Project 1: Principle Component Analysis

Μια από τις πιο σημαντικές παραγοντοποιήσεις πινάκων είναι η ***Singular Value Decomposition*** ή συντετμημένα SVD. Η SVD έχει πολλές χρήσιμες ιδιότητες, επιθυμητές σε πολλές εφαρμογές. Σε αυτή την εργασία, περιγράφεται μια συγκεκριμένη και δημοφιλής εφαρμογή της SVD: η ***Principle Components Analysis (PCA).***

Περιεχόμενα

1. Image SVD

Η PCA αξιοποιεί την καλύτερη χαμηλού βαθμού (low-rank) προσεγγιστική ιδιότητα της SVD. Έχοντας υπολογίσει την SVD ενός m επι n πίνακα A, αυτός μπορεί να αναπαρασταθεί σαν το άθροισμα των πινάκων πρώτου βαθμού:

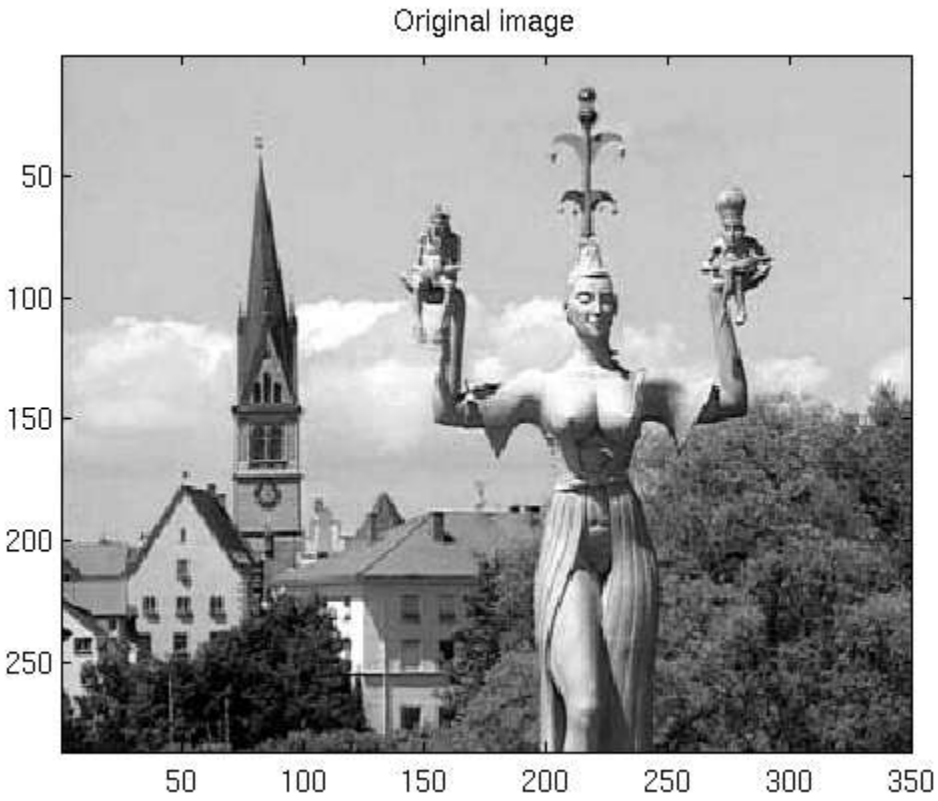
$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{j=1}^{n} \sigma_j u_j v_j^T,$$

όπου u και v είναι τα μοναδιαία διανύσματα και σ είναι οι μοναδιαίες τιμές. Ο προσεγγιστικός βαθμός r του πίνακα είναι επομένως το άθροισμα:

$$\mathbf{A} = \sum_{j=1}^{r} \sigma_j u_j v_j^T.$$

Ας εφαρμόσουμε αυτή την ιδέα σε μια απλή άσκηση, όπου αναζητούμε την προσέγγιση ενός πίνακα που αποτελεί μια ασπρόμαυρη εικόνα.

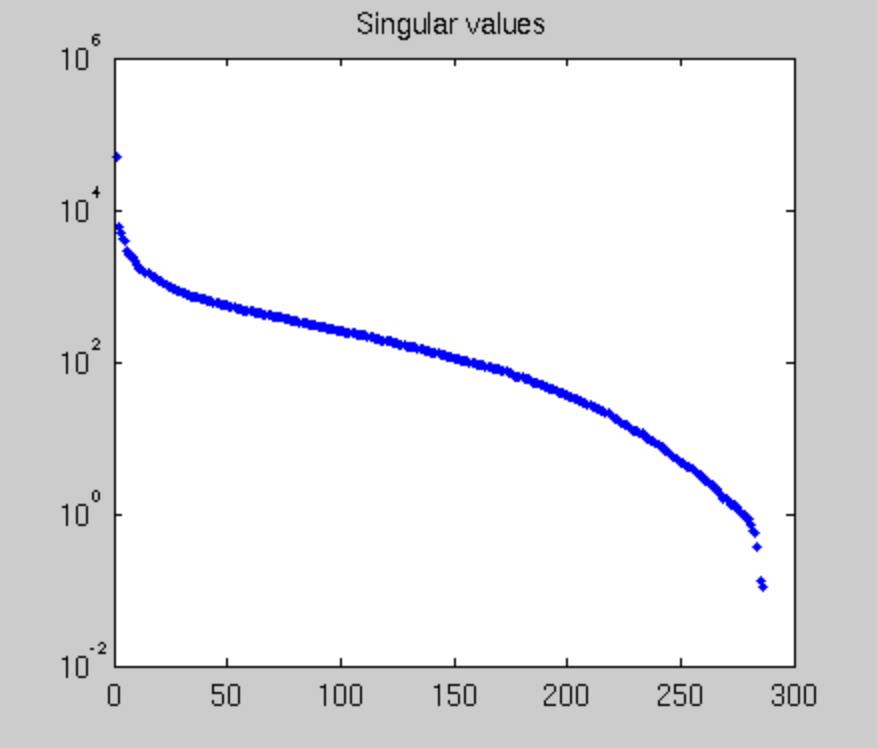
a) Φόρτωμα και εμφάνιση της αυθεντικής εικόνας



b) Υπολογισμός της SVD:

χρησιμοποιώντας την MATLAB [U, S, V] = svd(A, 0) συνάρτηση. Το μηδέν υποδηλώνει το "οικονομικό μέγεθος" της SVD. Στην δεύτερη εικόνα σχεδιάζονται οι μοναδιαίες τιμές σε λογαριθμική κλίμακα.

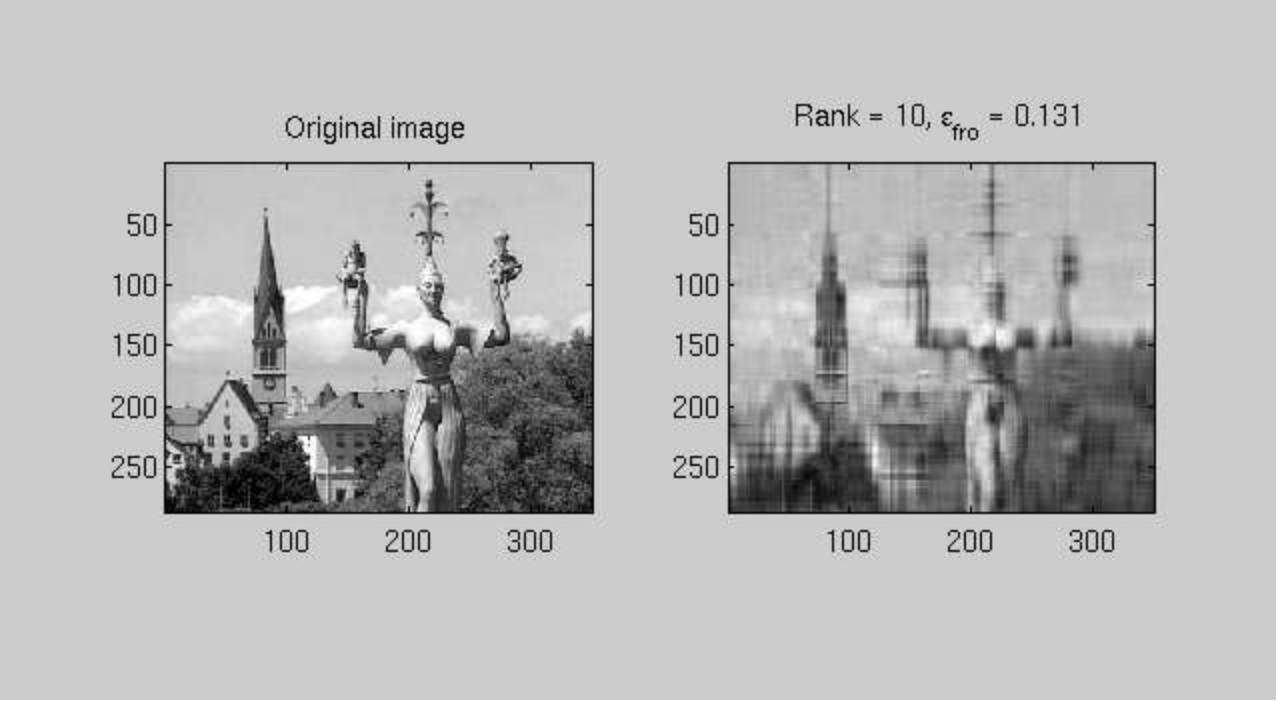
Ο ΚΩΔΙΚΑΣ ΣΑΣ ΕΔΩ:



c) Υπολογισμός της προσέγγισης χαμηλού βαθμού (low-rank):

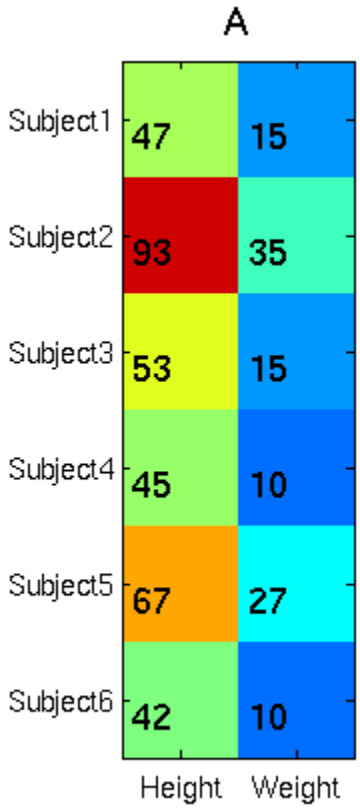
Υπολογίστε επίσης το σχετικό προσεγγιστικό σφάλμα (relative approximation error) w.r.t. Frobenius norm (χρησιμοποιείστε τη συνάρτηση `norm(A,'fro')` του MATLAB). Σχεδιασμός της αρχικής εικόνας στο δεξί μέρος της 3ης εικόνας και της βαθμού r (rank-r) προσέγγισης στο αριστερό. Απεικόνιση της αρχικής εικόνας και των τιμών σφαλμάτων στον τίτλο. Σε ποιό βαθμό και για ποιές τιμές σφαλμάτων η αρχική εικόνα και η προσέγγισή της παύουν να είναι ευδιάκριτες;

Ο ΚΩΔΙΚΑΣ ΣΑΣ ΕΔΩ:



2. PCA

Θα μάθουμε την PCA μέσω ενός απλού αλλά τυπικού παραδείγματος. Υποθέστε ότι μετρούμε το ύψος και βάρος έξι υποκειμένων και γράφουμε τα δεδομένα σε στήλες ένός πίνακα A



Κάθε ένα από τα υποκείμενα χαρακτηρίζεται από δυο στοιχεία (συντεταγμένες), ονομαστικά, ύψος, H και βάρος, W. Το ερώτημα στο οποίο απαντά η PCA είναι:

υπάρχει κάποιο συστατικό στοιχείο – ας το αποκαλέσουμε "μέγεθος", S = το οποίο προβλέπει γραμμικά (linearly) (προσεγγίζει) το ύψος και το βάρος ταυτόχρονα;

Ας επιχειρήσουμε αρχικά μια ποιοτική απάντηση στο ερώτημα. Θα επιθυμούσαμε να γνωρίζουμε αν υπάρχει κάποια ικανοποιητική προσέγγιση του ύψους και του βάρους στη μορφή:

$$H_{approx}(S) = a \cdot S, \qquad (1)$$

$$W_{approx}(S) = b \cdot S,$$

όπου α και b είναι κάποιες σταθερές. Για να το θέσουμε διαφορετικά, θα θέλαμε να μάθουμε αν ο λόγος W προς H είναι προσεγγιστική σταθερά:

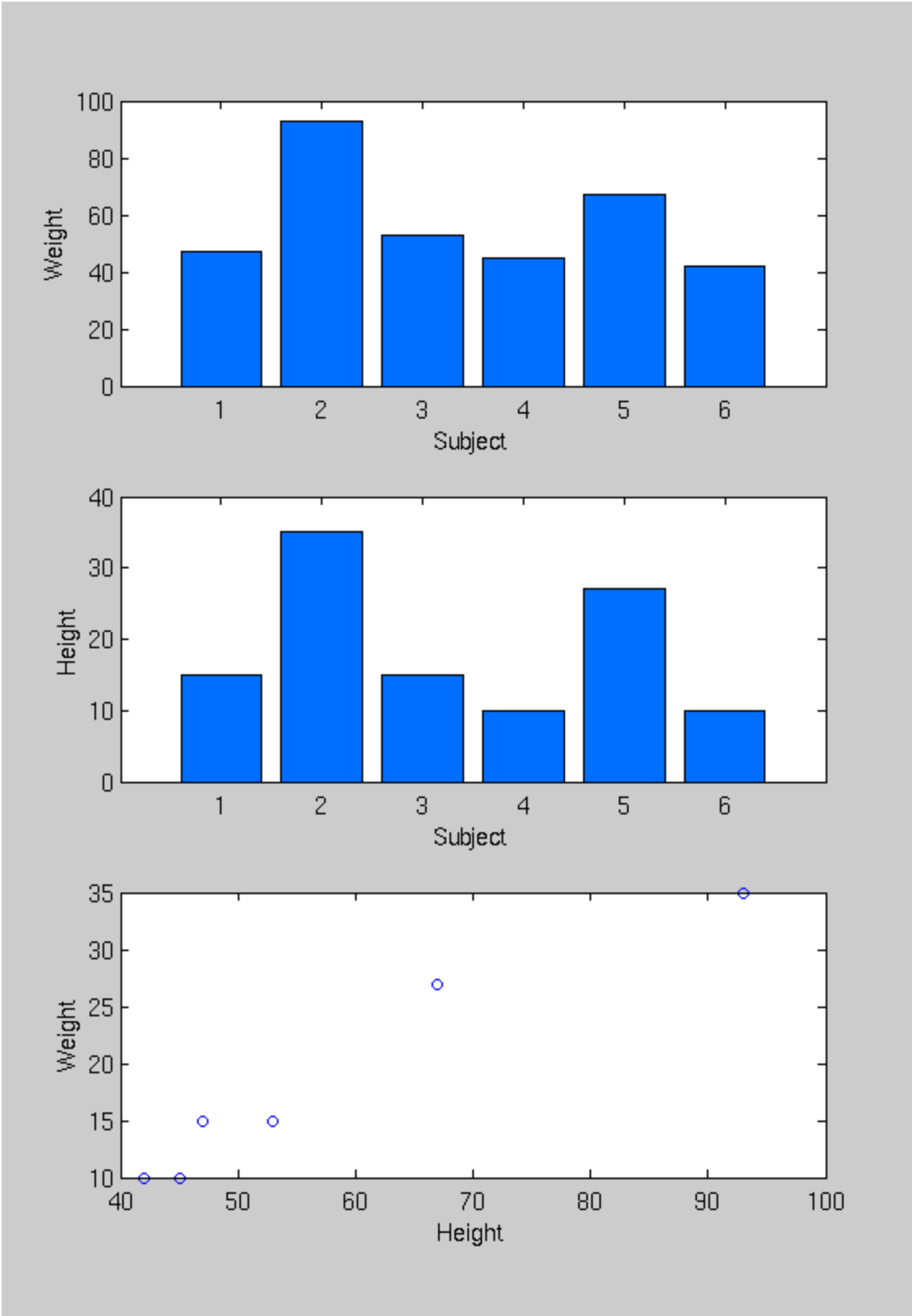
$$\frac{W}{H} \approx \frac{b}{a}$$

Ας εξετάσουμε τα δεδομένα για να διαπιστώσουμε αν κάτι τέτοιο ισχύει.

a) Διερεύνηση δεδομένων:

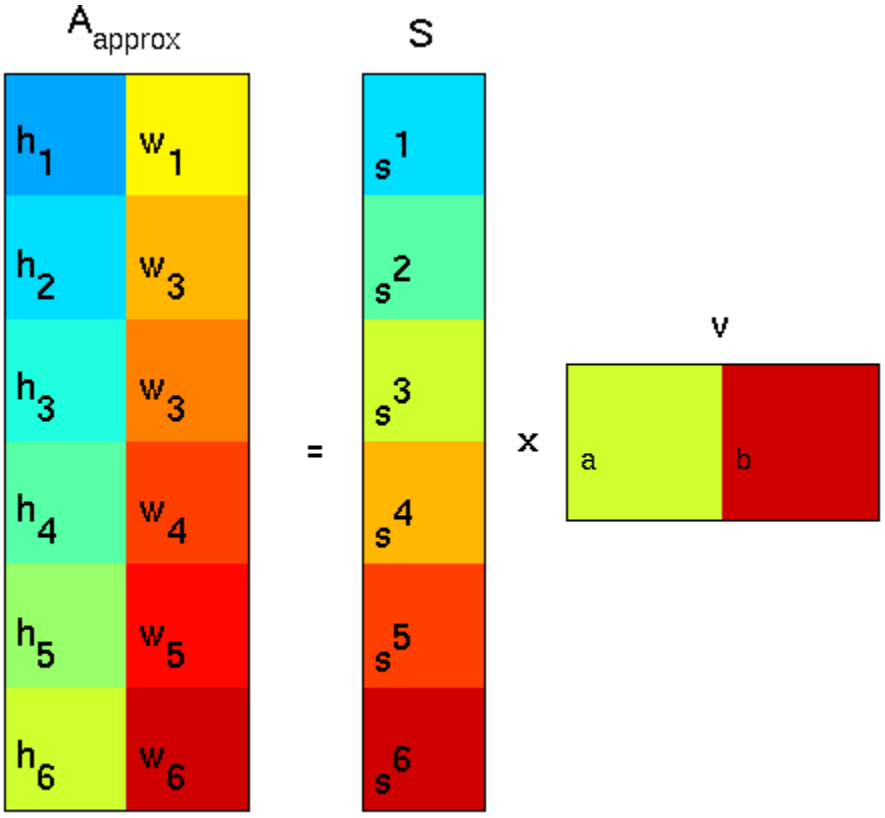
Σε μια εικόνα, αλλά σε διαφορετικά διαγράμματα, σχεδιάστε τις τιμές ύψους και βάρους και το λόγο βάρους προς ύψος.

Ο ΚΩΔΙΚΑΣ ΣΑΣ ΕΔΩ:


Weight
Subject
Height
Subject
Weight
Height

Από τα διαγράμματα γίνεται αντιληπτό ότι τα δεδομένα “ύψος” και “βάρος” είναι στενά συσχετισμένα. π.χ γραμμικώς εξαρτημένα. Επομένως, θα πρέπει να υπάρχει ένα απλό στοιχείο πρόβλεψης όπως το “Μέγεθος”.

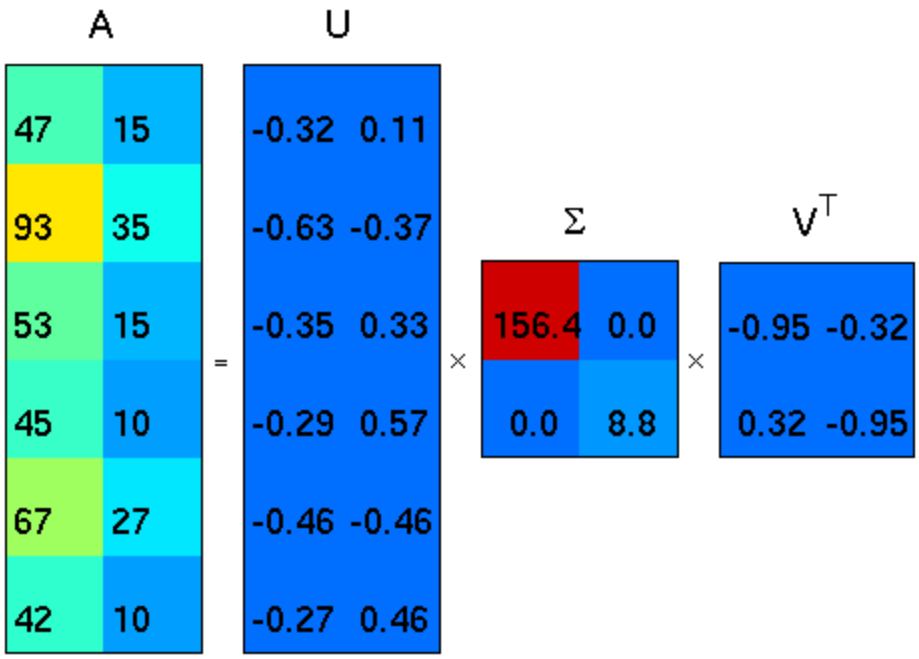
Τώρα το ερώτημα γίνεται ποσοτικό: πώς θα υπολογιστεί αυτό το στοιχείο. Ως συνήθως υπάρχουν περισσότεροι από έναν τρόπο για να επιτευχθεί κάτι τέτοιο. Μιά περίπτωση είναι να χρησιμοποιηθεί η SVD για να το υπολογίσει. Ας γράψουμε πάλι την εξίσωση (1) σε μορφή πίνακα:



Έτσι, θα θέλαμε να βρούμε έναν διάνυσμα στήλης s και ένα διάνυσμα γραμμής v=[a,b] το οποίο θα παράγει την βέλτιστη δυνατή προσέγγιση πρώτου βαθμού του A. Αυτό ακριβώς κάνει η SVD! Ας το υπολογίσουμε.

b) Υπολογισμός της SVD:

Ο ΚΩΔΙΚΑΣ ΣΑΣ ΕΔΩ:



Παρατηρήστε ότι η πρώτη μοναδιαία τιμή είναι πολύ μεγαλύτερη από την δεύτερη, υποδεικνύοντας ότι η πρώτου βαθμού προσέγγιση είναι αρκετά ακριβής.

c) Το πρωταρχικό στοιχείο:

Συγκρίνοντας τις δύο παραπάνω ισότητες, είναι φανερό ότι το στοιχείο “Μεγεθος” είναι υπο κλίμακα, το πρώτο αριστερό μοναδιαίο διάνυσμα *u*.

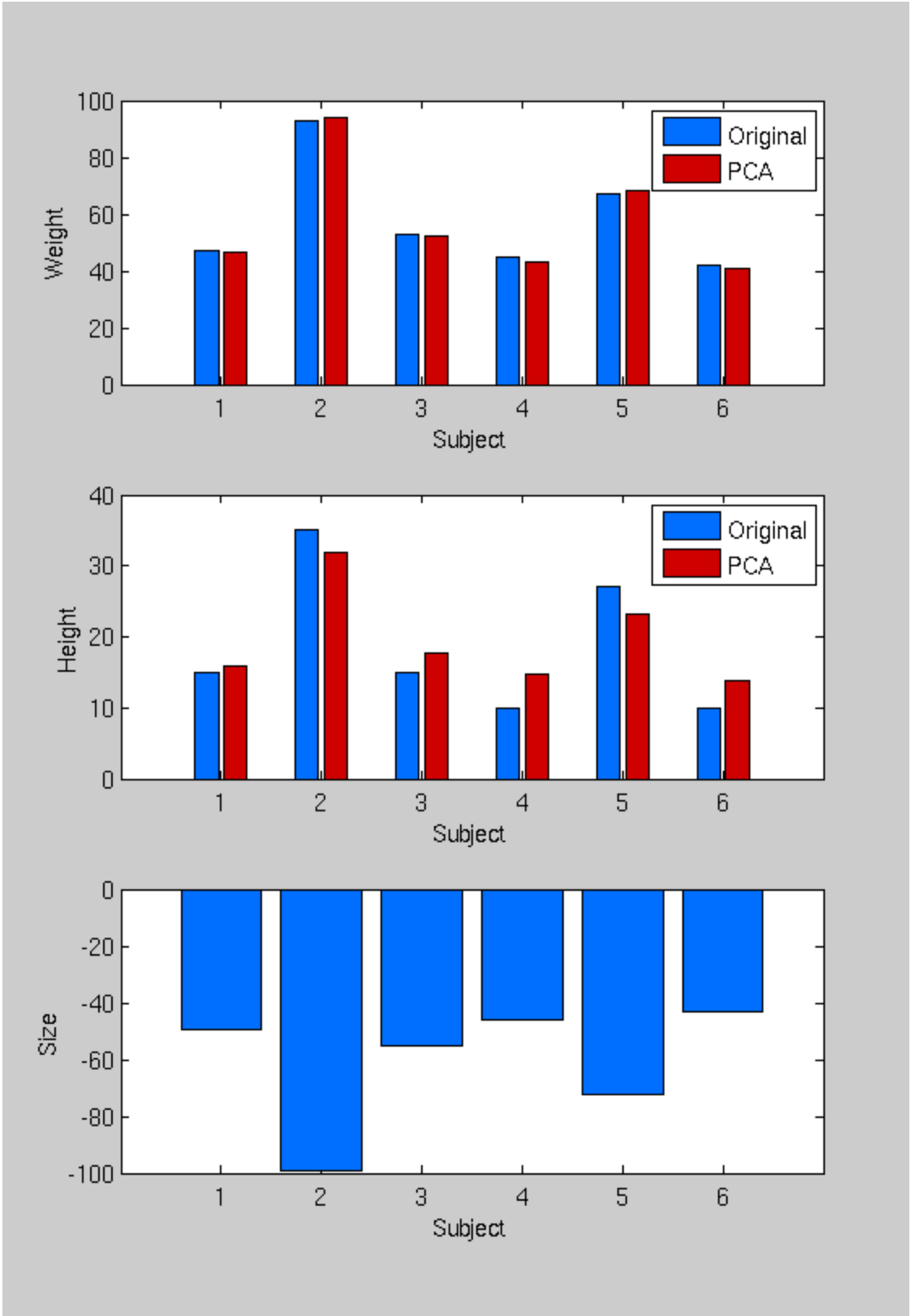
$$S = u_1 * \sigma_1$$

και οι αντίστοιχοι συντελεστές α και b είναι τα στοιχεία του πρώτου δεξιού μοναδιαίου διανύσματος v.

d) Η πρώτου βαθμού προσέγγιση:

Δεδομένου του “Μεγέθους” και διανυσμάτων συντελεστών, μπορούμε να υπολογίσουμε τα προσεγγιστικά δεδομένα “βάρος” και “ύψος”. Υπολογίστε τα και σχεδιάστε (σε μια σχέδιο αλλά διαφορετικά διαγράμματα) τις προσεγγιστικές ποσότητες μαζί με τα αρχικά δεδομένα. Προσθέστε μια λεζάντα στα διαγράμματά σας. Σε ένα ξεχωριστό διάγραμμα σχεδιάστε επίσης της τιμές “μεγέθους”.

Ο ΚΩΔΙΚΑΣ ΣΑΣ ΕΔΩ:


Weight
100
80
60
40
20
0
1
2
3
4
5
6
Subject
Original
PCA
Height
40
30
20
10
0
1
2
3
4
5
6
Subject
Original
PCA
Size
0
-20
-40
-60
-80
-100
1
2
3
4
5
6
Subject