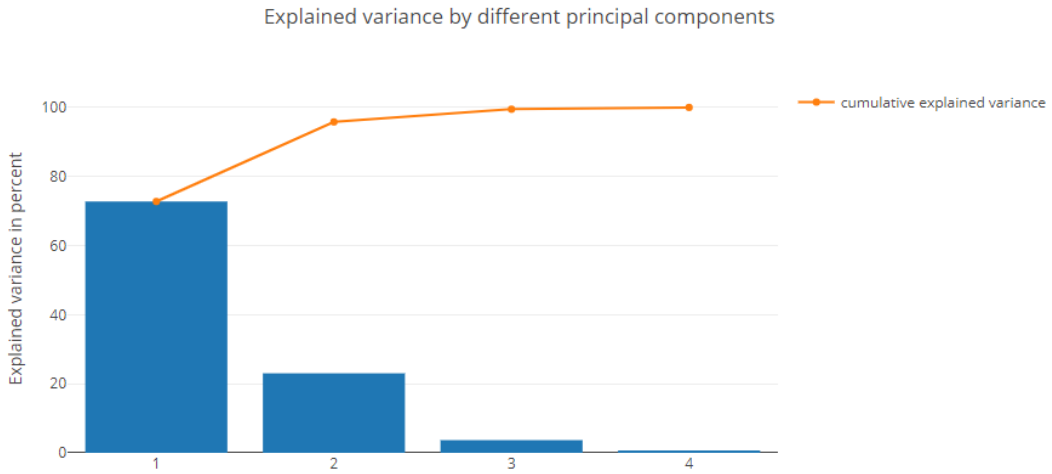
2η εργασία στο μάθημα ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Μέρος 1ο

Πόσες διαστάσεις θα πρέπει να λάβουμε υπόψη για να διαχωρίσουμε ευκρινώς τα δείγματα;

Με το PCA, μειώνουμε τις διαστάσεις ενός d-διαστατικού συνόλου δεδομένων προβάλλοντάς το σε ένα k - διαστατικό υποσύστημα (όπου k < d) προκειμένου να αυξηθεί η υπολογιστική αποτελεσματικότητα διατηρώντας παράλληλα τις περισσότερες πληροφορίες. Εκτελώντας τα βήματά του, φτάνουμε στο παρακάτω γράφημα όπου η εξηγηθείσα διακύμανση μας λέει πόση πληροφορία μπορεί να αποδοθεί σε κάθε κύρια συνιστώσα.



Η παραπάνω γραφική παράσταση δείχνει σαφώς ότι το μεγαλύτερο μέρος της διακύμανσης (72,77% της διακύμανσης είναι ακριβές) μπορεί να εξηγηθεί από το πρώτο κύριο συστατικό μόνο. Το δεύτερο κύριο στοιχείο εξακολουθεί να έχει μερικές πληροφορίες (23,03%), ενώ το τρίτο και τέταρτο βασικό στοιχείο μπορεί να πέσει με ασφάλεια χωρίς να χάσει πολλές πληροφορίες. Μαζί, τα πρώτα δύο κύρια στοιχεία περιέχουν 95,8% των πληροφοριών. Επομένως, η μείωση των διαστάσεων από 4 σε 2 οδηγεί στην διατήρηση του 95,8% της διασποράς δηλαδή μπορεί να επιτευχθεί ικανοποιητικός διαχωρισμός.

Ποιο μέτρο απόστασης θεωρείται καλύτερο για το συγκεκριμένο πρόβλημα;

Για το συγκεκριμένα πρόβλημα, το μέτρο απόστασης Mahalanobis θεωρείται καλύτερο. Με την mahalanobis όλες οι πληροφορίες λαμβάνονται υπόψη στην τελική τιμή οπότε θα έχουμε καλύτερο αποτέλεσμα.