Singular Value Decomposition for genome-wide expression data processing and modeling

Orly Alter, Patrick O. Brown and David Botstein
Why?

To approach any problem we formulate a hypothesis

Hypothesis:
- Genes reflect the state of the cell (and thereby the phenotype)
- Genes are classified into groups of similar regulation and function
How?

• What we have?
  gene expression data in the form of micro array
• What do we do with this?
  Clustering, Bi-clustering
• Why care about SVD?
  Data space is diagonalized
  Each eigen gene expressed only in the corresponding eigen array
  Dimensionality reduction
SVD

• $A = USV^T$
• $A$ (n by m) is any rectangular matrix (n rows and m columns)
• $U$ (n by m) is an “orthogonal” matrix
• $S$ (m by m) is a diagonal matrix
• $V$ (m by m) is another orthogonal matrix
• Such decomposition always exists
• All matrices are real; $n \geq m$
\[ A = USV^T \]

- U is an “orthogonal” matrix (n ≥ m)
- Column vectors of U form an orthonormal basis for the column space of A: \( U^T U \)
- \( u_1, \ldots, u_m \) in \( U \) are eigenvectors of \( AA^T \)
  - \( AA^T = USV^T VSU^T = US^2 U^T \) - “Left singular vectors”
\[ A = USV^T \]

- V is an orthogonal matrix (m by m)

- Column vectors of V form an orthonormal basis for the row space of A: \[ V^TV = VV^T = I \]

- \( v_1, \ldots, v_m \) in V are eigenvectors of \( A^TA \)
  - \( A^TA = VSU^T \)
  - \( USV^T = VS^2 V^T \)
  - “Right singular vectors”
\[ A = U S V^T \]

- S is a diagonal matrix (n by n) of non-negative singular values

- Typically sorted from largest to smallest

- Singular values are the non-negative square root of corresponding eigenvalues of \( A^T A \) and \( A A^T \)
Is SVD same as PCA?

• U is the eigen vector matrix of $A^*A^T$
• But $A^*A^T$ is proportional to the covariance matrix of the genes
• The eigen vector matrix (U) of $A^*A^T$ gives the PCA of the genes
• $V^T$ is the eigen vector matrix of $A^T*A$
• But $A^T*A$ is proportional to the covariance matrix of the arrays
• The eigen vector matrix (V) of $A^T*A$ gives the PCA of the arrays
Apply SVD to microarray data of yeast cell cycle
Apply SVD to microarray data
(elutriation synchronized cell cycle of yeast)

• 5981 genes of yeast observed throughout its cell cycle
• 14 microarrays correspond to different stages in the yeast cell cycle
Apply SVD to microarray data
(elutriation synchronized cell cycle of yeast)

Each eigen gene (V) is expressed only in the corresponding eigen array (U) with the corresponding eigen expression value (S)
• Eigen genes and eigen arrays are orthonormal superpositions of the genes and arrays
• Expression of each eigen gene is decoupled and decorrelated from that of all other eigen genes
Filtering

• Filter out the noisy eigen genes and normalize the data

• Noise occurs due to uncontrolled experimental conditions

• To filter out an eigen gene – make its corresponding eigen expression 0
Normalization

- Adjusting values in a micro array experiment to improve consistency and reduce bias
- Normalized dataset centered around the zero arithmetic mean
Fig. 1. Normalized elutriation eigengenes. (a) Raster display of $\Phi_n^T$, the expression of 14 eigengenes in 14 arrays. (b) Bar chart of the fractions of eigenexpression, showing that $|\gamma_1)_n$ and $|\gamma_2)_n$ capture about 20% of the overall normalized expression each, and a high entropy $d = 0.88$. (c) Line-joined graphs of the expression levels of $|\gamma_1)_n$ (red) and $|\gamma_2)_n$ (blue) in the 14 arrays fit dashed graphs of normalized sine (red) and cosine (blue) of period $T = 390$ min and phase $\theta = 2\pi/13$, respectively.
• Variable expression of the first and the second eigen gene (principal components) across 14 arrays fit the graphs of sine and cosine wave with $T = 390$ min
Data Sorting

• Given two eigen arrays calculate the correlation of each of the 14 eigen arrays with both the eigen array 1 and eigen array 2.
• Plot the correlation of eigen array 1 vs correlation of eigen array 2 (dimensionality reduction) for each of the 14 arrays
• Reduced dimensionality
Distance from the origin is the amplitude of expression

- The angular distance gives the degree of correlation
• We know different arrays capture different time intervals during the yeast cell cycle – They represent the changing cellular states during the cell cycle.
• We also know each eigen gene is expressed only in the corresponding eigen array.
• Can we somehow relate this to genes?
• Compute the correlation of all 5981 genes with the top 2 significant eigen genes and plot it the same way.
Yeast cell cycle
• We now know the genes that are expressed during different cell cycle stages
• How about the expression levels of the other genes during the cell cycle stages?
Normalized eigen arrays

Fig. 3. Genes sorted by relative correlation with \(|\gamma_1|_N\) and \(|\gamma_2|_N\) of normalized elutriation. (a) Normalized elutriation expression of the sorted 5,981 genes in the 14 arrays, showing travelling wave of expression. (b) Eigenarrays expression; the expression of \(|\alpha_1|_N\) and \(|\alpha_2|_N\), the eigenarrays corresponding to \(|\gamma_1|_N\) and \(|\gamma_2|_N\), displays the sorting. (c) Expression levels of \(|\alpha_1|_N\) (red) and \(|\alpha_2|_N\) (green) fit normalized sine and cosine functions of period \(Z = N - 1 = 5,980\) and phase \(\theta \sim 2\pi/13\) (blue), respectively.
• Normalized elutriation expression fits a sinusoidal traveling wave of expression both across genes and arrays

• The expression of the $n^{th}$ gene in the $m^{th}$ array satisfies 

$$-2\cos\left[2\pi\left(t/T-z/Z\right)\right]/\sqrt{ZT}$$

where $Z= N-1$ and $z = n - 1$
Conclusion

• We are able to infer the set of genes that are expressed during different stages of the yeast cell cycle
• This helps to infer the functional aspects of the genes
• In the broader perspective groups of genes having similar regulation and function reflect the state of the cell