Άλλη μία εφαρμογή των προηγουμένων στη Στατιστική:

Ανάλυση κύριων συνιστωσών (Principal Components Analysis: PCA)

Από: Bishop (2006): Pattern recognition and machine Learning, Chapter 12:

- Έστω διαθέτουμε N παρατηρήσεις $X_1, \dots, X_N$, που η καθεμία έχει πολλές (n) διαστάσεις: $X_i \in \mathbb{R}^n$

  Αν για παράδειγμα το X_i παριστάνει την i εικόνα και κάθε συνιστώσα X_{ij} του X_i ένα pixel αυτής της εικόνας, με $X_{ij}=0$ για άσπρο και $X_{ij}=1$ για μαύρο, τότε αν η εικόνα έχει 100×100 pixel, θα έχουμε $n=10.000$ συνιστώσες σε κάθε X_i.

- Σε κάποιες εφαρμογές τα N δεδομένα του $\mathbb{R}^n$, $X_1, \dots, X_N$, βρίσκονται "κοντά" σε έναν υπόχωρο χαμηλότερης διάστασης (k) του $\mathbb{R}^n$: τον κύριο υπόχωρο, U_k.

  Απλό παράδειγμα $n=2$, $k=1$: Figure 12.2

- Τότε η "περισσότερη πληροφορία" των δεδομένων μας μπορεί να περιγράφεται "επαρκώς" από την προβολή των X_i στον κύριο υπόχωρο U_k, $\tilde{X}_i$. Οι αποκλίσεις $X_i - \tilde{X}_i$ θα μπορούσαν (για κατάλληλη επιλογή του k) να είναι "θόρυβος" μικρής σημασίας.

  – Κύριος υπόχωρος, διάστασης k, θα είναι εκείνος που αναπαριστά την περισσότερη διακύμανση των X_i, δηλαδή εκείνος που μεγιστοποιεί την διακύμανση των $\tilde{X}_i$.

  – Ισοδύναμα (θα δούμε): εκείνος που ελαχιστοποιεί τις αποκλίσεις $\sum \|X_i - \tilde{X}_i\|^2$.

Η PCA χρησιμοποιείται για

- μείωση της διάστασης που είναι αναγκαία για να περιγράψω τα δεδομένα (dimensionality reduction)
- συμπίεση δεδομένων (data compression)
- εξαγωγή χαρακτηριστικών (feature extraction)
- οπτικοποίηση δεδομένων (data visualization)

Για τον χαρακτηρισμό και την εύρεση των κύριων συνιστωσών υπάρχουν δύο δυνατές ισοδύναμες διατυπώσεις.

- κατ' αρχάς ας υποθέσουμε ότι τα δεδομένα μας έχουν μηδενικό δειγματικό μέσο $\bar{X} = \frac{1}{N}\sum X_i = \begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}$. Αν όχι θα αρκούσε να αφαιρέσουμε τό $\bar{X}$ από τα δεδομένα μας, οπότε τα "νέα" δεδομένα θα είχαν μηδενικό δειγματικό μέσο.

* Πρώτη διατύπωση: Βρες υπόχωρο U_k, για σταθερή και δεδομένη διάσταση k, που μεγιστοποιεί τη διακύμανση των προβολών $\tilde{X}_i = P_{U_k} X_i$ των δεδομένων.

- Ας υποθέσουμε $k=1$. και $U_k = \mathrm{Span}(q_1)$, $\|q_1\|=1$ για κατάλληλο q_1 : ~~[illegible]~~ Θα βρούμε ποιο πρέπει να είναι τό q_1.

  Έστω $\tilde{X}_i = P_{q_1} X_i = (q_1^T X_i) \cdot q_1$. ~~[illegible]~~

  Τότε η "δειγματική διακύμανση" των $\tilde{X}_i$ ορίζεται ως το $\frac{1}{N}\sum_{i=1}^{N}\|\tilde{X}_i\|^2$, το οποίο ισούται με

$$\frac{1}{N}\sum_{i=1}^{N}\|\tilde{X}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i^T q_1)^2 = \frac{1}{N}\sum_{i=1}^{n} q_1^T X_i X_i^T q_1$$

$$= q_1^T S q_1 \quad \text{με } S := \frac{1}{N}\sum X_i X_i^T$$

Άρα ψάχνουμε q_1, με $\|q_1\|=1$ πού νά μεγιστοποιεί τήν $q_1^T S q_1$. Η παράγραφος για το πηλίκο του Rayleigh (βλέπε Strang ή Σημειώσεις) δίνει τήν απάντηση: Το μέγιστο επιτυγχάνεται όταν q_1 είναι το ιδιοδιάνυσμα που αντιστοιχεί στή μέγιστη ιδιοτιμή λ_1, του S. Τότε $q_1^T S q_1 = \lambda_1$

- Αν $k=2$ ψάχνουμε U_2 και έστω τόν περιγράφουμε ως $U_2 = \mathrm{Span}(q_1, q_2)$ με q_1, q_2 ορθοκανονικά. Τότε με $\tilde{X}_i = P_{U_2} X_i$

$$\max_{U_2} \frac{1}{N} \sum \|\tilde{X}_i\|^2 = \max_{q_1, q_2 \text{ ορθ.}} \left(\frac{1}{N} \sum \|(q_1^T X_i) q_1 + (q_2^T X_i) q_2\|^2 \right)$$

$$= \max_{\substack{q_1, q_2 \\ \text{ορθ.}}} \left\{ \frac{1}{N} \sum \|(q_1^T X_i) q_1\|^2 + \frac{1}{N} \sum \|(q_2^T X_i) q_2\|^2 \right\}$$

$$= \max_{q_1} q_1^T S q_1 + \max_{q_2 \perp q_1} q_2^T S q_2.$$

Τό μέγιστο αυτό επιτυγχάνεται για q_1, q_2 νά είναι τά ιδιοδιανύσματα που αντιστοιχούν στίς δύο μεγαλύτερες ιδιοτιμές του S. και ισούται με $\lambda_1 + \lambda_2$.

- Γενικά, $k \geq 2$: Ο U_k, ο υπόχωρος που μεγιστοποιεί τή δειγματική διακύμανση των προβολών των X_i στον U_k θα είναι ο $U_k = \mathrm{Span}(q_1, \dots, q_k)$ όπου $q_1, \dots, q_k$ θά είναι τα ιδιοδιανύσματα των k μεγαλύτερων ιδιοτιμών του S $\lambda_1 \geq \dots \geq \lambda_k$.

Τότε $\tilde{X}_i = (q_1^T X_i) \cdot q_1 + \dots + (q_k^T X_i) q_k$

$\boxed{\text{* Δεύτερη διατύπωση: Βρες υπόχωρο } U_k \text{, για σταθερή και δοσμένη διάσταση } k \text{, που ελαχιστοποιεί την τετραγωνική απόσταση ανάμεσα στα δεδομένα } x_i \text{ και στην προβολή τους } \tilde{x}_i = P_{U_k} x_i \text{ δηλαδή την } J = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \tilde{x}_i\|^2.}$

- Έστω $p_1, \dots, p_n$ μια ορθοκανονική βάση του $\mathbb{R}^n$

Τότε $$x_i = \sum_{j=1}^{n} (x_i^T p_j) \cdot p_j$$

- Αν $V_k = \operatorname{Span}(p_1, \dots, p_k)$ τότε το σημείο του V_k που απέχει ελάχιστα από το x_i είναι το $\tilde{x}_i = P_{V_k} x_i$ και τότε $\tilde{x}_i = \sum_{j=1}^{k} (x_i^T p_j) p_j$ και η απόσταση

$$x_i - \tilde{x}_i = \sum_{j=k+1}^{n} (x_i^T p_j) \cdot p_j .$$

- Τότε θα έχουμε $$J = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \tilde{x}_i\|^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=k+1}^{n} (p_j^T x_i)^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=k+1}^{n} p_j^T x_i x_i^T p_j$$

$$= \sum_{j=k+1}^{n} p_j^T S p_j , \quad \text{όπου } S = \frac{1}{N} \sum x_i x_i^T$$

Άρα το θέμα μας είναι να βρούμε μια βάση $p_1, \dots, p_n$ του $\mathbb{R}^n$ PCA5

έτσι ώστε για συγκεκριμένο k να ελαχιστοποιείται η

$$J = \sum_{j=k+1}^{n} p_j^T S p_j$$

• Παράδειγμα: $n=2$, $k=1$. : Ελαχιστοποιούμε την $p^T S p$

Απάντηση (πηλίκο του Rayleigh) για $p = q_2$ το ιδιοδιάνυσμα με μικρότερης ιδιοτιμή λ_2. Τότε θα έχουμε και $J = q_2^T S q_2 = \lambda_2$

Άρα $p_1 = q_1$ και $U_1 = \mathrm{Span}(q_1)$, όπως έχουμε βρει και με τη προηγούμενη διατύπωση.

• <u>Γενικώς</u> $(k>1)$: Διαλέξτε $p_{k+1}, \dots, p_n$ να είναι τα ιδιοδιανύσματα ~~του~~ S που αντιστοιχούν στις μικρότερες ιδιοτιμές $\lambda_{k+1} \geq \dots \geq \lambda_n$. Τότε $J = \sum_{j=k+1}^{n} \lambda_j$

και $U_k = \mathrm{Span}(q_1, \dots, q_k)$ τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές $\lambda_1, \dots, \lambda_n$.

<u>Παρατήρηση</u>. Αν έχω $k=n$ δεν έχω μείωση διάστασης αλλά $\tilde{x}_i = x_i$

• Σε πολλές εφαρμογές φτάνουν "αρκετά λίγες" ιδιοτιμές $k<n$ για να αναπαραστήσουν αρκετά καλά τα δεδομένα

$\longrightarrow$ Παραδείγματα.