Ανάλυση Ανεξάρτητων, Θετικών Συνιστώσων

Αρχές και Μέθοδοι Μηχανικής Μάθησης

Σημειώσεις στις σημειώσεις του κ. Διαμαντάρα

Ανεξάρτητοι Παράγοντες

Από το PCA γνωρίζουμε ότι οι παράγοντες y_i είναι **Στατιστικά Ασυσχέτιστοι** μεταξύ τους. Δηλαδή ισχύει:

$$\boxed{E\{y_i y_j\} = 0, i \neq j} \tag{1}$$

Αλλά,

Σε κάποιες εφαρμογές είναι σωστότερο να υποθέσουμε ότι οι **κρυφοί παράγοντες** είναι **στατιστικά ανεξάρτητοι**, δηλαδή έχουμε:

$$\boxed{p(y_i, y_j) = p(y_i) * p(y_j)} \tag{2}$$

Και αυτό γιατί η ανεξαρτησία είναι **πιο ισχυρή** από την έλλειψη συσχέτισης:

$$\boxed{y_i, y_j \text{Ανεξάρτητοι} \Rightarrow y_i, y_j \text{Ασυσχέτιστοι}} \tag{3}$$

Δηλαδή οι ασυσχέτιστοι **δεν σημαίνει** απαραίτητα ότι είναι και ανεξάρτητοι.

$$\boxed{y_i, y_j \text{Ανεξάρτητοι} \not\Leftarrow y_i, y_j \text{Ασυσχέτιστοι}} \tag{4}$$

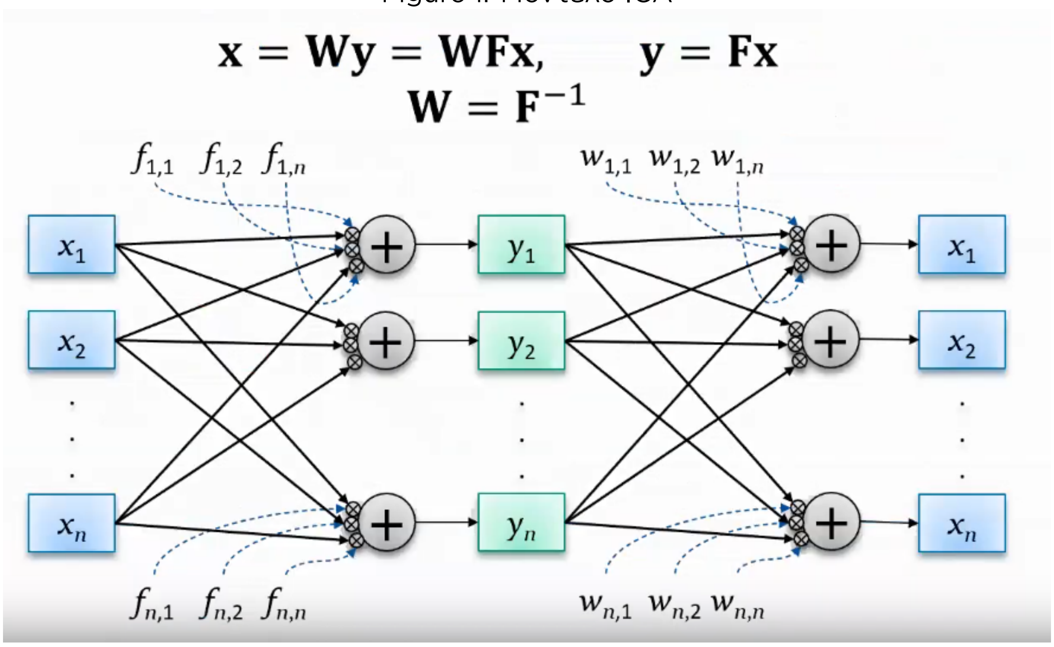
Πως μπορεί να χρησιμοποιηθεί

Έστω ότι υπάρχουν 2 παρουσιαστές σε μια σκηνή με 1 μικρόφωνο ο καθένας. Δυστυχώς, επειδή είναι πολύ κοντά, **οι φωνές και των 2 απορροφούνται και από τα 2 μικρόφωνα!** Με το ICA, θα μπορούσαμε αφού έχουμε τα μπλεγμένα δεδομένα, να τα **ξεμπλέξουμε και να ξεχωρίσουμε τις 2 φωνές.**

Ανάλυση Ανεξαρτήτων Συνιστώσων

- <u>Δεδομένα</u>:
  - Παρατηρήσεις: $\boldsymbol{x}(1), ..., \boldsymbol{x}(N) \in \mathbb{R}^n$
  - Δεν χρησιμοποιούνται στόχοι t (Χωρίς Επίβλεψη)
  - Πλήθος παραγόντων n
- <u>Πρόβλημα</u>: Βρες τον πίνακα $\boldsymbol{F} \in \mathbb{R}^n * n$ έτσι ώστε οι παράγοντες $y_i = \boldsymbol{f}_i^T \boldsymbol{x}$ να είναι στατιστικά ανεξάρτητοι μεταξύ τους.
  - $\boldsymbol{f}_i^T$ οι γραμμές του $\boldsymbol{F}$
- <u>Παρατήρηση</u>: Αν θέσουμε
  - $\boldsymbol{W} = \boldsymbol{F}^{-1}$
  - $\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x}$
  - τότε $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y} = \boldsymbol{w_1} y_1 + ... + \boldsymbol{w}_n y_n$

Figure 1: Μοντέλο ICA



Συσσωρεύτριες (cumulants)

Οι Συσσωρεύτριες είναι συναρτήσεις αντίθεσης.

Ορισμός 0.1. Μια συσσωρεύτρια k τάξης μιας τυχαίας μεταβλητής x ορίζεται ως

$$\boxed{c_k(x) = (-j)^k \frac{d^k \phi_x(\omega)}{d\omega^k}} \tag{5}$$

Όπου $\phi_x(\omega) = lnE\{e^{j\omega x}\}$

Και πρακτικά έχουμε:

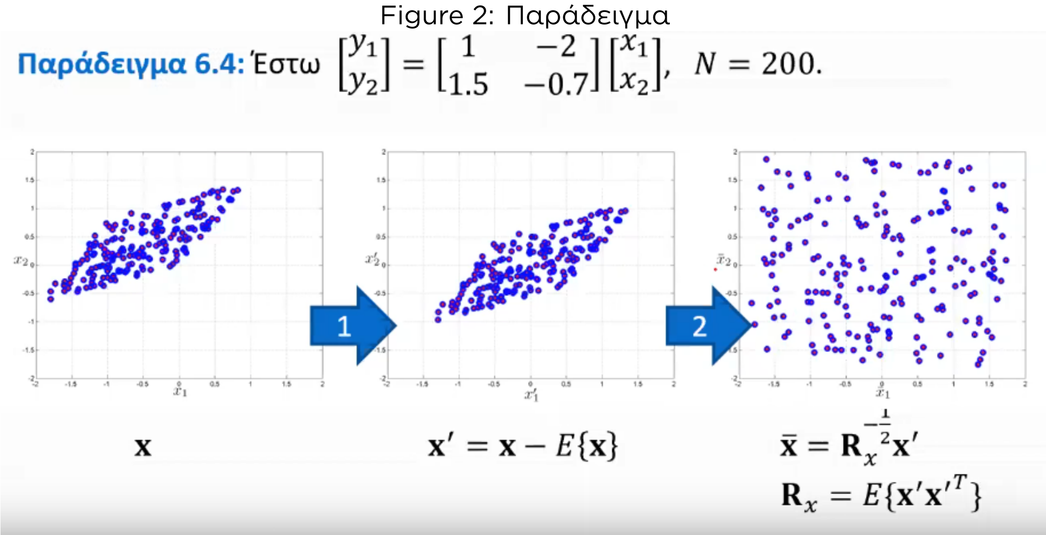
- $c_1 = E\{x\}$
- $c_2 = E\{x^2\}$ (διακύμανση)
- $c_3 = E\{x^3\}$
- $c_4 = E\{x^4\} - 3E\{x^2\}$ (κύρτωση, αρκετά χρήσιμη)
- ...κλπ

Προεπεξεργασία Δεδομένων

1. <u>Αφαίρεση μέσου όρου</u>: $\boldsymbol{x} \Rightarrow \boldsymbol{x'} : E\{\boldsymbol{x'}\} = 0$
2. <u>Λεύκανση</u>: $\boldsymbol{x'} \Rightarrow \bar{\boldsymbol{x}} : E\{\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\boldsymbol{T}}\} = \boldsymbol{I}$

Ορισμός 0.2. Λεύκανση ονομάζεται η τετραγωνοποίηση διαγράμματος μιας ομάδας δεδομένων.

Figure 2: Παράδειγμα



Εξαγωγή ενός παράγοντα

Έστω για κάποιο $\boldsymbol{f}$: $z(k) = \boldsymbol{f^T x(k)}$

θα έχουμε $z(k) = \boldsymbol{f^T W y}(k)$

όπου $\boldsymbol{f^T} = \boldsymbol{v^T} \Rightarrow z(k) = \boldsymbol{v^T y(k)}$

Θεώρημα 0.1. Αν οι παράγοντες δεν είναι Γκαουσσανές τυχαίες μεταβλητές, τότε η μεγιστοποίηση της κύρτωσης $J(\boldsymbol{f}) = |c_4(z)|$ υπό τον περιορισμό $c_2(z) = 1$ επιτυγχάνεται για

$\boldsymbol{v}$ = [0 0 … 0 α 0 … 0]

Περιορισμοί

- Χάνεται η σειρά των παραγόντων, δηλαδή οι εκτιμώμενοι παράγοντες y_i μπορεί να είναι ανακατωμένοι.
- Οι παράγοντες μπορεί να εξαχθούν με αυθαίρετη κλιμάκωση. Πιθανή **αλλαγή προσήμου**.
- Κάθε ανάλυση ICA σε ίδια δεδομένα μπορεί να εξάγει παράγοντες με διαφορετική σειρά και κλιμάκωση.

PCA εναντίον ICA

Ζουμί	PCA	ICA
Πλήθος παραγόντων(συνιστώσων)	m (≤ διάσταση διανύσματος παρατήρησης)	n (= διάσταση διανύσματος παρατήρησης)
Σχέση παραγόντων	Ασυσχέτιστοι	Ανεξάρτητοι
Συνάρτηση Κόστους	Μέσο τετραγωνικό σφάλμα	Συναρτήσεις αντίθεσεις