ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

PRINCIPLE COMPONENT ANALYSIS (PCA)

Η pca είναι μια μεθοδος ανάλυσης πολυμεταβλητων δεδομένων.

Βασίζεται στην μελετη των ιδιοτιμών κ των ιδιοδιανυσματων του πίνακα διακύμανσης/συνδιακύμανσης.

Α) Προετοιμασία:

Ξεκινώντας, έχουμε τον πίνακα με τα Ν δεδομένα (σε στήλες). Υπολογίζουμε πρώτα το διάνυσμα (κ × 1)του μέσου όρου Μ και το αφαιρούμε απο καθε στήλη, και έτσι παίρνουμε ενα νεο πίνακα (τον Β,κ ×Ν) στον οποίο τα δεδομένα έχουν μεση τιμή 0. (mean- deviation form)

Στη συνέχεια υπολογίζουμε τον πίνακα διακύμανσης/συνδιακύμανσης Σ.

$$\Sigma = \frac{1}{N-1} BB^T$$

Στη διαφώνιο είναι η διακύμανση την δεδομένων κ έτσι το άθροισμα των διαγώνιον στοιχείων (ιχνος του Σ, tr(Σ)) μας δίνει την ολική διακύμανση (total variance)

2) PCA:

Υπολογίζουμε τις ιδιοτιμές κ τα αντίστοιχα ιδιοδιανύσματα (Κανονικοποιημένα) του Σ.

Η μεγαλύτερη ιδιοτιμή είναι ο πρώτος παράγοντας, ο πιο σημαντικός.

Το αντίστοιχο ιδιοδιάνυσμα μας δίνει το πρώτο principal component κλπ. Το πρώτο κύριο συστατικό (principal component) διατηρεί περισσότερες πληροφορίες δεδομένων σε σύγκρίση με το δεύτερο το οποίο δεν διατηρεί πληροφορίες οι οποίες έχουν εισέλθει νωρίτερα (στο πρώτο συστατικό). Τα principal components δεν συσχετίζονται.

Αυτα μπαίνουν σε στήλες και παίρνουμε τον πίνακα Ρ. Θέτουμε Υ = P^T Χ και τα νέα δεδομένα Υ είναι ασυσχέτιστα, οποτε ο νέος πίνακας Σ’ θα είναι διαγώνιος.

Παράδειγμα:

Εστα τα 3-διαστατα δεδομένα : $X_{ι,}$ ι = 1,2,3,4.

$$Xi = \begin{bmatrix}1\\2\\1\end{bmatrix}, \begin{bmatrix}4\\2\\13\end{bmatrix}, \begin{bmatrix}7\\8\\1\end{bmatrix}, \begin{bmatrix}8\\4\\5\end{bmatrix}.$$

Ο μέσος όρος τους είναι το διάνυσμα Μ = $\begin{bmatrix}5\\4\\5\end{bmatrix}$

Τα δεδομένα σε mean- deviation form γίνονται:

$\widehat{X_i} = \begin{bmatrix}-4\\-2\\-4\end{bmatrix}, \begin{bmatrix}-1\\-2\\8\end{bmatrix}, \begin{bmatrix}2\\4\\-4\end{bmatrix}, \begin{bmatrix}3\\0\\0\end{bmatrix}$ και βαζοντας τα σε στήλες παίρνουμε τον πίνακα Β.

Ο Variance- Covariance matrix Σ = $\frac{1}{3}BB^T$ είναι τελικά ο

$\Sigma = \begin{bmatrix}10 & 6 & 0\\6 & 8 & -8\\0 & -8 & 32\end{bmatrix}$ και η ολική διακύμανση είναι 50.

Οπως βλέπουμε, η πρώτη συνιστώσα έχει το 10/50 = 20% της πληροφορίας, η δευτερη το 8/50 = 16% και η τρίτη το 32/50 = 64%.

Οι ιδιοτιμές του Σ είναι λ_1= 34,5 λ_2 = 13,8 και λ_3 = 1,6.

Προφανως η πρώτη ιδιοτιμή μας δίνει την κύρια συνιστώσα, και το αντίστοιχο ιδιοδιάνυσμα κανονικοποιημένο είναι το $u = \begin{bmatrix}0{,}074\\0.303\\-0.95\end{bmatrix}$

Ετσι λοιπον ενας δείκτης για τα συγκεκριμένα δεδομένα είναι ο γραμμικος συνδυασμός είναι ο **$y = 0.074x_1 + 0.303x_2 - 0.95x_3$** .

Ο νέος πίνακας Variance- Covariance θα είναι διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές, επομένως τώρα η πληροφορία θα μοιραστεί ως εξής:

34,5 /50 = 69%, 13,8/50 = 27,6% και 1,6/50 = 3,2%.