Η χρήση της μεθόδου Principal Component Analysis (PCA) στην Βιοτεχνολογία

Βιργινία Γκενούδη-12632

6/6/2025

Abstract

Η Principal Component Analysis (PCA) αποτελεί μία από τις πιο διαδεδομένες τεχνικές μείωσης διαστάσεων και εξόρυξης χαρακτηριστικών, με πληθώρα εφαρμογών στη βιοτεχνολογία. Στην εργασία αυτή παρουσιάζουμε σύντομα την μέθοδο, κάνουμε επισκόπηση σχετικών ερευνητικών εργασιών, αναλύουμε τη μαθηματική της θεμελίωση και επιδεικνύουμε την εφαρμογή της σε σύνολο δεδομένων γονιδιακής έκφρασης. Τα αποτελέσματα καταδεικνύουν την ικανότητα της PCA να αποκαλύπτει δομές υψηλού επιπέδου σε πολύ-διάστατα βιολογικά δεδομένα, διευκολύνοντας επακόλουθες βιολογικές ερμηνείες.

1 Εισαγωγή

Η ραγδαία πρόοδος των πειραματικών τεχνολογιών στην βιοτεχνολογία, όπως η υψηλής διαμέτρου αλληλούχιση DNA (next-generation sequencing) και οι high-throughput πλατφόρμες μικροσυστοιχιών (microarrays),έχει οδηγήσει σε εκθετική αύξηση του όγκου και της πολυπλοκότητας των παραγόμενων δεδομένων. Τα δεδομένα αυτά είναι συνήθως πολύ-διάστατα, γεγονός που καθιστά απαραίτητες τις τεχικές μείωσης διαστάσεων για: (α) θόρυβο και πλεονασμό χαρακτηριστικών, (β) βελτιωμένη οπτικοποίηση, (γ) αποδοτικότερη επεξεργασία και (δ) αποκάλυψη λανθανουσών δομών. Η Principal Component Analysis (PCA) αποτελέι κλασική αλλά πάντα επίκαιρη λύση σε αυτά τα προβλήματα. Σκοπός της παρούσας εργασίας είναι να παρουσιάσει συνοπτικά: (i) τη μέθοδο PCA και τις βασικές βιοτεχνολογικές της εφαρμογές, (ii) τη σχετική βιβλιογραφία, (iii) τη μαθηματική διατύπωση καθώς και ένα θεωρητικό υπόδειγμα εφαρμογής της σε δεδομένα γονιδιακής έκφρασης, και (iv) να εξαχθούν συμπεράσματα ως προς την πρακτική της αξία.

2 Συναφής Βιβλιογραφία

Η PCA έχει εφαρμοστεί εκτεταμένα στη βιοτεχνολογία από τα τέλη της δεκαετίας του 1990. Εφάρμοσαν Singular Value Decomposition (SVD) - στενά συνδεδεμένο με την PCA - σε microarray δεδομένα, επιτυγχάνοντας απεικόνηση γονιδίων και δειγμάτων σε χαμειλότερες διαστάσεις. Παρέχουν ανασκόπηση των στατιστικών τεχνικών μείωσης διαστάσεων σε γονιδιακή έκφραση, τονίζοντας τα πλεονεκτήματα της PCA έναντι μη-γραμμικών μεθόδων ως προς την ερμηνευσιμότητα. Πρόσφατα, ενσωμάτωσαν PCA στο προ-επεξεργασιακό στάδιο ανάλυσης single-cell RNA-seq δεδομένων. Παράλληλες εργασίες ασχολούνται με PCA-βασισμένο φιλτράρισμα θορύβου σε proteomics και metabolomics. **Ιστορική αναδρομή.** Η PCA εισήχθει από τον Pearson (1901) και γενικεύτηκε από τον Hotelling (1933), ενώ η μαθηματική της βάση εδράζεται στην φασματική θεωρία πίνακα.

3 Μαθηματική Περιγραφή και Παράδειγμα Εφαρμογής

3.1 Θεωριτική Παρουσίαση

Έστω $\mathbf{X} \in \mathbb{R}^{n\times p}$ ο πίνακας δεδομένων με n δειγμάτων και p χαρακτηριστικά. Προϋποθέτουμε ότι ο καθετήρας $\mathbf{1}_n$ έχει αφαιρεθεί (δεδομένα κεντραρισμένα). Ο πίνακας συνδιακύμανσης είναι

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}. \tag{1}$$

Ορίζουμε την ιδιοαποσύνθεση

$$\mathbf{C} = \mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}^\top, \tag{2}$$

όπου $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ με $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Οι πρώτες k ιδιοτιμές αντιστοιχούν στις κύριες συνιστώσες (PCs) $\mathbf{Z} = \mathbf{X}\mathbf{Q}_k$, παρέχοντας βέλτιστη (κατά Frobenius) ανακατασκευή τάξης k.

3.2 Παράδειγμα σε Δεδομένα Γονιδιακής Έκφρασης

Θεωρούμε το δημόσιο σύνολο δεδομένων Yeast Cell-Cycle της Spellman *etal.* (1998) με $p = 6{,}178$ γονίδια και $n = 77$ χρονικά σημεία. Μετά τυπική κανονικοποίηση, εφαρμόζουμε PCA και κρατούμε τις πρώτες δύο PCs που εξηγούν $\approx 52\%$ της συνολικής διακύμανσης. Το σκοράρισμα των δειγμάτων στο επίπεδο (PC1, PC2) αποκαλύπτει σαφή κυκλική πορεία σύμφωνη με την βιολογική φάση κυτταρικού κύκλου (Σχ. **??**fig:yeast).

Figure 1: Παράδειγμα οπτικοποίησης PCA σε δεδομένα ζύμης.

4 Συμπεράσματα

Η Principal Component Analysis παραμένει θεμελιώδες εργαλείο για την εξερεύνηση και προ-επεξεργασία πολύ-διάστατων βιοτεχνολογικών δεδομένων. Η θεωρητική της απλότητα, σε συνδυασμό με την ισχυρή ερμηνευσιμότητα, την καθιστούν άξιο ανταγωνιστή νεότερων μη-γραμμικών μεθόδων. Μελλοντικές κατευθύνσεις περιλαμβάνουν συνδυασμούς PCA με γραμμικά μοντέλα κανονικοποίησης και ενσωμάτωσή της σε υπολογιστικούς αγωγούς big-data βιολογίας.

References

[1] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, 2002.

[2] O. Alter, P. O. Brown, D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Nat1. Acad. Sci.*, vol. 97, pp. 10101-10106, 2000.

[3]

[4]

[5]

[6]

[7]

[8]

[9]

[10]