άξονες της πρώτης και της δεύτερης κύριας συνιστώσας.

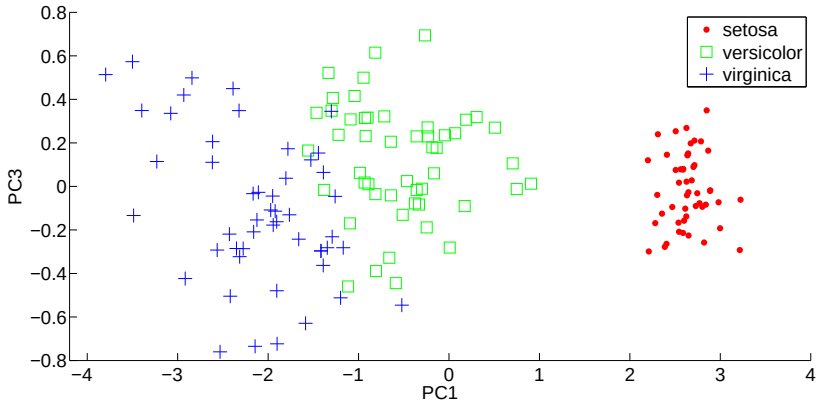

Figure 3.8: Προβολή δεδομένων στους άξονες της πρώτης και της τρίτης κύριας συνιστώσας.

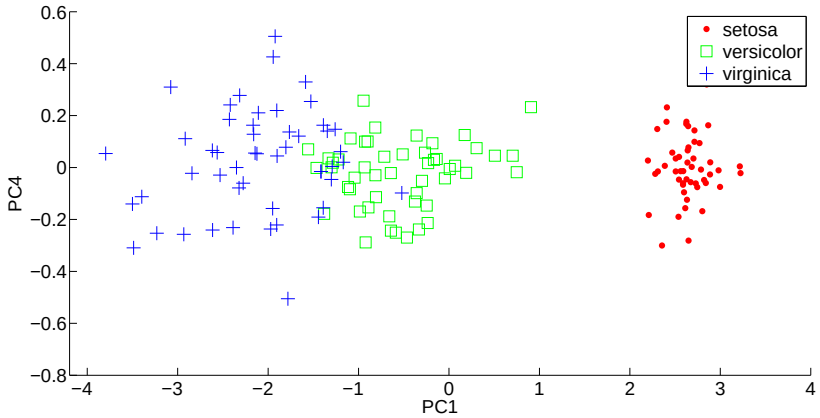

Figure 3.9: Προβολή δεδομένων στους άξονες της πρώτης και της τέταρτης κύριας συνιστώσας.

Αντίθετα, αν κοιτάξουμε τις προβολές μεταξύ των κυρίων συνιστωσών 2, 3 και 4 δεν υπάρχει καμία εμφανής ομαδοποίηση. Έτσι λοιπόν, επαληθεύεται η θεωρία της ανάλυσης κυρίων συνιστωσών, κατά την οποία, αν ακολουθήσουμε την κατεύθυνση στην οποία μεγιστοποιείται η διακύμανση των αρχικών δεδομένων, τότε μπορούμε να εξάγουμε χρήσιμες πληροφορίες γι'αυτά. Συμπερασματικά, η πρώτη κύρια συνιστώσα, η οποία αντιστοιχεί στο μήκος των σεπάλων, είναι η συνιστώσα που δίνει την πιο ξεκάθαρη κατηγοριοποίηση των λουλουδιών στις τρεις οικογένειες της ίριδας.

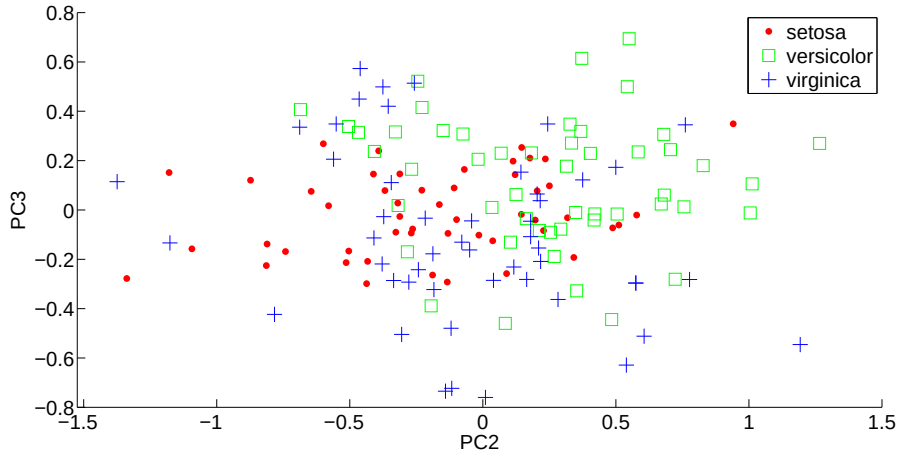

Figure 3.10: Προβολή δεδομένων στους άξονες της δεύτερης και της τρίτης κύριας συνιστώσας.

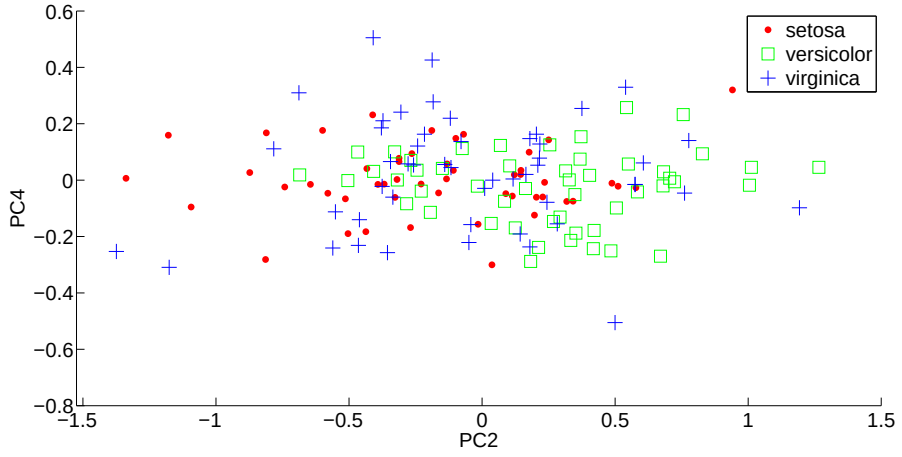

Figure 3.11: Προβολή δεδομένων στους άξονες της δεύτερης και της τέταρτης κύριας συνιστώσας.

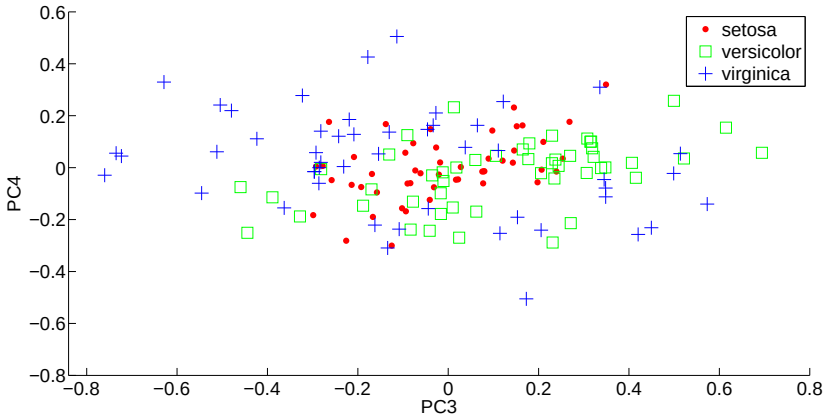

Figure 3.12: Προβολή δεδομένων στους άξονες της τρίτης και της τέταρτης κύριας συνιστώσας.

3.2.3 Συμπίεση εικόνας με χρήση της PCA

Η Ανάλυση Κυρίων Συνιστωσών έχει μεγάλο εύρος εφαρμογών, το οποίο φτάνει μέχρι και την επεξεργασία και συμπίεση εικόνας. Στο παράδειγμα αυτό θα εφαρμόσουμε τη μέθοδο με τη χρήση της SVD και θα δούμε ότι αποδεικνύεται μία πολύ καλή μέθοδος για να κάνουμε συμπίεση μίας εικόνας. Αρχικά, έχουμε την εικόνα Butterfly, η οποία έχει διαστάσεις 512×512 εικονοστοιχείων (pixels).

Figure 3.13: Η ασπρόμαυρη εικόνα 'Butterfly' με την οποία θα κάνουμε δοκιμή της PCA.

Η Matlab χειρίζεται την εικόνα ως δύο πίνακες: στον έναν, έστω X, κρατάει πληροφορίες για κάθε εικονοστοιχείο και στον άλλον πληροφορίες για το εύρος χρώματος (colormap). Εφαρμόζουμε τη μέθοδο στον πίνακα X ο οποίος έχει διαστάσεις 512×512. Τα στοιχεία του X εκφράζουν την απόχρωση του γκρι σε κάθε εικονοστοιχείο, με τιμές ανάμεσα στο 0 (μαύρο) και στο 1 (άσπρο). Εφαρμόζουμε την ανάλυση κυρίων συνιστωσών με την ίδια διαδικασία όπως στα προηγούμενα παραδείγματα και βρίσκουμε τις κύριες συνιστώσες. Εάν

προβάλουμε τα δεδομένα στην πρώτη κύρια συνιστώσα, η οποία αντιστοιχεί στα στοιχεία με τη μεγαλύτερη διακύμανση, θα έχουμε τον βασικό "σκελετό" της εικόνας χωρίς όμως λεπτομέρειες. Αν προσθέσουμε παραπάνω κύριες συνιστώσες, παίρνουμε μία πιο ξεκάθαρη απόδοση της αρχικής εικόνας. Ακολουθούν γραφήματα στα οποία βλέπουμε τα αποτελέσματα της ανάλυσης κυρίων συνιστωσών.

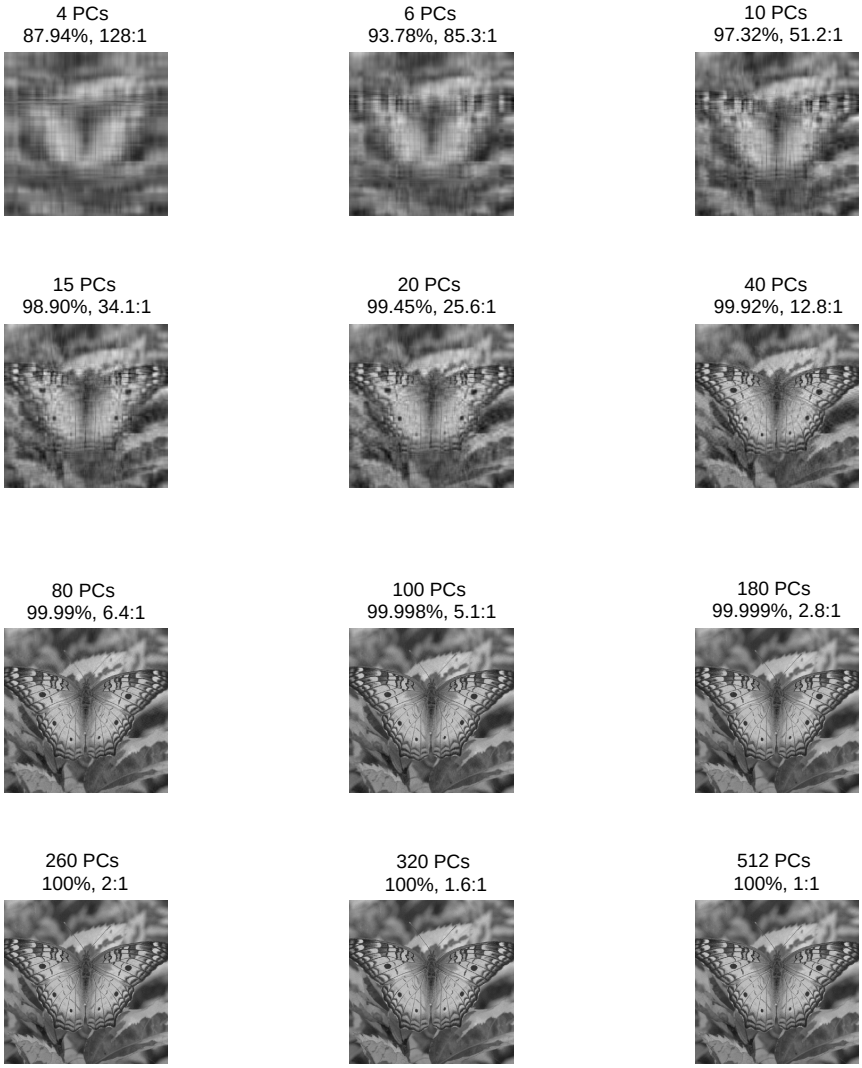

Figure 3.14: Βλέπουμε την εικόνα που παράγουμε με την χρήση των 4, 6, 10, 15 κλπ. πρώτων κυρίων συνιστωσών, το ποσοστό της πληροφορίας που παίρνουμε καθώς και τον βαθμό συμπίεσης της αρχικής εικόνας.

Είναι αξιοσημείωτο πως παίρνοντας μόλις τις 40 πρώτες κύριες συνιστώσες από τις 512 αρχικές, η εικόνα που παίρνουμε είναι αρκετά κοντά στην αρχική (με ποσοστό 99.92% της συνολικής

διακύμανσης), ενώ με 80 κύριες συνιστώσες η διαφορά από την αρχική εικόνα είναι ελάχιστη (έχουμε το 99.99% της αρχικής διακύμανσης). Αυτό επιβεβαιώνει ότι η μέθοδος ανάλυσης κυρίων συνιστωσών δίνει τη δυνατότητα απλούστευσης ενός μεγάλου συνόλου δεδομένων σε ένα αρκετά μικρότερο, χωρίς σημαντική απώλεια πληροφορίας.

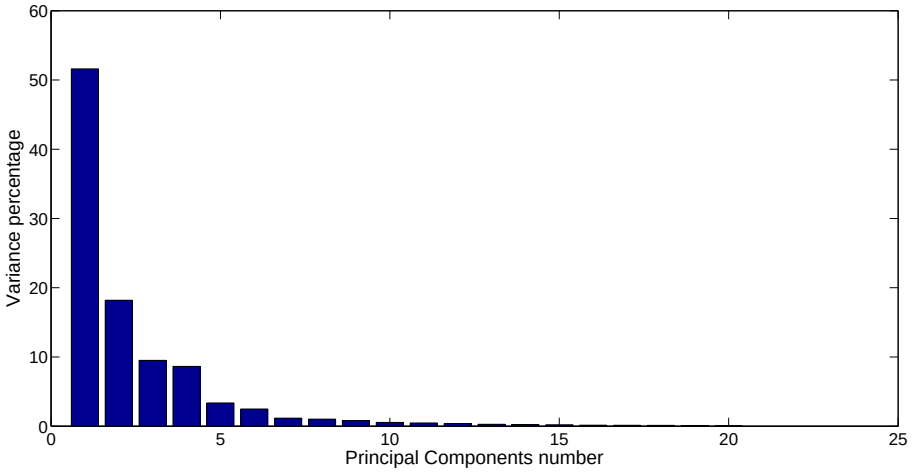

Figure 3.15: Διάγραμμα με το ποσοστό της διακύμανσης των πρώτων 20 κυρίων συνιστωσών.

Στο διάγραμμα παρατηρούμε ότι η μέγιστη διακύμανση των δεδομένων βρίσκεται στις πρώτες κύριες συνιστώσες. Έχουμε εμφανίσει τις πρώτες 20 και φαίνεται ότι ήδη από την 14^{η} συνιστώσα κι έπειτα, η διακύμανση είναι πάρα πολύ μικρή. Συμπερασματικά, μπορούμε να επιλέξουμε μόνο τις n πρώτες κύριες συνιστώσες, οι οποίες αθροιστικά δίνουν πάνω από το 99% της αρχικής διακύμανσης, κι έτσι να απλοποιήσουμε κατά πολύ τις διαστάσεις των δεδομένων.

Σημείωση:
Μπορούμε να εργαστούμε ανάλογα και στις έγχρωμες εικόνες, οι οποίες ακολουθούν το πρότυπο RGB[3]. Η Matlab χειρίζεται την έγχρωμη εικόνα με έναν τρισδιάστατο πίνακα διαστάσεων $N \times N \times 3$. Έτσι, αν έχουμε μία εικόνα με διαστάσεις 512×512 εικονοστοιχεία, ο πίνακας θα είναι $512 \times 512 \times 3$, όπου η τρίτη διάσταση αναφέρεται σε κάθε ένα από τα τρία χρώματα. Για να εφαρμόσουμε την PCA όπως στο προηγούμενο παράδειγμα, δουλεύουμε στους τρεις υποπίνακες που ορίζουν τα τρία χρώματα. Μετά τη διαδικασία αυτή, συνδυάζουμε τους πίνακες που προκύπτουν σε μία εικόνα, κι έτσι έχουμε τα παρακάτω ενδεικτικά αποτελέσματα.

[3]Το πρότυπο χρώματος RGB είναι ένα προσθετικό πρότυπο στο οποίο τα χρώματα κόκκινο, πράσινο και μπλε συνδυάζονται με διάφορους τρόπους για να αναπαραχθούν τα άλλα χρώματα. Το όνομα του προτύπου και η σύντμηση RGB προέρχονται από τα τρία βασικά χρώματα, το κόκκινο (Red), το πράσινο (Green) και το μπλέ (Blue).

Figure 3.16: Τα αντίστοιχα αποτελέσματα της PCA για την έγχρωμη εικόνα που παράγουμε με την χρήση των 4, 6, 10, 15 κλπ. πρώτων κυρίων συνιστωσών. Στις εικόνες βλέπουμε τον αριθμό των κυρίων συνιστωσών που έχουμε μαζί με το βαθμό συμπίεσης της αρχικής εικόνας, καθώς και το ποσοστό της διακύμανσης για κάθε ένα από τα Red, Green, Blue χρώματα αντίστοιχα.

4 Η μέθοδος σε χρονοεξαρτώμενα δεδομένα

4.1 Χρονοεξαρτώμενα σύνολα δεδομένων

Η ανάλυση κυρίων συνιστωσών εφαρμόζεται ευρύτατα στις ατμοσφαιρικές και γεωφυσικές επιστήμες. Σε τέτοιου είδους σύνολα δεδομένων, εμφανίζονται μεγάλες συσχετίσεις μεταξύ των ποσοτήτων που μετρώνται, οπότε η PCA δίνει μία πιο συμπαγή αναπαράσταση της μεταβλητότητάς τους. Η πλειοψηφία των εφαρμογών της, αφορούν στην ανάλυση πεδίων όπως μετρήσεις θερμοκρασίας, βροχόπτωσης, ταχύτητας του ανέμου, κλπ. Σε αυτές τις περιπτώσεις, τα δεδομένα περιέχουν παρατηρήσεις ενός ή περισσοτέρων πεδίων και δίνονται σε μορφή χρονοσειρών.
Έστω X, ένα σύνολο με δεδομένα τα οποία έχουν συλλεχθεί σε K διαφορετικούς σταθμούς (locations), σε καθέναν από τους οποίους έχουμε μετρήσεις για μία μεταβλητή. Τα δεδομένα των K σταθμών σε μία δεδομένη χρονική στιγμή, δίνονται υπό τη μορφή $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K]^T$. Ο πίνακας δεδομένων X έχει διαστάσεις $n \times K$ (ή αλλιώς time $\times$ space), εφόσον οι μετρήσεις στους K σταθμούς έχουν γίνει σε n διαφορετικές χρονικές στιγμές. Τα δεδομένα του X σε μορφή χρονοσειράς μπορούν να γραφτούν ως

$$h(s,t) = \sum_{k=1}^{K} \alpha_k(t) e_k(s) = \alpha_s(t), \qquad s = 1, \ldots, K \ \text{ και } \ t = 1, \ldots, n,$$

όπου $\alpha_k(t)$ είναι το μέγεθος που μετράμε στον σταθμό k και $e_k(s)$ το κ-οστό μοναδιαίο διάνυσμα στον $\mathbb{R}^K$.
Η ανάλυση κυρίων συνιστωσών γίνεται με τον ίδιο τρόπο που έχουμε δει στην προηγούμενη ενότητα. Αν $X \in \mathbb{R}^{n\times K}$ τότε ο ανάστροφός του είναι $X^T \in \mathbb{R}^{K\times n}$, οπότε ο πίνακας συνδιακύμανσης δίνεται ως εξής:

$$(XX^T)_{ij} = \sum_{k=1}^{K} X_{ik} \big(X^T\big)_{kj} = \sum_{k=1}^{K} X_{ik} X_{jk}.$$

Έτσι, τελικά έχουμε

$$C(s,s') = \sum_{t} h(s,t) h(s',t).$$

Κάνουμε την ανάλυση ιδιαζουσών τιμών του πίνακα συνδιακύμανσης κι έπειτα βρίσκουμε τις κύριες συνιστώσες οι οποίες μπορούν να γραφούν ως

$$u(t) = [E]^T \mathrm{x}(t),$$

όπου E είναι ο πίνακας των ιδιοδιανυσμάτων που έχουμε βρει, και $\mathrm{x}(t)$ τα αρχικά δεδομένα. Εφόσον τα αρχικά δεδομένα είναι σε μορφή χρονοσειράς, τότε και οι κύριες συνιστώσες θα έχουν την ίδια μορφή.
Σκοπός είναι να μειώσουμε τον αριθμό των μεταβλητών των αρχικών δεδομένων. Μπορούμε να το πετύχουμε αυτό, αν αποκόψουμε έναν αριθμό συνιστωσών οι οποίες δεν δίνουν σημαντική πληροφορία για τα δεδομένα. Αν κρατήσουμε τις $M << K$ πρώτες κύριες συνιστώσες, τότε

θα έχουμε τα στοιχεία u_m από το u τα οποία αντιστοιχούν στο μεγαλύτερο ποσοστό της μεταβλητότητας των αρχικών δεδομένων. Έτσι λοιπόν, η $m-$οστή κύρια συνιστώσα, δηλαδή η προβολή των αρχικών δεδομένων x πάνω στο $m-$οστό ιδιοδιάνυσμα e_m, θα είναι

$$u_m(t) = \mathrm{e}_m^T \mathrm{x}(t) = \sum_{k=1}^{K} \mathrm{e}_{km}\ \mathrm{x}_k(t), \qquad m = 1, \ldots, M \ \text{ και } \ t = 1, \ldots, n.$$

Αν επιλέξουμε $M = K$, τότε κρατάμε όλες τις κύριες συνιστώσες, κι έχουμε το 100% της αρχικής διακύμανσης (άρα όλη την πληροφορία), χωρίς μείωση των διαστάσεων των δεδομένων, εφόσον περιέχονται όλες οι κύριες συνιστώσες, ακόμη κι εκείνες που παρουσιάζουν πολύ μικρή διακύμανση.

4.1.1 Παράδειγμα χρονοεξαρτώμενων δεδομένων

Έχουμε ένα σύνολο δεδομένων το οποίο περιέχει μετρήσεις για την καθημερινή βροχόπτωση, τις μέγιστες και ελάχιστες θερμοκρασίες των περιοχών Ithaca και Canandaigua της Νέας Υόρκης, για τον Ιανουάριο του 1987 [1]. Ο πίνακας δεδομένων αποτελείται από $n = 31$ γραμμές (όσες είναι οι μέρες του μήνα) και $k = 6$ στήλες στις οποίες αντιστοιχούν οι ποσότητες που μελετάμε. Οι τρεις πρώτες στήλες αναφέρονται στην βροχόπτωση (σε ίντσες), τις μέγιστες και ελάχιστες θερμοκρασίες (σε βαθμούς F) της Ithaca, ενώ οι τρεις τελευταίες αναφέρονται στις ίδιες ποσότητες για την Canandaigua.

Ορίζουμε τρεις πίνακες με τους οποίους θα δουλέψουμε: τον Ppt $\in \mathbb{R}^{31\times 2}$ με τις μετρήσεις της βροχόπτωσης, τον Tmax $\in \mathbb{R}^{31\times 2}$ με τις μέγιστες θερμοκρασίες, τον Tmin $\in \mathbb{R}^{31\times 2}$ με τις ελάχιστες θερμοκρασίες. Κάθε πίνακας στην πρώτη στήλη περιέχει τη χρονοσειρά για την Ithaca και στη δεύτερη τη χρονοσειρά για την Canandaigua.
Κάνουμε την ανάλυση κυρίων συνιστωσών σε κάθε έναν πίνακα ξεχωριστά, ακολουθώντας τη γνωστή διαδικασία:

• Υπολογίζουμε και αφαιρούμε τις μέσες τιμές του πίνακα (κατά στήλες) κι έπειτα υπολογίζουμε τον πίνακα συνδιακύμανσης C των δεδομένων.
• Κάνουμε την ανάλυση ιδιαζουσών τιμών του C και βρίσκουμε το διαγώνιο πίνακα $S \in \mathbb{R}^{2\times 2}$ με τις ιδιάζουσες τιμές στη διαγώνιο, και τον πίνακα $V \in \mathbb{R}^{2\times 2}$ με στήλες τα ιδιάζοντα διανύσματα που τους αντιστοιχούν.
• Προβάλουμε τα δεδομένα στους άξονες που ορίζουν οι κύριες συνιστώσες, ορίζουμε $u = XV$.
• Ανακατασκευάζουμε τα δεδομένα χρησιμοποιώντας μόνο την πρώτη κύρια συνιστώσα $XX = u_1 V_1$).

Παράγουμε τα ακόλουθα γραφήματα, στα οποία έχουμε τις χρονοσειρές των αρχικών δεδομένων σε σύγκριση με τις χρονοσειρές που κατασκευάσαμε χρησιμοποιώντας μόνο την πρώτη κύρια συνιστώσα και παρατηρούμε ότι οι χρονοσειρές σχεδόν ταυτίζονται. Αυτό συμβαίνει επειδή η πρώτη κύρια συνιστώσα αντιστοιχεί στη μεγαλύτερη ιδιάζουσα τιμή, η

οποία εκφράζει το μεγαλύτερο ποσοστό της διακύμανσης των αρχικών δεδομένων, το οποίο είναι μεγαλύτερο από το 96%.

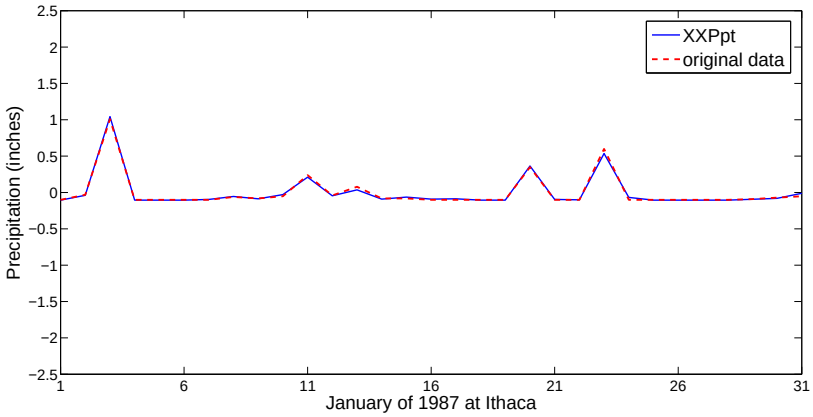

Figure 4.1: Χρονοσειρές μετρήσεων βροχόπτωσης της Ithaca.

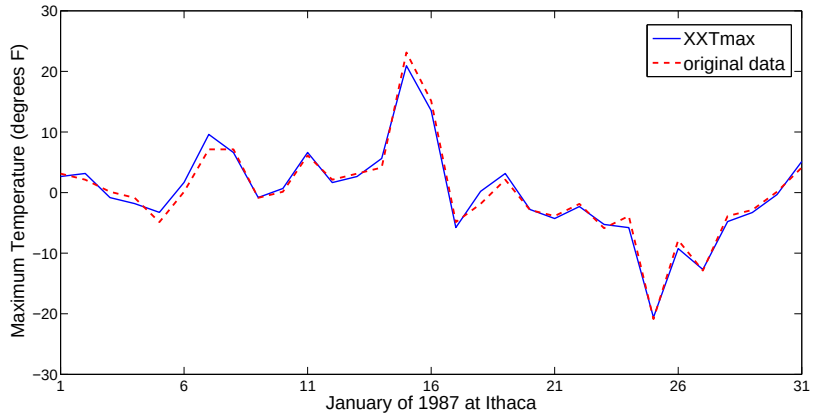

Figure 4.2: Χρονοσειρές μετρήσεων μέγιστης θερμοκρασίας της Ithaca.

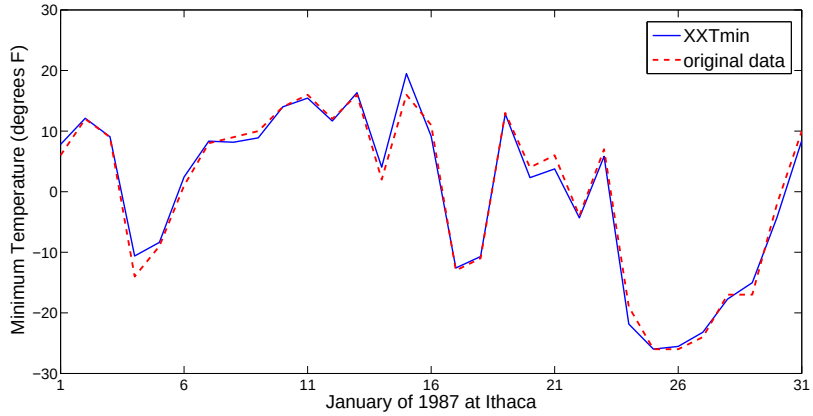

Figure 4.3: Χρονοσειρές μετρήσεων ελάχιστης θερμοκρασίας της Ithaca.

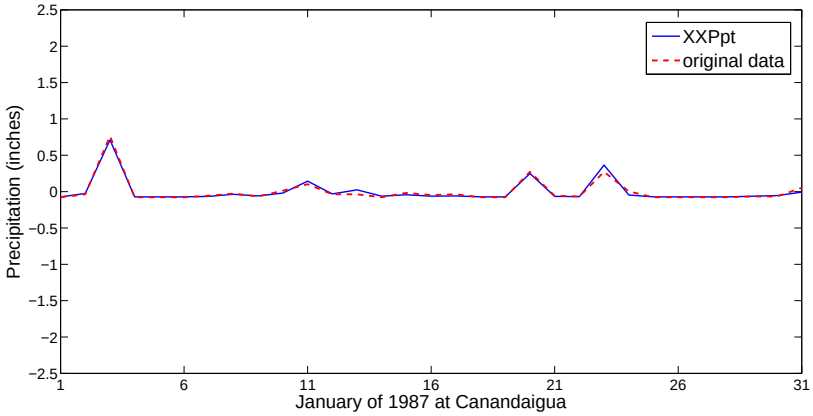

Figure 4.4: Χρονοσειρές μετρήσεων βροχόπτωσης της Canandaigua.

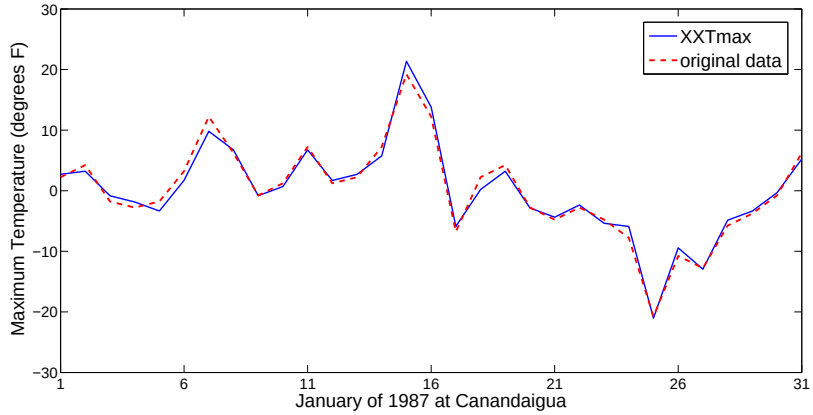

Figure 4.5: Χρονοσειρές μετρήσεων μέγιστης θερμοκρασίας της Canandaigua.

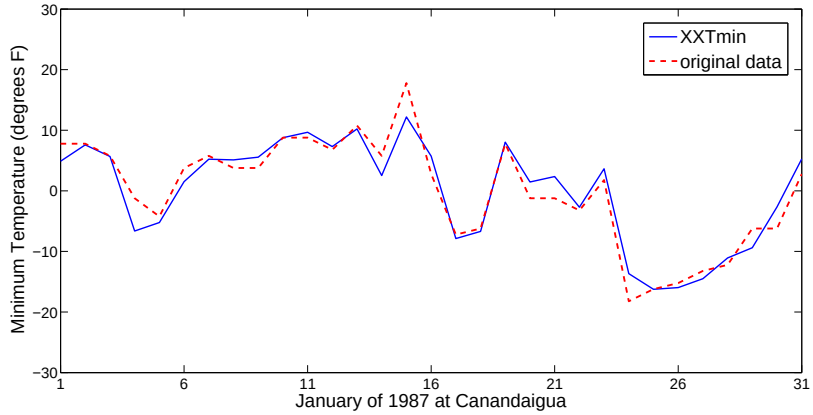

Figure 4.6: Χρονοσειρές μετρήσεων ελάχιστης θερμοκρασίας της Canandaigua.

4.2 Πίνακας συνδιακύμανσης - πίνακας συσχέτισης

Η ανάλυση κυρίων συνιστωσών μπορεί να διεξαχθεί είτε με τη χρήση του πίνακα συνδιακύμανσης είτε με τη χρήση του πίνακα συσχέτισης[4] των δεδομένων. Οι δύο αυτές εναλλακτικές κατά κύρια βάση, δε δίνουν ισοδύναμη πληροφορία για τα δεδομένα. Είναι σημαντικό να γίνει έξυπνη επιλογή ανάμεσα στους δύο πίνακες κι αυτό εξαρτάται από το σύνολο δεδομένων που εξετάζουμε κάθε φορά.
Αν ο στόχος είναι να εντοπίσουμε τη μεγαλύτερη μεταβλητότητα στα δεδομένα, τότε συνιστάται η χρήση του πίνακα συνδιακύμανσης C. Διαφορετικά, αν η ανάλυση γίνεται σε μεταβλητές που δεν μετρώνται στις ίδιες μονάδες, τότε είναι προτιμότερη η χρήση του πίνακα συσχέτισης R. Ο πίνακας συσχέτισης δεν επηρεάζεται από αλλαγή κλίμακας στα δεδομένα, εφόσον όλα τα στοιχεία στην κύρια διαγώνιο είναι 1 κι έτσι οι συσχετίσεις παραμένουν ίδιες. Από τον πίνακα συνδιακύμανσης C, έχουμε δει ότι μπορούμε να υπολογίσουμε εύκολα τον πίνακα συσχέτισης R, οπότε σε κάθε πρόβλημα μένει να επιλέγουμε κάθε φορά τον σωστό πίνακα με τον οποίο θα εργαστούμε.

Στους πίνακες που ακολουθούν, έχουμε τα αποτελέσματα της ανάλυσης κυρίων συνιστωσών, πρώτα με τη χρήση του πίνακα συνδιακύμανσης κι έπειτα με την χρήση του πίνακα συσχέτισης. Στο συγκεκριμένο παράδειγμα, είναι καλύτερη η χρήση του πίνακα συνδιακύμανσης εφόσον εξετάζουμε τα δεδομένα ξεχωριστά για κάθε μία από τις τρεις ποσότητες που μετράμε.

Αποτελέσματα με χρήση του πίνακα συνδιακύμανσης:

Μεταβλητή	Διακύμανση	e_1	e_2
Ithaca Tmin	185.4667	−0.8479	−0.5302
Canandaigua Tmin	77.5806	−0.5302	0.8479
Ιδιοτιμές		254.7571	8.2902
Ποσοστό (%)		96.8484	100.0000
Ithaca Tmax	59.5161	−0.7000	−0.7142
Canandaigua Tmax	61.8473	−0.7142	0.7000
Ιδιοτιμές		118.7633	2.6001
Ποσοστό (%)		97.8576	100.0000
Ithaca Ppt	0.0590	−0.8263	−0.5632
Canandaigua Ppt	0.0281	−0.5632	0.8263
Ιδιοτιμές		0.0858	0.0013
Ποσοστό (%)		98.4945	100.0000

Οι διαφορές δεν είναι σημαντικά μεγάλες στους αριθμούς που βλέπουμε στους παρακάτω πίνακες. Ωστόσο, εάν για τις ελάχιστες θερμοκρασίες (Tmin) βρούμε τις κύριες συνιστώσες μέσω του πίνακα συσχέτισης, τότε θα διαπιστώσουμε ότι τα γραφήματα των χρονοσειρών δεν είναι τόσο κοντά στα αρχικά δεδομένα, όσο στα αντίστοιχα γραφήματα των χρονοσειρών που είδαμε νωρίτερα.

[4] Ο ορισμός δίνεται στην ενότητα 2.1

Αποτελέσματα με χρήση του πίνακα συσχέτισης:

Μεταβλητή	Διακύμανση	e_1	e_2
Ithaca Tmin	1.0000	−0.7071	−0.7071
Canandaigua Tmin	1.0000	−0.7071	0.7071
Ιδιοτιμές		1.9237	0.0763
Ποσοστό (%)		96.1849	100.0000
Ithaca Tmax	1.0000	−0.7071	−0.7071
Canandaigua Tmax	1.0000	−0.7071	0.7071
Ιδιοτιμές		1.9571	0.0429
Ποσοστό (%)		97.8568	100.0000
Ithaca Ppt	1.0000	−0.7071	−0.7071
Canandaigua Ppt	1.0000	−0.7071	0.7071
Ιδιοτιμές		1.9655	0.0345
Ποσοστό (%)		98.2740	100.0000

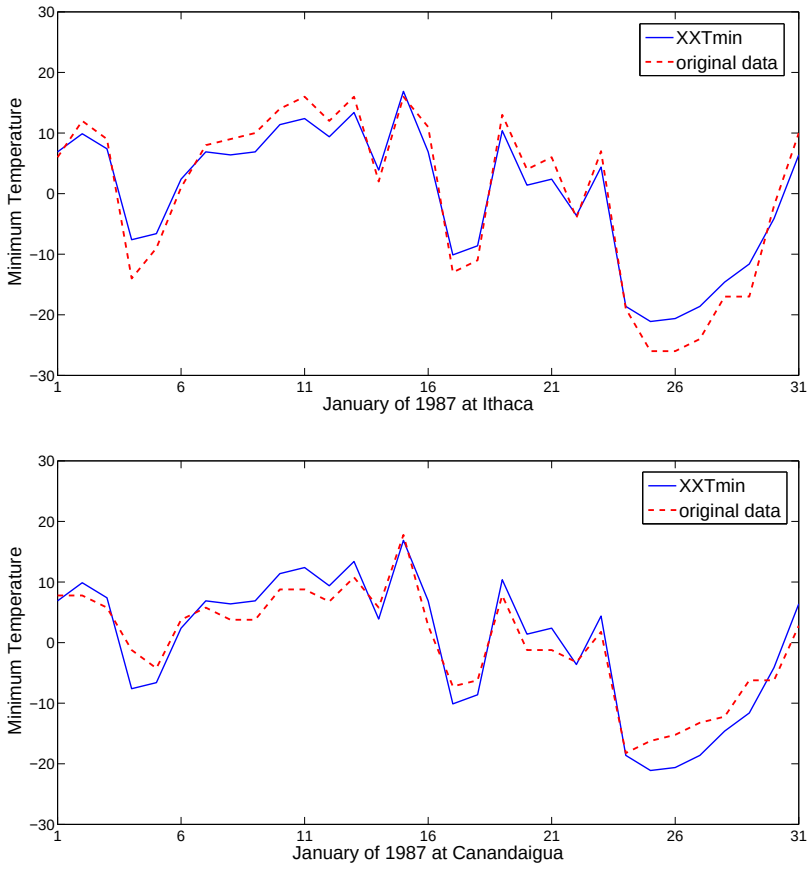

Figure 4.7: Χρονοσειρές μετρήσεων ελάχιστης θερμοκρασίας της Ithaca (πρώτο γράφημα) και της Canandaigua (δεύτερο γράφημα). Οι χρονοσειρές αυτές κατασκευάστηκαν από την πρώτη κύρια συνιστώσα, που προέκυψε από την PCA με τη χρήση του πίνακα συσχέτισης.

5 Παράρτημα

5.1 Δεδομένα

Table 1: Παρατηρήσεις καθημερινής βροχόπτωσης (σε ίντσες) και θερμοκρασίας (σε °F) των περιοχών Ithaca και Canandaigua της Νέας Υόρκης για τον Ιανουάριο του 1987 [1].

	Ithaca			**Canandaigua**		
Date	Precip.	Max Temp.	Min Temp.	Precip.	Max Temp.	Min Temp.
1	0.00	33	19	0.00	34	28
2	0.07	32	25	0.04	36	28
3	1.11	30	22	0.84	30	26
4	0.00	29	−1	0.00	29	19
5	0.00	25	4	0.00	30	16
6	0.00	30	14	0.00	35	24
7	0.00	37	21	0.02	44	26
8	0.04	37	22	0.05	38	24
9	0.02	29	23	0.01	31	24
10	0.05	30	27	0.09	33	29
11	0.34	36	29	0.18	39	29
12	0.06	32	25	0.04	33	27
13	0.18	33	29	0.04	34	31
14	0.02	34	15	0.00	39	26
15	0.02	53	29	0.06	51	38
16	0.00	45	24	0.03	44	23
17	0.00	25	0	0.04	25	13
18	0.00	28	2	0.00	34	14
19	0.00	32	26	0.00	36	28
20	0.45	27	17	0.35	29	19
21	0.00	26	19	0.02	27	19
22	0.00	28	9	0.01	29	17
23	0.70	24	20	0.35	27	22
24	0.00	26	−6	0.08	24	2
25	0.00	9	−13	0.00	11	4
26	0.00	22	−13	0.00	21	5
27	0.00	17	−11	0.00	19	7
28	0.00	26	−4	0.00	26	8
29	0.01	27	−4	0.01	28	14
30	0.03	30	11	0.01	31	14
31	0.05	34	23	0.13	38	23

Πηγές για δεδομένα/εικόνες που χρησιμοποιήσαμε στα παραδείγματα:

Iris Data Set: `http://en.wikipedia.org/wiki/Iris_flower_data_set`.
Εικόνα "Butterfly": `http://decsai.ugr.es/cvg/CG/images/base/35.gif`.
Εικόνα "Mandrill": `http://pngnq.sourceforge.net/testimages/mandrill.png`.

5.2 Προγράμματα στη Matlab

Κώδικας για το παράδειγμα "Διατροφικές Συνήθειες".

Αρχείο: habits.m

```
% load data
load('mydata.dat')
mydata

% get the size of mydata
[m,n] = size(mydata)
format shortg

% compute the mean value of each row
mv = mean(mydata,2)

% subtract the means (centering the data)
X = mydata - repmat(mv,1,n)

% create matrix Z
Z = X' / sqrt(n-1);

% compute the covariance matrix of Z
covz = Z'*Z;

% find the eigenvalues and eigenvectors of the covariance matrix
[vec,val] = eig(covz);
% extract diagonal of eigenvalues-matrix as vector
val = diag(val);

% sort the eigenvalues in decreasing order
[junk,I] = sort(-1*val);
val = val(I);
% sort eigenvectors in order of their eigenvalues
vec = vec(:,I);
```

```
% project the original data set
Y = vec'*X

% reconstruct the data (back to the original)
XX_A = vec*Y;
XX_A = XX_A + repmat(mv,1,n);


% plot eigenvalues
val= (val ./ sum(val)) * 100;
figure();
bar(val(1:4),0.4);
ylim([0 100])
hold all
plot(1:4, val(1:4), 'r','markersize',20)
plot(1:4, val(1:4), 'r.','markersize',12)
xlabel('Eigenvector number','fontsize',20)
ylabel('Eigenvalue','fontsize',20)

% Boxplot of the original dataset
figure();
names = {'Cheese';
    'Carcase meat';
    'Meat';
    'Fish';
    'Fats';
    'Sugar';
    'Potatoes';
    'Green veg.';
    'Fresh veg.';
    'Proc. pot.';
    'Proc. veg.';
    'Fruit';
    'Cereals';
    'Beverages';
    'Soft drinks';
    'Alc. drinks';
    'Confect.'};
em = {'';'';'';'';'';'';'';'';'';'';'';'';'';'';'';'';'';'';'';''};

boxplot(mydata', 'labels', em)
txt = text(1:17, -50*ones(1,17), names(: ),'fontsize',13);
set(txt,'HorizontalAlignment','right','VerticalAlignment','top',
```

```
    'Rotation',45);
ylabel('Consumption','fontsize',18)

% project on the 1st PC
figure
plot(Y(1,1),0,'b.',Y(1,2),0,'b.',Y(1,3),0,'b.',Y(1,4),0,'r.',
     'markersize',30)
xlabel('Principal Component 1','fontsize',20)
ylim([-0.4,0.4])
text(Y(1,1)-25,0.05,'England','fontsize',16)
text(Y(1,2)-20,-0.05,'Wales','fontsize',16)
text(Y(1,3)-30,-0.05,'Scotland','fontsize',16)
text(Y(1,4),-0.05,'N. Ireland','fontsize',16)

% project on the axes of the 1st and 2nd PCs
figure
plot(Y(1,1),Y(2,1),'b.',Y(1,2),Y(2,2),'b.',Y(1,3),Y(2,3),'b.',
     Y(1,4),Y(2,4),'r.','markersize',30)
xlabel('Principal Component 1','fontsize',20)
ylabel('Principal Component 2','fontsize',20)
ylim([-400,400])
text(Y(1,1)-25,Y(2,1)-25,'England','fontsize',16)
text(Y(1,2)-20,Y(2,2)-25,'Wales','fontsize',16)
text(Y(1,3)-30,Y(2,3)-25,'Scotland','fontsize',16)
text(Y(1,4),Y(2,4)-25,'N. Ireland','fontsize',16)
```

Κώδικας Matlab για το παράδειγμα "Iris Flower Data set".

Αρχείο: iris.m

```
% load iris data set (returns meas and species matrices)
load fisheriris

% reverse data for easier use
mydata = meas';

% species info matrix
sp = species;

% get the size of mydata
[m,n]  = size(mydata);
```

```
% compute the mean value of each row
means = mean(mydata,2);

% subtract the means
X = mydata - repmat(means,1,n);

% create matrix Z
Z = X' / sqrt(n-1);

% compute the covariance matrix of Z
covz = Z'*Z;

% compute SVD:
[U,S,V] = svd(covz);

% vectors
vec = V;

% values
val = diag(S);

% compute and plot eigenspectrum
values = (val ./ sum(val)) * 100
figure
hold all
bar(values, 0.4)
plot(values, 'r','markersize',20)
plot(values, 'r.','markersize',12)
xlabel('Eigenvector Number','fontsize',20)
ylabel('Singular value','fontsize',20)

% signals (projection onto PCs)
Y = vec'*X;

% reconstruction of the data
XX_A = vec*Y;
format shortg
XX_A = XX_A + repmat(means,1,n);

% Scatter plots of PCs  (histogram plots of the variances in the diagonal)
figure()
plotmatrix(Y')
```

```
% Projections onto principal components
figure
gscatter(Y(1,:), Y(2,:), species,'rgb','.s+');
xlabel('PC1');
ylabel('PC2');
N = size(meas,1);
title('Projection on principal components 1 and 2')

figure
gscatter(Y(1,:), Y(3,:), species,'rgb','.s+');
xlabel('PC1');
ylabel('PC3');
N = size(meas,1);
title('Projection on principal components 1 and 3')

figure
gscatter(Y(1,:), Y(4,:), species,'rgb','.s+');
xlabel('PC1');
ylabel('PC4');
N = size(meas,1);
title('Projection on principal components 1 and 4')

figure
gscatter(Y(2,:), Y(3,:), species,'rgb','.s+');
xlabel('PC2');
ylabel('PC3');
N = size(meas,1);
title('Projection on principal components 2 and 3')

figure
gscatter(Y(2,:), Y(4,:), species,'rgb','.s+');
xlabel('PC2');
ylabel('PC4');
N = size(meas,1);
title('Projection on principal components 2 and 4')

figure
gscatter(Y(3,:), Y(4,:), species,'rgb','.s+');
xlabel('PC3');
ylabel('PC4');
N = size(meas,1);
title('Projection on principal components 3 and 4')
```

Κώδικας Matlab για το παράδειγμα συμπίεσης της εικόνας "Butterfly".

Αρχείο: imgpca.m

```
% Load original image, convert to double
  [fly,map] = imread('butterfly.gif');
  fly = double(fly);

% Display original image
   image(fly), colormap(map);
   axis off, axis equal;

% Size of the data
  [m,n] = size(fly);

% Compute the mean values
  means = mean(fly,2);
% Subtract the mean values
  X = fly - repmat(means,1,n);
% Compute Z
  Z = X' / sqrt(n-1);

% Compute covariance matrix of Z
  covz = Z'*Z;

% Compute SVD of Z'Z
  [U,S,V] = svd(covz);
  variances = diag(S).*diag(S);

  sum_all = sum(variances);
  vars = (variances ./ sum_all ) * 100;
  figure();
  bar(vars(1:20));
  xlabel('Principal Components number','fontsize',18);
  ylabel('Variance percentage','fontsize',18);

  sum_all = sum(variances) ;
  pos=1;
  ratio = 0;
  figure();

% N is number of principal components for each subplot
  for N = [4 6 10 15 20 40 80 100 180 260 320 512]
```

```
    pcs = N;
    % keep only the first N PCs
    VV = V(:,1:pcs);

    % change of basis
    Y = VV' * X;

    % compression ratio
    ratio = 512 / pcs;

    % reconstruct the data
    XX = VV*Y;
    XX = XX + repmat(means,1,n);

    sum_N = sum(variances(1:N));

    % calculate the percentage
    P = sum_N/sum_all*100;

    % Subplot
    subplot(4,3,pos)
    image(XX), colormap(map);
    title([num2str(N) ' PCs ' num2str(P) '%, '
           num2str(roundn(ratio,-1)) ':1'])
    axis off, axis equal;
    pos=pos+1;

end
```

Κώδικας Matlab για τη συμπίεση της έγχρωμης εικόνας "Mandrill".

Αρχείο: colorimg.m

```
% Load original image, convert to double
[init_img,map] = imread('mandrill.png');

mR = double(init_img(:,:,1));
mG = double(init_img(:,:,2));
mB = double(init_img(:,:,3));

[m,n] = size(mR);

% mean values
meansR = mean(mR,2);
```

```
meansG = mean(mG,2);
meansB = mean(mB,2);

% subtract means and compute covariance matrices
XR = mR - repmat(meansR,1,n);
ZR = XR' / sqrt(n-1);
covzR = ZR'*ZR;

XG = mG - repmat(meansG,1,n);
ZG = XG' / sqrt(n-1);
covzG = ZG'*ZG;

XB = mB - repmat(meansB,1,n);
ZB = XB' / sqrt(n-1);
covzB = ZB'*ZB;

% Compute SVD of Z'Z and variances for each color
[UR,SR,VR] = svd(covzR);
variancesR = diag(SR).*diag(SR);

[UG,SG,VG] = svd(covzG);
variancesG = diag(SG).*diag(SG);

[UB,SB,VB] = svd(covzB);
variancesB = diag(SB).*diag(SB);

sum_allR = sum(variancesR);
sum_allG = sum(variancesG);
sum_allB = sum(variancesB);

pos=1;
d = 0.001;
figure();
for N = [4 6 10 15 20 40 80 100 180 260 320 512]
    pcs = N;

    % keep only the first N PCs
    VVR = VR(:,1:pcs);
    VVG = VG(:,1:pcs);
    VVB = VB(:,1:pcs);

    % change of basis
    YR = VVR' * XR;
```

```
        YG = VVG' * XG;
        YB = VVB' * XB;

        % compression ratio
        ratio = 512 / pcs;

        % re-construct the data
        XXR = VVR*YR;
        XXR = XXR + repmat(meansR,1,n);

        XXG = VVG*YG;
        XXG = XXG + repmat(meansG,1,n);

        XXB = VVB*YB;
        XXB = XXB + repmat(meansB,1,n);

        % calculate the percentage of the variance for R, G, B colors
        sum_N1 = sum(variancesR(1:N));
        PR = sum_N1/sum_allR*100;

        sum_N2 = sum(variancesG(1:N));
        PG = sum_N2/sum_allG*100;

        sum_N3 = sum(variancesB(1:N));
        PB = sum_N3/sum_allB*100;

        % plot the image using N PCs
        subplot(4,3,pos)
        im(:,:,1) = nmlz(XXR./255);
        im(:,:,2) = nmlz(XXG./255);
        im(:,:,3) = nmlz(XXB./255);
        image(im), colormap(map);
        title([num2str(N) ' PCs,  ' num2str(roundn(ratio,-1)) ':1  '
               num2str(round(PR/d) *d) '% '
               num2str(round(PG/d) *d) '% '
               num2str( round(PB/d) *d) '%' ])
        axis off, axis equal;
        pos=pos+1;

    end
```

Βοηθητική συνάρτηση που καλείται στον προηγούμενο κώδικα (για την κανονικοποίηση μη αναμενόμενων τιμών που ίσως προκύψουν από τη διαίρεση με το 255).

Αρχείο: nmlz.m

```
function out = nmlz(out)

    for i=1:size(out,1)
       for j=1:size(out,2)
          if out(i,j) < 0.0
             out(i,j) = 0.0;
          end
          if out(i,j) > 1.0
             out(i,j) = 1.0;
          end
       end
    end

end
```

Κώδικας Matlab για το παράδειγμα "Ithaca - Canandaigua".

Αρχείο: ithaca.m

```
%  load data from TABLE A.1
load 'table1.txt'
% 31x6 (excluding the 1st column that contains the dates)
data = table1(:,2:end);

[n,k]  = size(data); % size of data: n days, k variables

m1 = ones(n,n);
means =  1/n * m1 * data; % compute mean values
X = data - means; % mean values subtracted

% matrix 31x2 with the precipitation at K Locations
Ppt  = [ X(:,1) X(:,4) ];

% matrix 31x2 with the max Temp. at K Locations
Tmax = [ X(:,2) X(:,5) ];

% matrix 31x2 with the precipitation at K Locations
Tmin = [ X(:,3) X(:,6) ];

n = length(Tmin);
```

```
%----------------------------------------
% compute covariance matrix
CovTmin = 1/(n-1) * Tmin' * Tmin

% compute SVD of the covariance matrix
[U1,S1,V1] = svd(CovTmin)
Values1 = diag(S1)

% cumulative variance of the singular values
percval1 = (Values1 ./ sum(Values1)) * 100

% compute signals for Tmin timeseries
YTmin = Tmin * V1;

format short
% reconstruct the Tmin - data using only the 1st Principal Component
XXTmin = YTmin(:,1)*V1(:,1)';


%----------------------------------------
% compute covariance matrix
CovTmax = 1/(n-1) * Tmax' * Tmax

% compute SVD of the covariance matrix
[U2,S2,V2] = svd(CovTmax)
Values2 = diag(S2)

%cumulative variance of the singular values
percval2 = (Values2 ./ sum(Values2)) * 100

% compute signals for Tmax timeseries
YTmax = Tmax * V2;

% reconstruct the Tmax - data using only the 1st Principal Component
XXTmax = YTmax(:,1)*V2(:,1)';

%----------------------------------------
% compute covariance matrix
CovPpt = 1/(n-1) * Ppt' * Ppt

% compute SVD of the covariance matrix
[U3,S3,V3] = svd(CovPpt)
Values3 = diag(S3)
```

```
%cumulative variance of the singular values
percval3 = (Values3 ./ sum(Values3)) * 100

% compute signals for Ppt timeseries
YPpt = Ppt * V3;

% reconstruct the Ppt - data using only the 1st Principal Component
XXPpt = YPpt(:,1)*V3(:,1)';

% ---------- Figures ----------
z = 1:n;
% plot precipitation at Ithaca
figure
plot(z,XXPpt(:,1),'b',z,Ppt(:,1),'r:')
xlim([1,31])
legend('XXPpt','original data','Location','NorthEast')
ylim([-2.5,2.5])
title('Precipitation')
xlabel('January of 1987 at Ithaca', 'fontsize', 20)
ylabel('Precipitation (inches)','fontsize', 20)

% plot max temp. at Ithaca
figure
plot(z,XXTmax(:,1),'b',z,Tmax(:,1),'r:')
xlim([1,31])
legend('XXTmax','original data','Location','NorthEast')
ylim([-30,30])
title('Maximum Temperatures')
xlabel('January of 1987 at Ithaca', 'fontsize', 20)
ylabel('Maximum Temperature (degrees F)','fontsize', 20)

% plot min temp. at Ithaca
figure
plot(z,XXTmin(:,1),'b',z,Tmin(:,1),'r:')
xlim([1,31])
legend('XXTmin','original data','Location','NorthEast')
ylim([-30,30])
title('Minimum Temperatures')
xlabel('January of 1987 at Ithaca', 'fontsize', 20)
ylabel('Minimum Temperature (degrees F)','fontsize', 20)
```

```
% plot precipitation at Canandaigua
figure
plot(z,XXPpt(:,2),'b',z,Ppt(:,2),'r:')
xlim([1,31])
legend('XXPpt','original data','Location','NorthEast')
ylim([-2.5,2.5])
title('Precipitation')
xlabel('January of 1987 at Canandaigua', 'fontsize', 20)
ylabel('Precipitation (inches)','fontsize', 20)

% plot max temp. at Canandaigua
figure
plot(z,XXTmax(:,2),'b',z,Tmax(:,2),'r:')
xlim([1,31])
legend('XXTmax','original data','Location','NorthEast')
ylim([-30,30])
title('Maximum Temperatures')
xlabel('January of 1987 at Canandaigua', 'fontsize', 20)
ylabel('Maximum Temperature (degrees F)','fontsize', 20)

% plot min temp. at Canandaigua
figure
plot(z,XXTmin(:,2),'b',z,Tmin(:,2),'r:')
xlim([1,31])
legend('XXTmin','original data','Location','NorthEast')
ylim([-30,30])
title('Minimum Temperatures')
xlabel('January of 1987 at Canandaigua', 'fontsize', 20)
ylabel('Minimum Temperature (degrees F)','fontsize', 20)
```

Βιβλιογραφία

[1] Wilks, D.S., second edition, 2006. Statistical Methods in the Atmospheric Sciences, vol. 91, International Geophysics Series.

[2] Plaut, G., and Vautard R., Spells of Low-Frequency Oscillations and Weather Regimes in the Northern Hemisphere, J. Atm. Sci., vol. 51, 210-236, 1994.

[3] Annual Report on Food Expenditure, Consumption and Nutrient Intakes, National Food Survey report for 1998, DEFRA website.

[4] Jonathon Shlens, December, 2005, version 2, A Tutorial on Principal Component Analysis.

[5] Iris Data Set: `http://en.wikipedia.org/wiki/Iris_flower_data_set`.