

Μείωση διάστασης

Δημήτρης Κουγιουμτζής

10 Δεκεμβρίου 2019

Εισαγωγή

Τεχνολογία αισθητήρων, αποθήκευσης, δικτύωσης →
καταγραφή δεδομένων από πολλές μεταβλητές. Τις θέλουμε
όλες; **Η κατάρα της διαστατικότητας**

Εισαγωγή

Τεχνολογία αισθητήρων, αποθήκευσης, δικτύωσης → καταγραφή δεδομένων από πολλές μεταβλητές. Τις θέλουμε όλες; **Η κατάρα της διαστατικότητας**

Μείωση διάστασης είναι η διαδικασία που από ένα υψηλο-διάστατο σύνολο δεδομένων δημιουργεί ένα άλλο μικρότερης διάστασης που διατηρεί το μεγαλύτερο μέρος της πληροφορίας των αρχικών δεδομένων.

Τεχνολογία αισθητήρων, αποθήκευσης, δικτύωσης → καταγραφή δεδομένων από πολλές μεταβλητές. Τις θέλουμε όλες; **Η κατάρα της διαστατικότητας**

Μείωση διάστασης είναι η διαδικασία που από ένα υψηλο-διάστατο σύνολο δεδομένων δημιουργεί ένα άλλο μικρότερης διάστασης που διατηρεί το μεγαλύτερο μέρος της πληροφορίας των αρχικών δεδομένων.

Δύο βασικές κατηγορίες μεθόδων μείωσης διάστασης:

1. Επιλογή υποσυνόλου μεταβλητών (προβολή στους φυσικούς άξονες).
2. Μετασχηματισμός των μεταβλητών των αρχικών δεδομένων σε νέες λιγότερες μεταβλητές (μετασχηματισμός ή/και προβολή).

Εισαγωγή

Τεχνολογία αισθητήρων, αποθήκευσης, δικτύωσης → καταγραφή δεδομένων από πολλές μεταβλητές. Τις θέλουμε όλες; **Η κατάρα της διαστατικότητας**

Μείωση διάστασης είναι η διαδικασία που από ένα υψηλο-διάστατο σύνολο δεδομένων δημιουργεί ένα άλλο μικρότερης διάστασης που διατηρεί το μεγαλύτερο μέρος της πληροφορίας των αρχικών δεδομένων.

Δύο βασικές κατηγορίες μεθόδων μείωσης διάστασης:

1. Επιλογή υποσυνόλου μεταβλητών (προβολή στους φυσικούς άξονες).
2. Μετασχηματισμός των μεταβλητών των αρχικών δεδομένων σε νέες λιγότερες μεταβλητές (μετασχηματισμός ή/και προβολή).

Προβλήματα όπου χρησιμοποιείται η μείωση διάστασης:

1. Διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις.
2. Μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης και ταξινόμησης

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;
2. Απόρριψη μεταβλητών που έχουν ισχυρή συσχέτιση με άλλες μεταβλητές. Σε μια τέτοια περίπτωση, ποια από τις δύο θα απορρίψουμε;

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;
2. Απόρριψη μεταβλητών που έχουν ισχυρή συσχέτιση με άλλες μεταβλητές. Σε μια τέτοια περίπτωση, ποια από τις δύο θα απορρίψουμε;
3. Απόρριψη μεταβλητών με 'μικρή σημαντικότητα'. Πως ορίζεται; π.χ. με την τεχνική του τυχαίου δάσους (random forest).

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;
2. Απόρριψη μεταβλητών που έχουν ισχυρή συσχέτιση με άλλες μεταβλητές. Σε μια τέτοια περίπτωση, ποια από τις δύο θα απορρίψουμε;
3. Απόρριψη μεταβλητών με 'μικρή σημαντικότητα'. Πως ορίζεται; π.χ. με την τεχνική του τυχαίου δάσους (random forest).

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;
2. Απόρριψη μεταβλητών που έχουν ισχυρή συσχέτιση με άλλες μεταβλητές. Σε μια τέτοια περίπτωση, ποια από τις δύο θα απορρίψουμε;
3. Απόρριψη μεταβλητών με 'μικρή σημαντικότητα'. Πως ορίζεται; π.χ. με την τεχνική του τυχαίου δάσους (random forest).

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

1. Όλα τα παραπάνω!

Επιλογή υποσυνόλου μεταβλητών

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Απόρριψη μεταβλητών με χαμηλή διασπορά. Πόσο χαμηλή διασπορά; Τυποποίηση των μεταβλητών;
2. Απόρριψη μεταβλητών που έχουν ισχυρή συσχέτιση με άλλες μεταβλητές. Σε μια τέτοια περίπτωση, ποια από τις δύο θα απορρίψουμε;
3. Απόρριψη μεταβλητών με 'μικρή σημαντικότητα'. Πως ορίζεται; π.χ. με την τεχνική του τυχαίου δάσους (random forest).

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

1. Όλα τα παραπάνω!
2. Τεχνικές επιλογής υποσυνόλου μεταβλητών που αποδίδουν καλύτερα (επιλογή προς τα εμπρός, πίσω, βηματική).

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**,
ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών
συσχετίσεων

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**, ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών συσχετίσεων
2. Μη-γραμμικοί μετασχηματισμοί, π.χ. πολυδιάστατη κλιμάκωση (multidimensional scaling), εκμάθηση πολλαπλότητας (manifold learning)

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**, ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών συσχετίσεων
2. Μη-γραμμικοί μετασχηματισμοί, π.χ. πολυδιάστατη κλιμάκωση (multidimensional scaling), εκμάθηση πολλαπλότητας (manifold learning)

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**, ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών συσχετίσεων
2. Μη-γραμμικοί μετασχηματισμοί, π.χ. πολυδιάστατη κλιμάκωση (multidimensional scaling), εκμάθηση πολλαπλότητας (manifold learning)

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

1. Όλα τα παραπάνω!

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**, ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών συσχετίσεων
2. Μη-γραμμικοί μετασχηματισμοί, π.χ. πολυδιάστατη κλιμάκωση (multidimensional scaling), εκμάθηση πολλαπλότητας (manifold learning)

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

1. Όλα τα παραπάνω!
2. Άλλοι γραμμικοί μετασχηματισμοί: γραμμική διαχωριστική ανάλυση (linear discriminant analysis), μερικά ελάχιστα τετράγωνα (partial least squares),

Μετασχηματισμός και προβολή

Για τη διερεύνηση σχέσεων σε μεταβλητές και παρατηρήσεις:

1. Γραμμικοί μετασχηματισμοί: **ανάλυση κύριων συνιστωσών**, ανάλυση ανεξάρτητων συνιστωσών, ανάλυση κανονικών συσχετίσεων
2. Μη-γραμμικοί μετασχηματισμοί, π.χ. πολυδιάστατη κλιμάκωση (multidimensional scaling), εκμάθηση πολλαπλότητας (manifold learning)

Για τη μείωση μεταβλητών εισόδου σε μοντέλα παλινδρόμησης, ταξινόμησης:

1. Όλα τα παραπάνω!
2. Άλλοι γραμμικοί μετασχηματισμοί: γραμμική διαχωριστική ανάλυση (linear discriminant analysis), μερικά ελάχιστα τετράγωνα (partial least squares),
3. Μη-γραμμικά μοντέλα όπως νευρωνικά δίκτυα.

Παράδειγμα προβολής στο επίπεδο

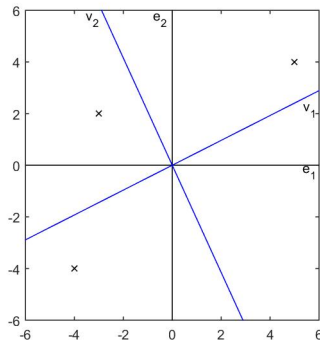
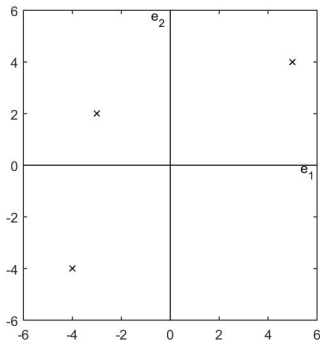
Δίνονται σημεία $\mathbf{x}_i = [x_{i1}, x_{i2}]'$ στο επίπεδο σε φυσικές συντεταγμένες (e_1, e_2) .

Ορίζεται η γραμμή v_1 με γωνία θ από την οριζόντια γραμμή e_1 , $y = \frac{\sin(\theta)}{\cos(\theta)}x$, και η γραμμή v_2 κάθετη στην v_1 στην αρχή των αξόνων, $y = -\frac{\cos(\theta)}{\sin(\theta)}x$.

Παράδειγμα προβολής στο επίπεδο

Δίνονται σημεία $x_i = [x_{i1}, x_{i2}]'$ στο επίπεδο σε φυσικές συντεταγμένες (e_1, e_2) .

Ορίζεται η γραμμή v_1 με γωνία θ από την οριζόντια γραμμή e_1 , $y = \frac{\sin(\theta)}{\cos(\theta)}x$, και η γραμμή v_2 κάθετη στην v_1 στην αρχή των αξόνων, $y = -\frac{\cos(\theta)}{\sin(\theta)}x$.



Ο πίνακας προβολής P στην ευθεία v_1 μπορεί να βρεθεί πως είναι

$$P = \begin{bmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{bmatrix}$$

Ο πίνακας προβολής P στην ευθεία v_1 μπορεί να βρεθεί πως είναι

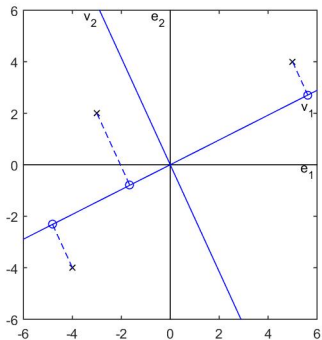
$$P = \begin{bmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{bmatrix}$$

Τα σημεία $\mathbf{y}_i = [y_{i1}, y_{i2}]'$ προβολής των \mathbf{x}_i στην ευθεία v_1 δίνονται ως $\mathbf{y}_i = P\mathbf{x}_i$

Ο πίνακας προβολής P στην ευθεία v_1 μπορεί να βρεθεί πως είναι

$$P = \begin{bmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{bmatrix}$$

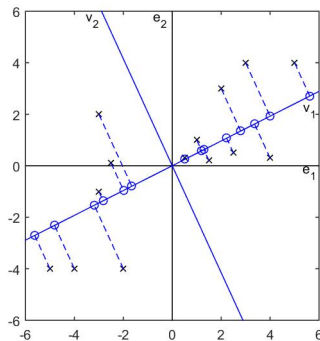
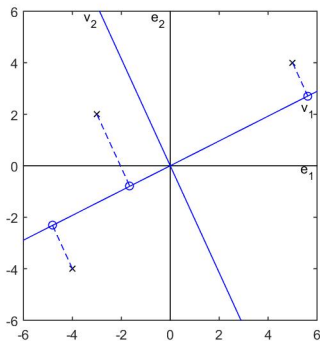
Τα σημεία $y_i = [y_{i1}, y_{i2}]'$ προβολής των x_i στην ευθεία v_1 δίνονται ως $y_i = Px_i$



Ο πίνακας προβολής P στην ευθεία v_1 μπορεί να βρεθεί πως είναι

$$P = \begin{bmatrix} \cos^2(\theta) & \cos(\theta)\sin(\theta) \\ \cos(\theta)\sin(\theta) & \sin^2(\theta) \end{bmatrix}$$

Τα σημεία $y_i = [y_{i1}, y_{i2}]'$ προβολής των x_i στην ευθεία v_1 δίνονται ως $y_i = Px_i$



Ανάλυση κύριων συνιστωσών

Η **ανάλυση κύριων συνιστωσών** (principal component analysis, PCA) σκοπό έχει να βρει έναν υποχώρο διάστασης $d < p$ (p η διάσταση του χώρου των δεδομένων) που εξηγεί όσο το δυνατόν περισσότερο τη διασπορά των σημείων.

Ανάλυση κύριων συνιστωσών

Η **ανάλυση κύριων συνιστωσών** (principal component analysis, PCA) σκοπό έχει να βρει έναν υποχώρο διάστασης $d < p$ (p η διάσταση του χώρου των δεδομένων) που εξηγεί όσο το δυνατόν περισσότερο τη διασπορά των σημείων.

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Ανάλυση κύριων συνιστωσών

Η **ανάλυση κύριων συνιστωσών** (principal component analysis, PCA) σκοπό έχει να βρει έναν υποχώρο διάστασης $d < p$ (p η διάσταση του χώρου των δεδομένων) που εξηγεί όσο το δυνατόν περισσότερο τη διασπορά των σημείων.

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Αρχικά, επεξεργασία των δεδομένων:

1. **Κεντράρισμα**: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

Ανάλυση κύριων συνιστωσών

Η **ανάλυση κύριων συνιστωσών** (principal component analysis, PCA) σκοπό έχει να βρει έναν υποχώρο διάστασης $d < p$ (p η διάσταση του χώρου των δεδομένων) που εξηγεί όσο το δυνατόν περισσότερο τη διασπορά των σημείων.

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Αρχικά, επεξεργασία των δεδομένων:

1. **Κεντράρισμα**: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.
2. Κατά περίπτωση **τυποποίηση** των τιμών διαιρώντας με την τυπική απόκλιση σε κάθε συνιστώσα, x_{ij}/s_j .

Ανάλυση κύριων συνιστωσών

Η **ανάλυση κύριων συνιστωσών** (principal component analysis, PCA) σκοπό έχει να βρει έναν υποχώρο διάστασης $d < p$ (p η διάσταση του χώρου των δεδομένων) που εξηγεί όσο το δυνατόν περισσότερο τη διασπορά των σημείων.

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Αρχικά, επεξεργασία των δεδομένων:

1. **Κεντράρισμα**: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.
2. Κατά περίπτωση **τυποποίηση** των τιμών διαιρώντας με την τυπική απόκλιση σε κάθε συνιστώσα, x_{ij}/s_j .

Βήματα της διαδικασίας PCA:

1. Υπολογισμός πίνακα διασπορών - συνδιασπορών:

$S = \frac{1}{n-1} X^T X$. Αν προηγηθεί τυποποίηση, S είναι πίνακας συσχέτισης.

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_i \mathbf{a}_i^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_i \mathbf{a}_i^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_i \mathbf{a}_i^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

Οι ιδιοτιμές είναι διαταγμένες σε φθίνουσα σειρά:

$$\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p.$$

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_i \mathbf{a}_i^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

Οι ιδιοτιμές είναι διαταγμένες σε φθίνουσα σειρά:

$$\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p.$$

3. Οι νέες μεταβλητές λέγονται **κύριες συνιστώσες** (principal components, PC) και ορίζονται από τον μετασχηματισμό $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$.

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_i \mathbf{a}_i^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

Οι ιδιοτιμές είναι διαταγμένες σε φθίνουσα σειρά:

$$\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p.$$

3. Οι νέες μεταβλητές λέγονται **κύριες συνιστώσες** (principal components, PC) και ορίζονται από τον μετασχηματισμό $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$.

Αυτός ο μετασχηματισμός δηλώνει περιστροφή γύρω από το μέσο όρο (κέντρο βάρους) των \mathbf{x} .

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_j \mathbf{a}_j^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

Οι ιδιοτιμές είναι διαταγμένες σε φθίνουσα σειρά:

$$\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p.$$

3. Οι νέες μεταβλητές λέγονται **κύριες συνιστώσες** (principal components, PC) και ορίζονται από τον μετασχηματισμό $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$.

Αυτός ο μετασχηματισμός δηλώνει περιστροφή γύρω από το μέσο όρο (κέντρο βάρους) των \mathbf{x} .

Τα στοιχεία του \mathbf{a}_j εκφράζουν τις διευθύνσεις συνημιτόνων που συνδέουν το αρχικό με το νέο σύστημα συντεταγμένων.

2. Υπολογισμός **ιδιοτιμών** $\lambda_1, \dots, \lambda_p$ του S : επίλυση της εξίσωσης $|S - \lambda I| = 0$, όπου $|x|$ είναι η ορίζουσα του πίνακα x και I ο μοναδιαίος πίνακας $p \times p$.

Υπολογισμός **ιδιοδιανυσμάτων** $\mathbf{a}_1, \dots, \mathbf{a}_p$ του S : επίλυση των p εξισώσεων $(S - \lambda_j I)\mathbf{a}_j = 0$, $j = 1, \dots, p$, με τη συνθήκη πως τα \mathbf{a}_j είναι ορθοκανονικά ($\mathbf{a}_j \mathbf{a}_j^T = 1$, $\mathbf{a}_j \mathbf{a}_i^T = 0$).

Ισχύει $L = A^T S A$, A πίνακας $p \times p$ των ιδιοδιανυσμάτων και L διαγώνιος $p \times p$ πίνακας των ιδιοτιμών.

Οι ιδιοτιμές είναι διαταγμένες σε φθίνουσα σειρά:

$$\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p.$$

3. Οι νέες μεταβλητές λέγονται **κύριες συνιστώσες** (principal components, PC) και ορίζονται από τον μετασχηματισμό $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$.

Αυτός ο μετασχηματισμός δηλώνει περιστροφή γύρω από το μέσο όρο (κέντρο βάρους) των \mathbf{x} .

Τα στοιχεία του \mathbf{a}_j εκφράζουν τις διευθύνσεις συνημιτόνων που συνδέουν το αρχικό με το νέο σύστημα συντεταγμένων.

Τα PC δίνονται ως γραμμικός συνδυασμός των αρχικών μεταβλητών.

4. Ο μετασχηματισμός των αρχικών παρατηρήσεων στο νέο σύστημα συντεταγμένων PC: $Y = XA$

4. Ο μετασχηματισμός των αρχικών παρατηρήσεων στο νέο σύστημα συντεταγμένων PC: $Y = XA$

Y έχει τα **σκορ κυρίων συνιστωσών** (PC scores). Τα PC scores είναι και αυτά κεντραρισμένα όπως τα στοιχεία του X και ασυσχέτιστα.

4. Ο μετασχηματισμός των αρχικών παρατηρήσεων στο νέο σύστημα συντεταγμένων PC: $Y = XA$

Y έχει τα **σκορ κυρίων συνιστωσών** (PC scores). Τα PC scores είναι και αυτά κεντραρισμένα όπως τα στοιχεία του X και ασυσχέτιστα.

Μπορεί επίσης να εκφραστούν οι αρχικές μεταβλητές ως προς τα PCs: $x = \bar{x} + Ay$.

4. Ο μετασχηματισμός των αρχικών παρατηρήσεων στο νέο σύστημα συντεταγμένων PC: $Y = XA$

Y έχει τα **σκορ κυρίων συνιστωσών** (PC scores). Τα PC scores είναι και αυτά κεντραρισμένα όπως τα στοιχεία του X και ασυσχέτιστα.

Μπορεί επίσης να εκφραστούν οι αρχικές μεταβλητές ως προς τα PCs: $x = \bar{x} + Ay$.

Προβολή

5. Μόνο τα d PCs που αντιστοιχούν σε υψηλές ιδιοτιμές (εκφράζουν μεγάλη μέρος της διασποράς των δεδομένων) συμπεριλαμβάνονται για την ανάλυση και τα $p - d$ απορρίπτονται:

$$Y_d = XA_d$$

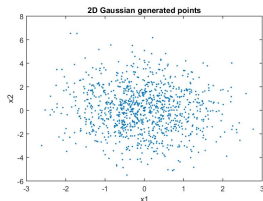
όπου A_d έχει τα πρώτα d ιδιοδιανύσματα και μέγεθος $p \times d$, και ο πίνακας Y_d έχει μέγεθος $n \times d$.

Παράδειγμα

Δημιουργούμε δείγμα $n = 1000$ παρατηρήσεων από διμεταβλητή κανονική κατανομή με διασπορά της πρώτης τ.μ. 1 και της δεύτερης τ.μ. 4. Τα δεδομένα μετασχηματίζονται από το \mathbb{R}^2 στο \mathbb{R}^3 με τον πίνακα

$$W = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$$

Αυτά είναι τα αρχικά δεδομένα στον $n \times p$ πίνακα X ($p = 3$).

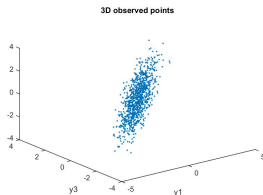
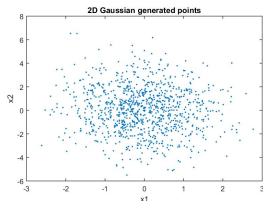


Παράδειγμα

Δημιουργούμε δείγμα $n = 1000$ παρατηρήσεων από διμεταβλητή κανονική κατανομή με διασπορά της πρώτης τ.μ. 1 και της δεύτερης τ.μ. 4. Τα δεδομένα μετασχηματίζονται από το \mathbb{R}^2 στο \mathbb{R}^3 με τον πίνακα

$$W = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$$

Αυτά είναι τα αρχικά δεδομένα στον $n \times p$ πίνακα X ($p = 3$).

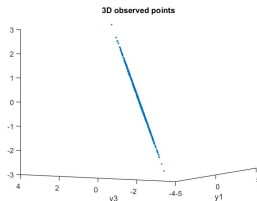
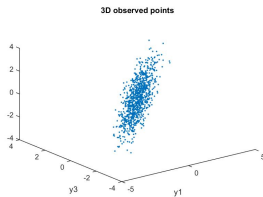
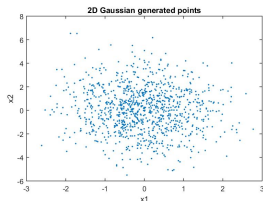


Παράδειγμα

Δημιουργούμε δείγμα $n = 1000$ παρατηρήσεων από διμεταβλητή κανονική κατανομή με διασπορά της πρώτης τ.μ. 1 και της δεύτερης τ.μ. 4. Τα δεδομένα μετασχηματίζονται από το \mathbb{R}^2 στο \mathbb{R}^3 με τον πίνακα

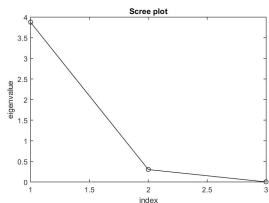
$$W = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$$

Αυτά είναι τα αρχικά δεδομένα στον $n \times p$ πίνακα X ($p = 3$).

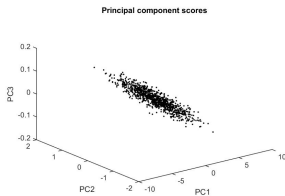
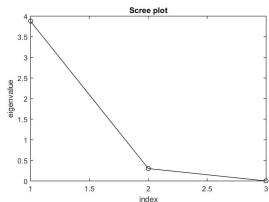


Κεντράρισμα \rightarrow Πίνακας διασπορών-συνδιασπορών \rightarrow Ιδιοτιμές
και ιδιοδιανύσματα \rightarrow Περιστροφή \rightarrow Προβολή στο \mathbb{R}^2

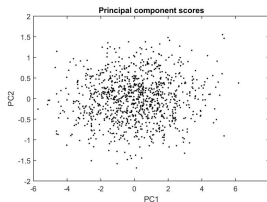
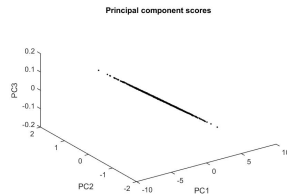
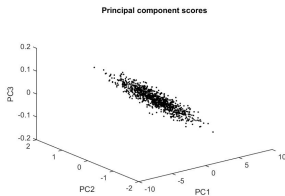
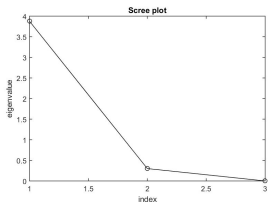
Κεντράρισμα \rightarrow Πίνακας διασπορών-συνδιασπορών \rightarrow Ιδιοτιμές
και ιδιοδιανύσματα \rightarrow Περιστροφή \rightarrow Προβολή στο \mathbb{R}^2



Κεντράρισμα \rightarrow Πίνακας διασπορών-συνδιασπορών \rightarrow Ιδιοτιμές και ιδιοδιανύσματα \rightarrow Περιστροφή \rightarrow Προβολή στο R^2



Κεντράρισμα \rightarrow Πίνακας διασπορών-συνδιασπορών \rightarrow Ιδιοτιμές και ιδιοδιανύσματα \rightarrow Περιστροφή \rightarrow Προβολή στο R^2



Επιλογή του d

Μέθοδοι επιλογής της διάστασης του υποχώρου προβολής:

Επιλογή του d

Μέθοδοι επιλογής της διάστασης του υποχώρου προβολής:

- Αθροιστικό ποσοστό διασποράς που εξηγείται

Ορίζεται ένα κατώφλι του αθροιστικού ποσοστού διασποράς, έστω t_d . Το d θα πρέπει να ικανοποιεί τη σχέση

$$t_d = 100 \frac{\sum_{i=1}^d \lambda_i}{\sum_{j=1}^p \lambda_j}$$

Επιλογή του d

Μέθοδοι επιλογής της διάστασης του υποχώρου προβολής:

- **Αθροιστικό ποσοστό διασποράς που εξηγείται**

Ορίζεται ένα κατώφλι του αθροιστικού ποσοστού διασποράς, έστω t_d . Το d θα πρέπει να ικανοποιεί τη σχέση

$$t_d = 100 \frac{\sum_{i=1}^d \lambda_i}{\sum_{j=1}^p \lambda_j}$$

- **Scree plot**

Γράφημα λ_i vs i .

Ανίχνευση κάποιας αλλαγής στην κλίση της πτωτικής τάσης ('αγκώνας') ή στην οριζοντίωση της.

Επιλογή του d

Μέθοδοι επιλογής της διάστασης του υποχώρου προβολής:

- **Αθροιστικό ποσοστό διασποράς που εξηγείται**

Ορίζεται ένα κατώφλι του αθροιστικού ποσοστού διασποράς, έστω t_d . Το d θα πρέπει να ικανοποιεί τη σχέση

$$t_d = 100 \frac{\sum_{i=1}^d \lambda_i}{\sum_{j=1}^p \lambda_j}$$

- **Scree plot**

Γράφημα λ_i vs i .

Ανίχνευση κάποιας αλλαγής στην κλίση της πτωτικής τάσης ('αγκώνας') ή στην οριζοντίωση της.

- **Μέγεθος διασποράς**

Διατηρούνται οι ιδιοτιμές που είναι πάνω από κάποιο όριο. Το όριο μπορεί να δίνεται ως ποσοστό του μέσου όρου των ιδιοτιμών, π.χ. $\lambda_i \geq 0.7 \bar{\lambda}$.

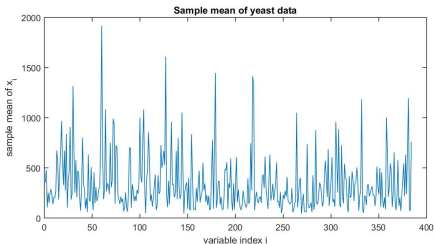
Εφαρμόζουμε PCA σε ένα σύνολο επιπέδων εκφράσεων $p = 384$ γονιδίων σε $n = 17$ χρονικές στιγμές.

Παράδειγμα

Εφαρμόζουμε PCA σε ένα σύνολο επιπέδων εκφράσεων $p = 384$ γονιδίων σε $n = 17$ χρονικές στιγμές.

Κεντράρισμα;

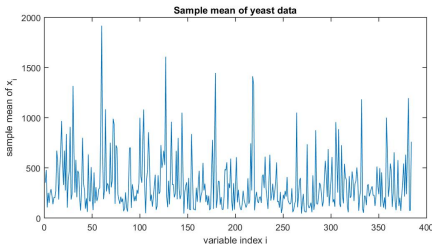
Τυποποίηση ;



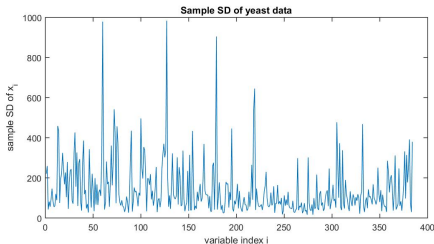
Παράδειγμα

Εφαρμόζουμε PCA σε ένα σύνολο επιπέδων εκφράσεων $p = 384$ γονιδίων σε $n = 17$ χρονικές στιγμές.

Κεντράρισμα;



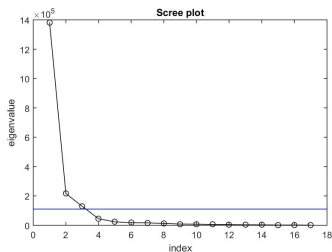
Τυποποίηση ;



Επιλογή διάστασης υποχώρου:

Scree plot

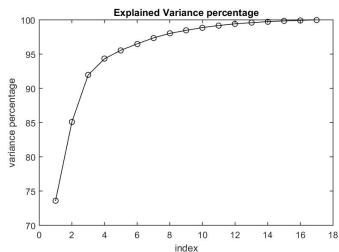
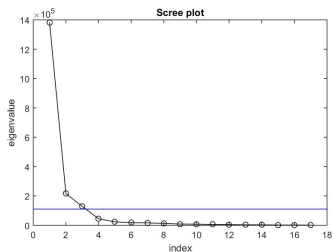
Αθροιστικό ποσοστό διασποράς



Επιλογή διάστασης υποχώρου:

Scree plot

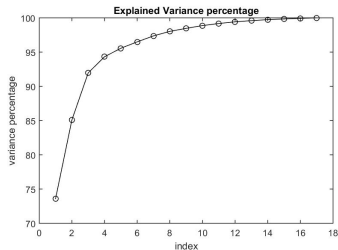
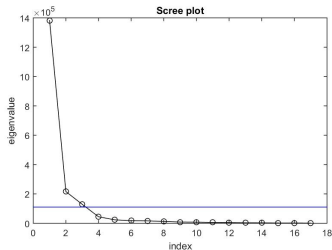
Αθροιστικό ποσοστό διασποράς



Επιλογή διάστασης υποχώρου:

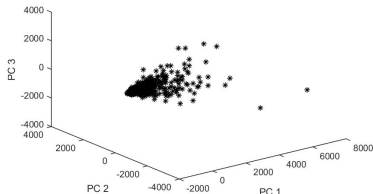
Scree plot

Αθροιστικό ποσοστό διασποράς



Μέγεθος διασποράς (όριο στη μέση τιμή των λ_i)

PCA, d=3



$d = 3, d = 5$ διάσταση του PC υποχώρου από τον αρχικό χώρο διάστασης $p = 17$.

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Κεντράρισμα: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Κεντράρισμα: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

SVD: $X = U \Sigma V^T$

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Κεντράρισμα: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

SVD: $X = U \Sigma V^T$

- U μεγέθους $n \times n$ έχει ορθοκανονικά διανύσματα βάσης του \mathbb{R}^n ($UU^T = I$), **αριστερά ιδιάζοντα διανύσματα**.

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Κεντράρισμα: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

SVD: $X = U \Sigma V^T$

- U μεγέθους $n \times n$ έχει ορθοκανονικά διανύσματα βάσης του \mathbb{R}^n ($UU^T = I$), **αριστερά ιδιάζοντα διανύσματα**.
- Σ διαγώνιος μεγέθους $n \times p$ έχει τις **ιδιάζουσες τιμές** στη διαγώνιο $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$.

Ανάλυση ιδιάζουσων τιμών

Η **ανάλυση ιδιάζουσων τιμών** (singular value decomposition, SVD) είναι ισοδύναμη με την PCA αλλά εφαρμόζεται απευθείας στον πίνακα δεδομένων X .

Έστω $X = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ μεγέθους $n \times p$ και $\mathbf{x}_i \in \mathbb{R}^p$

Κεντράρισμα: για κάθε συνιστώσα j , $j = 1, \dots, p$, το x_{ij} αντικαθίσταται από $x_{ij} - \bar{x}_j$, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$.

SVD: $X = U \Sigma V^T$

- U μεγέθους $n \times n$ έχει ορθοκανονικά διανύσματα βάσης του \mathbb{R}^n ($UU^T = I$), **αριστερά ιδιάζοντα διανύσματα**.
- Σ διαγώνιος μεγέθους $n \times p$ έχει τις **ιδιάζουσες τιμές** στη διαγώνιο $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$.
- V μεγέθους $p \times p$ έχει ορθοκανονικά διανύσματα βάσης του \mathbb{R}^p ($VV^T = I$), **δεξιά ιδιάζοντα διανύσματα**.

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow \text{SVD}$ 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow \text{SVD}$ 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Για τη μείωση διάστασης των δεδομένων σε $d < p$, θεωρούμε πως d προσεγγίζει την τάξη r του X . Η προσέγγιση χαμηλής τάξης του X (**low-rank approximation**):

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow \text{SVD}$ 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Για τη μείωση διάστασης των δεδομένων σε $d < p$, θεωρούμε πως d προσεγγίζει την τάξη r του X . Η προσέγγιση χαμηλής τάξης του X (**low-rank approximation**):

$$X_d = U_d \Sigma_d V_d^T$$

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow \text{SVD}$ 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Για τη μείωση διάστασης των δεδομένων σε $d < p$, θεωρούμε πως d προσεγγίζει την τάξη r του X . Η προσέγγιση χαμηλής τάξης του X (**low-rank approximation**):

$$X_d = U_d \Sigma_d V_d^T$$

- U_d μεγέθους $n \times d$, επιλέγονται μόνο τα πρώτα d αριστερά ιδιάζοντα διανύσματα.

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow$ SVD 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Για τη μείωση διάστασης των δεδομένων σε $d < p$, θεωρούμε πως d προσεγγίζει την τάξη r του X . Η προσέγγιση χαμηλής τάξης του X (**low-rank approximation**):

$$X_d = U_d \Sigma_d V_d^T$$

- U_d μεγέθους $n \times d$, επιλέγονται μόνο τα πρώτα d αριστερά ιδιάζοντα διανύσματα.
- Σ_d μεγέθους $d \times d$, επιλέγονται μόνο οι πρώτες d ιδιάζουσες τιμές.

Συνήθως έχουμε $n > p$ (περισσότερες παρατηρήσεις από μεταβλητές).

Τα $n - p$ αριστερά ιδιάζοντα διανύσματα είναι αυθαίρετα και αντιστοιχούν σε μηδενικές γραμμές του $\Sigma \Rightarrow$ SVD 'οικονομικού μεγέθους': $X = U\Sigma V^T$ με U μεγέθους $n \times p$ και Σ μεγέθους $p \times p$.

Για τη μείωση διάστασης των δεδομένων σε $d < p$, θεωρούμε πως d προσεγγίζει την τάξη r του X . Η προσέγγιση χαμηλής τάξης του X (low-rank approximation):

$$X_d = U_d \Sigma_d V_d^T$$

- U_d μεγέθους $n \times d$, επιλέγονται μόνο τα πρώτα d αριστερά ιδιάζοντα διανύσματα.
- Σ_d μεγέθους $d \times d$, επιλέγονται μόνο οι πρώτες d ιδιάζουσες τιμές.
- V_d μεγέθους $p \times d$, επιλέγονται μόνο τα πρώτα d δεξιά ιδιάζοντα διανύσματα.

Ισοδυναμία SVD και PCA

Ισοδυναμία SVD και PCA

- $\lambda_i = \sigma_i^2 / (n - 1), i = 1, \dots, p$
($X^T X = V(\Sigma^T \Sigma)V^T$ και $S = \frac{1}{n-1}X^T X$)

Ισοδυναμία SVD και PCA

- $\lambda_i = \sigma_i^2 / (n - 1)$, $i = 1, \dots, p$
($X^T X = V(\Sigma^T \Sigma)V^T$ και $S = \frac{1}{n-1}X^T X$)
- $A = V$, $\mathbf{a}_i = \mathbf{v}_i$, ιδιοδιανύσματα είναι τα δεξιά ιδιάζοντα διανύσματα.

Ισοδυναμία SVD και PCA

- $\lambda_i = \sigma_i^2 / (n - 1)$, $i = 1, \dots, p$
($X^T X = V(\Sigma^T \Sigma)V^T$ και $S = \frac{1}{n-1}X^T X$)
- $A = V$, $\mathbf{a}_i = \mathbf{v}_i$, ιδιοδιανύσματα είναι τα δεξιά ιδιάζοντα διανύσματα.

Κύριες συνιστώσες,

PCA: $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, SVD: $y_j = \mathbf{v}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$

Ισοδυναμία SVD και PCA

- $\lambda_i = \sigma_i^2 / (n - 1)$, $i = 1, \dots, p$
($X^T X = V(\Sigma^T \Sigma)V^T$ και $S = \frac{1}{n-1}X^T X$)
- $A = V$, $\mathbf{a}_i = \mathbf{v}_i$, ιδιοδιανύσματα είναι τα δεξιά ιδιάζοντα διανύσματα.

Κύριες συνιστώσες,

PCA: $y_j = \mathbf{a}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, SVD: $y_j = \mathbf{v}_j^T (\mathbf{x} - \bar{\mathbf{x}})$, $j = 1, \dots, p$

Σκορ κύριων συνιστωσών,

PCA: $Y = XA$, SVD: $Y = XV = U\Sigma$

Ασκήσεις Κεφαλαίου 6

1. Δημιουργείτε δείγμα $n = 1000$ παρατηρήσεων από διμεταβλητή κανονική κατανομή με διασπορά της πρώτης τ.μ. 1 και της δεύτερης τ.μ. 4. Στη συνέχεια μετασχηματίστε τα δεδομένα από το \mathbf{R}^2 στο \mathbf{R}^3 με τον πίνακα

$$W = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.5 \\ 0.7 & 0.3 \end{bmatrix}$$

- (α') Βρείτε τις ιδιοτιμές και ιδιοδιανύσματα του πίνακα διασπορών - συνδιασπορών. Σχηματίστε τα σκορ κυρίων συνιστωσών στο \mathbf{R}^3 .
- (β') Κάνετε το scree plot. Δείξτε γραφικά πως τα σκορ κυρίων συνιστωσών περιορίζονται στο \mathbf{R}^2 (κάνετε κατάλληλη περιστροφή στο τρισδιάστατο σχήμα).
- (γ') Σχηματίστε τα σκορ κυρίων συνιστωσών στο \mathbf{R}^2 και συγκρίνετε αυτό το σχήμα με το αρχικό των σημείων που δημιουργήσατε (στο \mathbf{R}^2).
2. Εφαρμόστε την ανάλυση PCA σε ένα σύνολο επιπέδων εκφράσεων $p = 384$ (μεταβλητές) γονιδίων σε $n = 17$ χρονικές στιγμές (παρατηρήσεις), που δίνονται στο αρχείο `yeast.dat`.
- (α') Εκτιμήστε τη διάσταση $d \leq p$ για τη μείωση διάστασης με ΠΑ.
- (β') Σχηματίστε τα σκορ κυρίων συνιστωσών στο \mathbf{R}^2 και \mathbf{R}^3 .
3. Εφαρμόστε την ανάλυση PCA στα δεδομένα της Άσκησης 5.8 (στο αρχείο `physical.txt`).
- (α') Εκτιμήστε τη διάσταση $d \leq p$ για τη μείωση διάστασης με ΠΑ.
- (β') Σχηματίστε τα σκορ κυρίων συνιστωσών στο \mathbf{R}^2 και \mathbf{R}^3 .

Η **ανάλυση ανεξάρτητων συνιστωσών** (independent component analysis, ICA) χρησιμοποιείται για την εξαγωγή κρυμμένων / υποκείμενων συνιστωσών του X .

Η **ανάλυση ανεξάρτητων συνιστωσών** (independent component analysis, ICA) χρησιμοποιείται για την εξαγωγή κρυμμένων / υποκείμενων συνιστωσών του X .

PCA βρίσκει τις **ασυσχέτιστες** συνιστώσες.

ICA βρίσκει τις συνιστώσες που είναι όσο το δυνατόν πιο **ανεξάρτητες** και με **μη-Γκαουσιανή** κατανομή.

Ανάλυση ανεξάρτητων συνιστωσών

Η **ανάλυση ανεξάρτητων συνιστωσών** (independent component analysis, ICA) χρησιμοποιείται για την εξαγωγή κρυμμένων / υποκείμενων συνιστωσών του X .

PCA βρίσκει τις **ασυσχέτιστες** συνιστώσες.

ICA βρίσκει τις συνιστώσες που είναι όσο το δυνατόν πιο **ανεξάρτητες** και με **μη-Γκαουσιανή** κατανομή.

ICA έχει χρησιμοποιηθεί κυρίως για το διαχωρισμό σημάτων σε πηγές (blind source separation),

Ανάλυση ανεξάρτητων συνιστωσών

Η **ανάλυση ανεξάρτητων συνιστωσών** (independent component analysis, ICA) χρησιμοποιείται για την εξαγωγή κρυμμένων / υποκείμενων συνιστωσών του X .

PCA βρίσκει τις **ασυσχέτιστες** συνιστώσες.

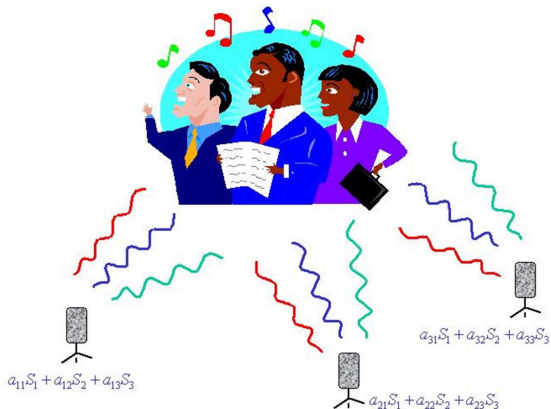
ICA βρίσκει τις συνιστώσες που είναι όσο το δυνατόν πιο **ανεξάρτητες** και με **μη-Γκαουσιανή** κατανομή.

ICA έχει χρησιμοποιηθεί κυρίως για το διαχωρισμό σημάτων σε πηγές (blind source separation),

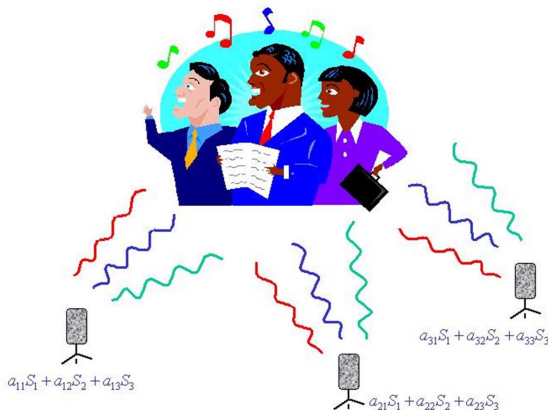
$X = [x_1, \dots, x_p]$ μεγέθους $n \times p$

p σήματα, $x_i = [x_i(1), x_i(2), \dots, x_i(n)]^T, i = 1, \dots, p$

Η ICA καλείται να λύσει το πρόβλημα του 'coctail party'



Η ICA καλείται να λύσει το πρόβλημα του 'coctail party'



Υποθέσεις:

1. Οι πηγές s_i (κρυφές συνιστώσες των x_i) είναι ανεξάρτητες.
2. Οι πηγές s_i έχουν μη-Γκαουσιανή κατανομή.

Πίνακας παρατηρούμενων σημάτων: $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ μεγέθους $n \times p$.

Πίνακας σημάτων πηγών: $S = [\mathbf{s}_1, \dots, \mathbf{s}_p]$ μεγέθους $n \times p$.

Πίνακας παρατηρούμενων σημάτων: $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ μεγέθους $n \times p$.

Πίνακας σημάτων πηγών: $S = [\mathbf{s}_1, \dots, \mathbf{s}_p]$ μεγέθους $n \times p$.

$X = SA$ (μίξη των ανεξάρτητων πηγών)

$$\mathbf{x}_i = a_{i,1}\mathbf{s}_1 + \dots + a_{i,p}\mathbf{s}_p, \quad i = 1, \dots, p$$

A : πίνακας μίξης (mixing matrix) μεγέθους $p \times p$.

Πίνακας παρατηρούμενων σημάτων: $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ μεγέθους $n \times p$.

Πίνακας σημάτων πηγών: $S = [\mathbf{s}_1, \dots, \mathbf{s}_p]$ μεγέθους $n \times p$.

$X = SA$ (μίξη των ανεξάρτητων πηγών)

$$\mathbf{x}_i = a_{i,1}\mathbf{s}_1 + \dots + a_{i,p}\mathbf{s}_p, \quad i = 1, \dots, p$$

A : πίνακας μίξης (mixing matrix) μεγέθους $p \times p$.

Αντίστροφα

$S = XW$ (διαχωρισμός των αναμεμειγμένων σημάτων)

$W = A^{-1}$: πίνακας διαχωρισμού (separating matrix) μεγέθους $p \times p$.

Πίνακας παρατηρούμενων σημάτων: $X = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ μεγέθους $n \times p$.

Πίνακας σημάτων πηγών: $S = [\mathbf{s}_1, \dots, \mathbf{s}_p]$ μεγέθους $n \times p$.

$$X = SA \quad (\text{μίξη των ανεξάρτητων πηγών})$$

$$\mathbf{x}_i = a_{i,1}\mathbf{s}_1 + \dots + a_{i,p}\mathbf{s}_p, \quad i = 1, \dots, p$$

A : πίνακας μίξης (mixing matrix) μεγέθους $p \times p$.

Αντίστροφα

$$S = XW \quad (\text{διαχωρισμός των αναμεμειγμένων σημάτων})$$

$W = A^{-1}$: πίνακας διαχωρισμού (separating matrix) μεγέθους $p \times p$.

Διαφορετικές μέθοδοι βρίσκουν τον πίνακα W που μεγιστοποιεί την ανεξαρτησία ή/και τη μη-Γκαουσιανή κατανομή.

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.
2. Η σειρά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη.

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.
2. Η σειρά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη.

Γιατί οι ανεξάρτητες συνιστώσες s_i πρέπει να είναι μη-Γκαουσιανές;

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.
2. Η σειρά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη.

Γιατί οι ανεξάρτητες συνιστώσες s_i πρέπει να είναι μη-Γκαουσιανές;

Κεντρικό οριακό θεώρημα: Ο μέσος όρος ανεξάρτητων τυχαίων μεταβλητών θα έχει κατανομή πιο κοντά σε Γκαουσιανή. \Rightarrow Η μίξη των συνιστωσών πηγής x_i θα είναι πιο κοντά σε Γκαουσιανή από ότι οι ανεξάρτητες συνιστώσες s_i .

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.
2. Η σειρά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη.

Γιατί οι ανεξάρτητες συνιστώσες s_i πρέπει να είναι μη-Γκαουσιανές;

Κεντρικό οριακό θεώρημα: Ο μέσος όρος ανεξάρτητων τυχαίων μεταβλητών θα έχει κατανομή πιο κοντά σε Γκαουσιανή. \Rightarrow Η μίξη των συνιστωσών πηγής x_i θα είναι πιο κοντά σε Γκαουσιανή από ότι οι ανεξάρτητες συνιστώσες s_i .

Για αυτό και η μέθοδος βασίζεται στη μεγιστοποίηση της απόκλισης από Γκαουσιανή της εκτιμώμενης πηγής $s_i = \sum_j w_{ij} x_j$.

Ασάφειες στη λύση της $S = XW$:

1. Η διασπορά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη. Μπορούμε αυθαίρετα να θέσουμε $\text{Var}(s_i^2) = 1$.
2. Η σειρά των ανεξάρτητων συνιστωσών δεν είναι καθορισμένη.

Γιατί οι ανεξάρτητες συνιστώσες s_i πρέπει να είναι μη-Γκαουσιανές;

Κεντρικό οριακό θεώρημα: Ο μέσος όρος ανεξάρτητων τυχαίων μεταβλητών θα έχει κατανομή πιο κοντά σε Γκαουσιανή. \Rightarrow Η μίξη των συνιστωσών πηγής x_i θα είναι πιο κοντά σε Γκαουσιανή από ότι οι ανεξάρτητες συνιστώσες s_i .

Για αυτό και η μέθοδος βασίζεται στη μεγιστοποίηση της απόκλισης από Γκαουσιανή της εκτιμώμενης πηγής $s_i = \sum_j w_{ij} x_j$.

Αν πάνω από μια συνιστώσα s_i είναι Γκαουσιανές, τότε δε μπορούν να ξεχωρίσουν.

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση σημαίνει ο X να μετασχηματιστεί σε X' που έχει $\text{Cov}(X') = I$:

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση σημαίνει ο X να μετασχηματιστεί σε X' που έχει $\text{Cov}(X') = I$:

$$X' = \Sigma^{-1} U X \text{ (}\Sigma \text{ και } U \text{ από SVD)}$$

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση σημαίνει ο X να μετασχηματιστεί σε X' που έχει $\text{Cov}(X') = I$:

$$X' = \Sigma^{-1} U X \text{ (}\Sigma \text{ και } U \text{ από SVD)}$$

s_i ανεξάρτητα (με διασπορά 1) $\Rightarrow \text{Cov}(S) = I$.

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση σημαίνει ο X να μετασχηματιστεί σε X' που έχει $\text{Cov}(X') = I$:

$$X' = \Sigma^{-1} U X \quad (\Sigma \text{ και } U \text{ από SVD})$$

s_i ανεξάρτητα (με διασπορά 1) $\Rightarrow \text{Cov}(S) = I$.

$$\text{Για } \mathbf{x}_i: \text{Cov}(X) = \frac{1}{n-1} X^T X = (SA)^T SA = A^T A.$$

Προλεύκανση του X για πιο αποτελεσματικό διαχωρισμό των πηγών.

Προλεύκανση σημαίνει ο X να μετασχηματιστεί σε X' που έχει $\text{Cov}(X') = I$:

$$X' = \Sigma^{-1}UX \text{ (}\Sigma \text{ και } U \text{ από SVD)}$$

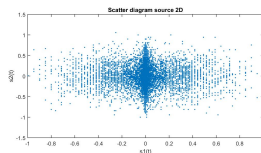
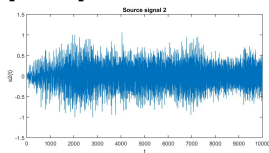
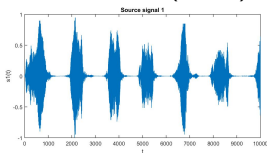
s_i ανεξάρτητα (με διασπορά 1) $\Rightarrow \text{Cov}(S) = I$.

$$\text{Για } \mathbf{x}_i: \text{Cov}(X) = \frac{1}{n-1}X^T X = (SA)^T SA = A^T A.$$

Αν ο X έχει $\text{Cov}(X) = I \Rightarrow A^T A = I \Rightarrow A$ ορθοκανονικός πίνακας (αντί να έχουμε να εκτιμήσουμε p^2 στοιχεία θα έχουμε $p(p-1)/2$).

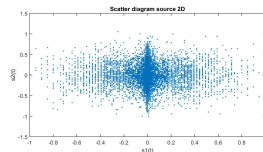
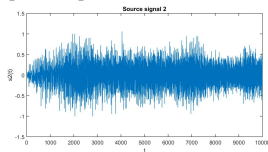
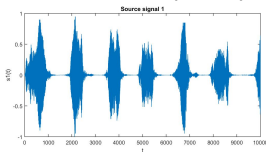
Παράδειγμα

Δύο ανεξάρτητα σήματα πηγών: 1) τερέτισμα (chirp), 2) ήχος κρουστού (gong) $S = [s_1, s_2]$



Παράδειγμα

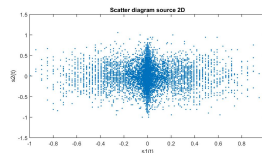
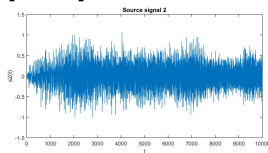
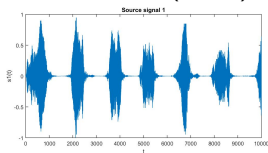
Δύο ανεξάρτητα σήματα πηγών: 1) τερέτισμα (chirp), 2) ήχος κρουστού (gong) $S = [s_1, s_2]$



Τυχάιος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 \\ -2.5 & -0.2 \end{bmatrix}$

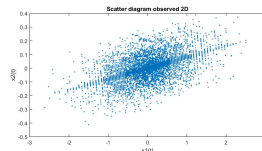
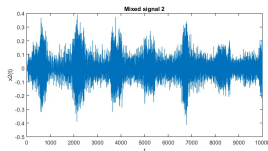
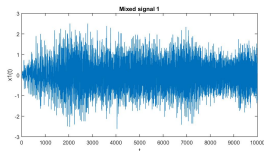
Παράδειγμα

Δύο ανεξάρτητα σήματα πηγών: 1) τερέτισμα (chirp), 2) ήχος κρουστού (gong) $S = [s_1, s_2]$



Τυχάιος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 \\ -2.5 & -0.2 \end{bmatrix}$

Τα παρατηρούμενα σήματα, $X = [x_1, x_2]$



Εφαρμογή της ICA (συνάρτηση `rica.m` στο Matlab)

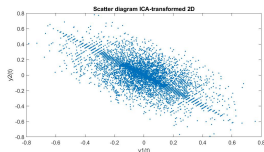
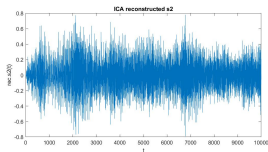
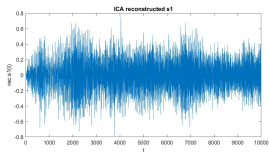
Εφαρμογή της ICA (συνάρτηση `rica.m` στο Matlab)

$$\text{Πίνακας διαχωρισμού } W = \begin{bmatrix} -0.1820 & 0.3317 \\ -0.9833 & -0.9434 \end{bmatrix}$$

Εφαρμογή της ICA (συνάρτηση rica.m στο Matlab)

Πίνακας διαχωρισμού $W = \begin{bmatrix} -0.1820 & 0.3317 \\ -0.9833 & -0.9434 \end{bmatrix}$

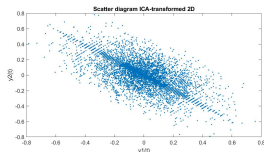
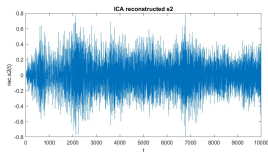
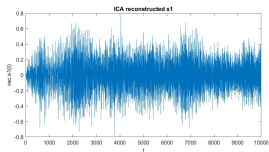
Τα σήματα από ICA, $S = XW$



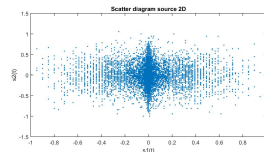
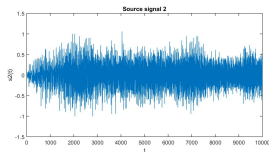
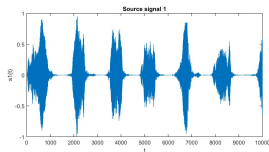
Εφαρμογή της ICA (συνάρτηση `rica.m` στο Matlab)

Πίνακας διαχωρισμού $W = \begin{bmatrix} -0.1820 & 0.3317 \\ -0.9833 & -0.9434 \end{bmatrix}$

Τα σήματα από ICA, $S = XW$

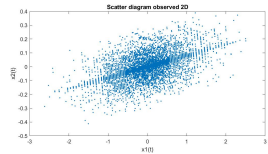
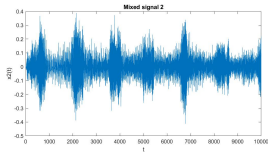
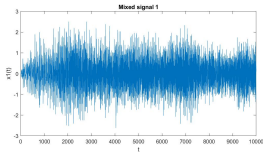


Δε φαίνεται να ανακτώνται τα αρχικά σήματα πηγής!

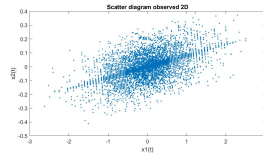
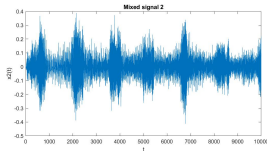
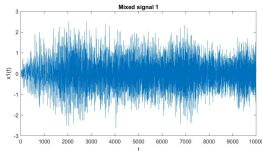


Με προλεύκανση, $X' = \Sigma^{-1}UX$ (Σ και U από SVD)

Με προλεύκανση, $X' = \Sigma^{-1}UX$ (Σ και U από SVD)

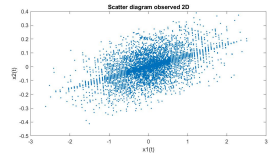
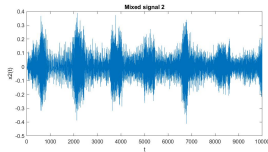
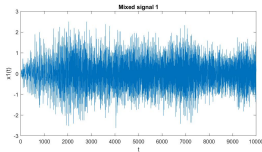


Με προλεύκανση, $X' = \Sigma^{-1}UX$ (Σ και U από SVD)



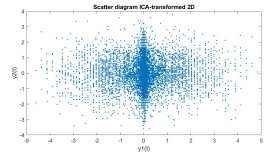
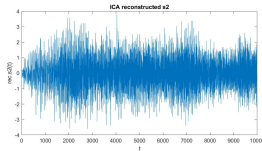
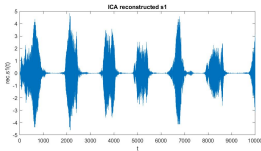
$$\text{Πίνακας διαχωρισμού } W' = \begin{bmatrix} -0.0290 & 0.9995 \\ -0.9996 & -0.0304 \end{bmatrix}$$

Με **προλεύκανση**, $X' = \Sigma^{-1}UX$ (Σ και U από SVD)

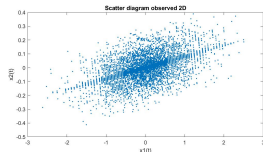
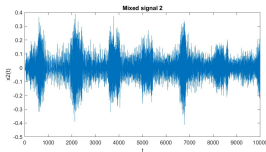
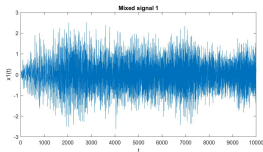


$$\text{Πίνακας διαχωρισμού } W' = \begin{bmatrix} -0.0290 & 0.9995 \\ -0.9996 & -0.0304 \end{bmatrix}$$

Τα σήματα από ICA, $S = X'W'$

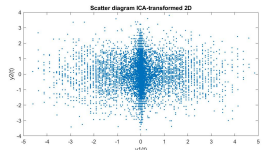
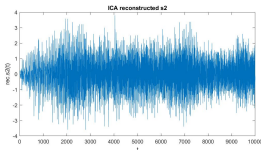
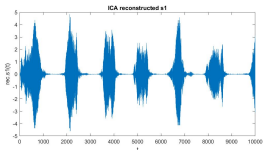


Με **προλεύκανση**, $X' = \Sigma^{-1}UX$ (Σ και U από SVD)

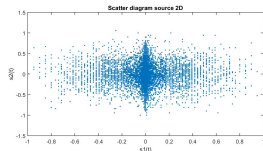
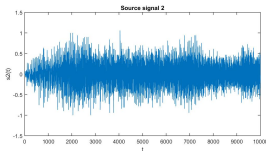
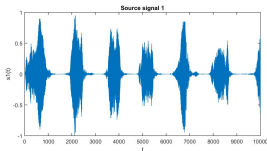


Πίνακας διαχωρισμού $W' = \begin{bmatrix} -0.0290 & 0.9995 \\ -0.9996 & -0.0304 \end{bmatrix}$

Τα σήματα από ICA, $S = X'W'$



Τα αρχικά σήματα ανακτώνται!



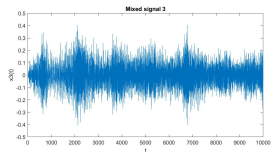
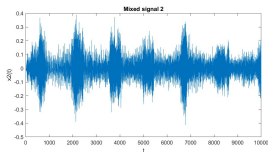
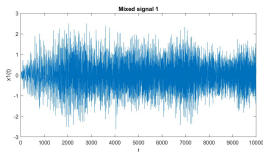
Η μείωση διάστασης μπορεί να γίνει στο στάδιο της
προλεύκανσης (μηδενικές ιδιοτιμές)

Η μείωση διάστασης μπορεί να γίνει στο στάδιο της προλεύκανσης (μηδενικές ιδιοτιμές)

Τυχαίος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 & -0.2 \\ -2.5 & -0.2 & -0.3 \end{bmatrix}$

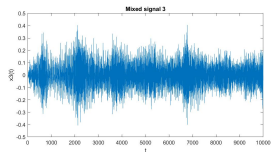
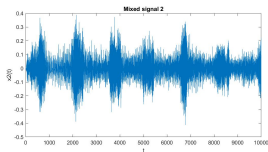
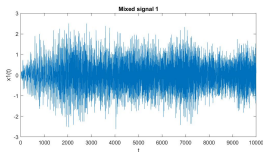
Η μείωση διάστασης μπορεί να γίνει στο στάδιο της προλεύκανσης (μηδενικές ιδιοτιμές)

Τυχαίος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 & -0.2 \\ -2.5 & -0.2 & -0.3 \end{bmatrix}$



Η μείωση διάστασης μπορεί να γίνει στο στάδιο της προλεύκανσης (μηδενικές ιδιοτιμές)

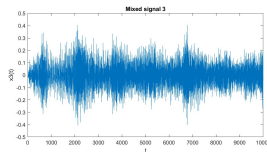
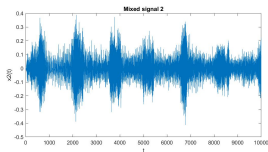
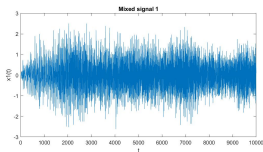
Τυχαίος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 & -0.2 \\ -2.5 & -0.2 & -0.3 \end{bmatrix}$



Πίνακας διαχωρισμού $W' = \begin{bmatrix} 0.0358 & 0.9993 \\ -0.9994 & 0.0372 \end{bmatrix}$

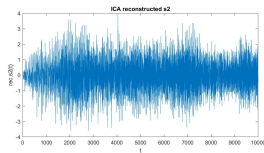
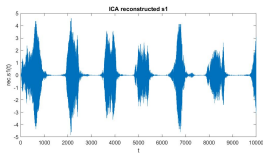
Η μείωση διάστασης μπορεί να γίνει στο στάδιο της προλεύκανσης (μηδενικές ιδιοτιμές)

Τυχάιος πίνακας μίξης $A = \begin{bmatrix} -0.1 & 0.3 & -0.2 \\ -2.5 & -0.2 & -0.3 \end{bmatrix}$



Πίνακας διαχωρισμού $W' = \begin{bmatrix} 0.0358 & 0.9993 \\ -0.9994 & 0.0372 \end{bmatrix}$

Τα σήματα από ICA, $S = X'W'$



4. Φορτώστε τα σήματα ήχου από 1) τερέτισμα και 2) κρουστό από τα αρχεία `chirp.mat` και `gong.mat` που υπάρχουν ως παραδείγματα στο Matlab. Κρατήστε τις πρώτες 10000 παρατηρήσεις από κάθε σήμα.
- (α') Επιλέξτε ένα τυχαίο πίνακα μίξης A μεγέθους 2×2 και μετασχηματίστε τα αρχικά σήματα σε δύο αναμεμιγμένα σήματα. Εφαρμόστε την μέθοδο ICA με και χωρίς προλεύκανση και ελέγξτε αν ανακτώνται τα δύο αρχικά σήματα (τερέτισμα και κρουστό).
- (β') Επαναλάβετε το ίδιο με παραπάνω αλλά για πίνακα μίξης A μεγέθους 2×3 .