

Η χρήση της μεθόδου PCA στη Βιοτεχνολογία

Κωνσταντίνος Μάστορας

May 4, 2025

1. Εισαγωγή

Η ανάλυση κύριων συνιστωσών (Principal Component Analysis, PCA) είναι μία τεχνική γραμμικού μετασχηματισμού που χρησιμοποιείται για τη μείωση της διαστασιμότητας σε δεδομένα με μεγάλο αριθμό μεταβλητών. Στη βιοτεχνολογία, εφαρμόζεται ευρέως στην ανάλυση έκφρασης γονιδίων, στη μελέτη πρωτεϊνικών αλληλουχιών, στην κατηγοριοποίηση ιστών και στη διάγνωση ασθενειών.

Η παρούσα εργασία παρουσιάζει τη βασική μαθηματική θεωρία της μεθόδου PCA, ιστορικές και σύγχρονες εφαρμογές της στον τομέα της βιοτεχνολογίας, καθώς και ένα θεωρητικό παράδειγμα εφαρμογής της μεθόδου.

2. Βιβλιογραφική Ανασκόπηση

Η μέθοδος PCA εισήχθη από τον Karl Pearson το 1901 και αναπτύχθηκε περαιτέρω από τον Harold Hotelling το 1933. Έκτοτε, έχει εφαρμοστεί σε πληθώρα επιστημών, με ιδιαίτερη σημασία στις βιολογικές επιστήμες.

Στη βιοτεχνολογία, η PCA έχει εφαρμοστεί στην επεξεργασία δεδομένων microarray, RNA-Seq και άλλων τεχνικών βιολογικής ανάλυσης. Μελέτες έχουν δείξει ότι η PCA μπορεί να αποκαλύψει υποκείμενες δομές και μοτίβα σε πολυδιάστατα δεδομένα.

3. Μαθηματική Περιγραφή

Έστω $X \in \mathbb{R}^{n \times p}$ ο πίνακας δεδομένων, όπου n οι παρατηρήσεις και p τα χαρακτηριστικά. Ο πρώτος βήμα είναι η κεντραρισμένη μήτρα $\tilde{X}$, με μέση τιμή μηδέν ανά στήλη.

Ο πίνακας συνδιακύμανσης είναι:

$$S = \frac{1}{n-1} \tilde{X}^T \tilde{X}$$

Η PCA υπολογίζει τα ιδιοδιανύσματα και τις ιδιοτιμές του S . Το ιδιοδιάνυσμα που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή ορίζει την πρώτη κύρια συνιστώσα, κ.ο.κ.

Παράδειγμα: Έστω δεδομένα έκφρασης 100 γονιδίων σε 10 δείγματα. Το PCA μπορεί να μειώσει τις 100 διαστάσεις σε 2 ή 3, διατηρώντας τη μεγαλύτερη δυνατή διακύμανση.

4. Συμπεράσματα

Η PCA είναι ένα ισχυρό εργαλείο για τη μείωση διαστασιμότητας και την εξαγωγή πληροφορίας από πολύπλοκα βιολογικά δεδομένα. Παρά τους περιορισμούς της (γραμμικότητα, ευαισθησία σε θόρυβο), η χρήση της σε σύγχρονες εφαρμογές της βιοτεχνολογίας παραμένει καθοριστική.

@articlepearson1901, author = Pearson, K., title = On Lines and Planes of Closest Fit to Systems of Points in Space, journal = Philosophical Magazine, year = 1901, volume = 2, pages = 559–572

@articlehotelling1933, author = Hotelling, H., title = Analysis of a complex of statistical variables into principal components, journal = Journal of Educational Psychology, year = 1933, volume = 24, pages = 417–441

@articlealter2000, author = Alter, O. and Brown, P.O. and Botstein, D., title = Singular value decomposition for genome-wide expression data processing and modeling, journal = Proceedings of the National Academy of Sciences, year = 2000, volume = 97, number = 18, pages = 10101–10106

@articleringner2008, author = Ringnér, M., title = What is principal component analysis?, journal = Nature Biotechnology, year = 2008, volume = 26, pages = 303–304

@articlejolliffe2002, author = Jolliffe, I.T., title = Principal Component Analysis, journal = Springer Series in Statistics, year = 2002

@articleyeung2001, author = Yeung, K.Y. and Ruzzo, W.L., title = Principal component analysis for clustering gene expression data, journal = Bioinformatics, year = 2001, volume = 17, number = 9, pages = 763–774

@articleleek2010, author = Leek, J.T. and Storey, J.D., title = The importance of adjusting for hidden confounders in gene expression studies, journal = PLoS Genetics, year = 2010, volume = 6, number = 9, pages = e1000970

@articlewold1987, author = Wold, S. and Esbensen, K. and Geladi, P., title = Principal component analysis, journal = Chemometrics and Intelligent Laboratory Systems, year = 1987, volume = 2, pages = 37–52

@articleringner2013, author = Ringnér, M., title = How to use PCA for exploration of high-dimensional data, journal = Nature Methods, year = 2013, volume = 10, pages = 641–642

@articlegoh2017, author = Goh, W.W.B. and Wang, W. and Wong, L., title = Why batch effects matter in omics data, and how to avoid them, journal = Trends in Biotechnology, year = 2017, volume = 35, number = 6, pages = 498–507