Unsupervised Data Mining – SVD

February 26, 2025

Lecture Objectives:

1. **Foundations of SVD**
   - Understand factorization of data matrix $\mathbf{A}$ into (U, S, V^T) and the mathematical relationship to eigendecomposition
   - Interpret the meaning of left/right singular vectors and singular values in terms of row/column spaces
   - Explain SVD as an unsupervised learning method for dimensionality reduction in high-throughput/ high dimensional data
2. **Applications of SVD**
   - Apply SVD for data compression via low-rank approximation
   - Show how SVD preserves essential information while reducing complexity
3. **SVD in Biological Data**
   - Interpret eigenarrays and eigengenes to identify major expression patterns
   - Apply SVD to uncover cyclical biological processes like cell-cycle phases
   - Explore real-world applications in gene expression analysis

Key Concepts and Definitions

- **Data Matrix (A)** A rectangular array of data. Rows and columns often represent different entities(e.g., genes × samples in our case).
- **Singular Value Decomposition (SVD)** A matrix factorization technique where a data matrix $\mathbf{A}$ is decomposed into $\mathbf{A} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$. It reveals key "directions" (patterns) in both row and column spaces.
- **Left Singular Vectors (U)** The columns of $\mathbf{U}$; form an orthonormal basis for the **column space** of $\mathbf{A}$. Interpreted as "eigen arrays" (if columns are samples/conditions).
- **Right Singular Vectors (V)** The columns of $\mathbf{V}$; form an orthonormal basis for the **row space** when transposed ($\mathbf{V}^T$). Interpreted as "eigen genes" (if rows are genes).
- **Singular Values (S)** The diagonal entries in $\mathbf{S}$. They are non-negative numbers sorted in descending order, indicating the "importance" (variance captured) of each singular vector pair.
- **Rank** Represents the dimension of the subspace spanned by the matrix's rows or columns and the maximum number of linearly independent rows or columns of the Data Matrix. Also, it's the number of non-zero singular values in the SVD of a matrix. The rank is always $r \leq \min(m, n)$, meaning if $m > n$, then the matrix can have rank at most n.
- **Row Space vs. Column Space**

  - **Row Space**: The space spanned by row vectors of $\mathbf{A}$.
  - **Column Space**: The space spanned by column vectors of $\mathbf{A}$.
- **Eigenvectors and Eigenvalues** In the context of SVD:
  - **Eigenvectors** (left or right) correspond to directions of maximum variance in row or column space.
  - **Eigenvalues** relate to the singular values (for $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$).
- $\mathbf{A}\mathbf{A}^T$ **and** $\mathbf{A}^T\mathbf{A}$
  - $\mathbf{A}\mathbf{A}^T$ is used to find eigenvectors in the column space (related to $\mathbf{U}$).
  - $\mathbf{A}^T\mathbf{A}$ is used to find eigenvectors in the row space (related to $\mathbf{V}$).
- **Dimensionality Reduction** The process of representing high-dimensional data with fewer variables (principal "directions" or components) while preserving most of the important variation.
- **Low-Rank Approximation** By truncating SVD to the top k singular values/vectors, one obtains the "best" rank-k approximation of $\mathbf{A}$. This is crucial for noise reduction and data compression.
- **Principal Component Analysis (PCA)** A closely related technique typically applied to covariance/correlation matrices. PCA can be viewed as performing an SVD on mean-centered data, where the principal components correspond to the singular vectors that capture the most variance.
- **Variance Captured** Each singular value σ_i (or σ_i^2 in PCA) represents the amount of variance in the data captured by its corresponding singular vector. Larger singular values indicate more important patterns in the data.
- **Orthonormal Vectors** Vectors that are mutually perpendicular and each have unit length ($\|\mathbf{u}\| = 1$). The columns in $\mathbf{U}$ and $\mathbf{V}$ are orthonormal bases.
- **Projection** Multiplying the data matrix $\mathbf{A}$ by a particular singular vector (e.g., $\mathbf{v}_i$) to see how the data "projects" onto that direction. Used to interpret major trends or groups.
- **Eigen Arrays / Eigen Genes** Terminology used in genomics for the left and right singular vectors, showing how genes or samples group/cluster together along principal directions of variation.
- **Cell Cycle** The series of events that take place in a cell leading to its division and duplication.
- **Gene Expression Data** Quantification of the activity (expression) of genes under various time frames or conditions.

1 Main Content

1.1 Introduction to SVD for Dimensionality Reduction

High-dimensional datasets present significant challenges in modern data analysis, particularly in biological contexts. With numerous features (genes, variables), visualizing and extracting meaningful structure becomes difficult using standard clustering methods alone. Dimensionality reduction offers a solution by creating simpler data representations while preserving essential information.

SVD is a fundamental linear algebra technique that decomposes an $m \times n$ data matrix $\mathbf{A}$ into three matrices capturing major variation sources in both rows and columns simultaneously. This approach effectively compresses data and reveals underlying patterns such as time-series progressions or biological processes (e.g., cell-cycle phases in yeast).

1.2 The SVD Factorization

At the heart of SVD is the factorization of the data matrix $\mathbf{A}$ into three components:

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m}\,\mathbf{S}_{m\times n}\,\mathbf{V}^T_{n\times n} \tag{1}$$

More explicitly, this can be written as:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \ddots & & \vdots \\ x_{m1} & \cdots & \cdots & x_{mn} \end{pmatrix}_{m\times n} = \begin{pmatrix} u_{11} & \cdots & u_{m1} \\ \vdots & \ddots & \vdots \\ u_{1m} & \cdots & u_{mm} \end{pmatrix}_{m\times m} \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n \\ 0 & \cdots & 0 \end{pmatrix}_{m\times n} \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nn} \end{pmatrix}_{n\times n} \tag{2}$$

Where:

- $\mathbf{A}$ is any rectangular matrix of size $m \times n$ (with $m \geq n$ in our example). In genomics, m might be the number of genes, and n might be the number of experiments or time points.
  - **Row space:** The vector subspace generated by the row vectors of $\mathbf{A}$.
  - **Column space:** The vector subspace generated by the column vectors of $\mathbf{A}$.
  - The dimension of both the row and column space is the rank of matrix $\mathbf{A}$: r (where $r \leq n$)
  - $\mathbf{A}$ represents a linear transformation that maps a vector $\mathbf{x}$ in row space into vector $\mathbf{Ax}$ in column space.
- $\mathbf{U}$ is an "orthogonal" matrix ($m \geq n$) whose columns $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ form an orthonormal basis for the **column space** of $\mathbf{A}$: $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Only the first r columns (corresponding to non-zero singular values) are needed, therefore we have columns indexed up to n.

$$U = \begin{pmatrix} | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \\ | & | & & | \end{pmatrix} \tag{3}$$

  - The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ in $\mathbf{U}$ are eigenvectors of $\mathbf{AA}^T$
  - $\mathbf{AA}^T = \mathbf{USV}^T\mathbf{VSU}^T = \mathbf{US}^2\mathbf{U}^T$
  - These are called "Left singular vectors" and in our genomic example referred to as "eigen arrays"
- $\mathbf{V}$ is an orthogonal matrix ($n \times n$) whose columns $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ form an orthonormal basis for the **row space** of $\mathbf{A}$: $\mathbf{V}^T\mathbf{V} = \mathbf{VV}^T = \mathbf{I}$.

$$V = \begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & & | \end{pmatrix} \tag{4}$$

  - The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$
  - $\mathbf{A}^T\mathbf{A} = \mathbf{VSU}^T\mathbf{USV}^T = \mathbf{VS}^2\mathbf{V}^T$
  - These are called "Right singular vectors" and in our genomic example referred to as "eigen genes"

- $\mathbf{S}$ is an $m \times n$ diagonal matrix containing the singular values ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$) in descending order. The singular values correspond to the magnitude or importance of each principal direction.
  - The singular values can be represented as:

$$\begin{pmatrix} \sqrt{\lambda_1} & & & & \\ & \sqrt{\lambda_2} & & & \\ & & \ddots & & \\ & & & \sqrt{\lambda_r} & \\ & & & & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ & & & & 0 \end{pmatrix} \tag{5}$$

- The singular values σ_i are arranged in descending order: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$
- Each singular value σ_i corresponds to a specific left singular vector $\mathbf{u}_i$ and right singular vector $\mathbf{v}_i$
- The singular values are the square roots of the eigenvalues (λ_i) of $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$

- **Key properties:** Both $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns, ensuring singular vectors capture unique directions of variation. Matrix $\mathbf{A}$ functions as a linear transformation mapping vector $\mathbf{x}$ from row space to $\mathbf{A}\mathbf{x}$ in column space.

1.3 Low-Rank Approximation and Dimensionality Reduction

SVD decomposes a matrix as $A = USV^T$ with singular values $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$. This outer product uv^T gives a matrix rather than the scalar of the inner product.

The SVD decomposition can be written as:

$$A = s_1\mathbf{u}_1\mathbf{v}_1^T + s_2\mathbf{u}_2\mathbf{v}_2^T + \ldots + s_n\mathbf{u}_n\mathbf{v}_n^T \tag{6}$$

The rank-r matrix $\hat{A}$ that best approximates A is obtained by truncating this sum:

$$\hat{A} = s_1\mathbf{u}_1\mathbf{v}_1^T + s_2\mathbf{u}_2\mathbf{v}_2^T + \ldots + s_r\mathbf{u}_r\mathbf{v}_r^T \tag{7}$$

This minimizes the least squares error:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(A_{ij} - \hat{A}_{ij})^2 \tag{8}$$

When $r = 1$, this corresponds to a line fit. This approach is very useful for matrix approximation and data compression.

An important property of SVD relates to how the matrix acts on the singular vectors:

$$A\mathbf{v}_i = s_i\mathbf{u}_i \tag{9}$$

1.4 Geometry of SVD in Row Space

Geometrically, we can view A as a collection of m row vectors (points) in the row space of A. The term $s_1\mathbf{u}_1\mathbf{v}_1^T + s_2\mathbf{u}_2\mathbf{v}_2^T$ represents the best rank-2 matrix approximation for A.

In this geometric interpretation:

- Vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are the directions of the best approximating rank-2 subspace that passes through the origin
- $s_1\mathbf{u}_1$ and $s_2\mathbf{u}_2$ give coordinates for row vectors in the rank-2 subspace
- $\mathbf{v}_1$ and $\mathbf{v}_2$ provide coordinates for row space basis vectors in the rank-2 subspace

This is captured by the equations:

$$A\mathbf{v}_i = s_i\mathbf{u}_i \tag{10}$$

$$I\mathbf{v}_i = \mathbf{v}_i \tag{11}$$

1.5 Connection to Principal Component Analysis (PCA)

PCA is typically performed on a covariance matrix (often $\mathbf{A}^T\mathbf{A}$ if rows are mean-centered). Mathematically, PCA is almost the same as SVD applied to the mean-centered data matrix. Both methods yield principal directions of maximum variance, but SVD is more general because it can be applied directly to any rectangular matrix, without requiring the explicit covariance calculation.

From Genes × Arrays to Eigengenes × Eigenarrays

In genome-wide microarray experiments, each element (i, j) of a data matrix $\mathbf{A}$ represents the measured expression level of gene i under experimental condition (or time point) j. In some genomic contexts, rows represent genes and columns represent experimental conditions (arrays), in this context columns of $\mathbf{U}$ is called "**eigenarrays**," reflecting variation across genes, whereas the columns of $\mathbf{V}$ are called "**eigengenes**,". **Singular Value Decomposition (SVD)** factors this matrix as

$$\mathbf{A} = \underbrace{\mathbf{U}}_{\text{(eigenarrays)}} \quad \underbrace{\mathbf{S}}_{\text{(singular values)}} \quad \underbrace{\mathbf{V}^{\mathsf{T}}}_{\text{(eigengenes)}}, \tag{12}$$

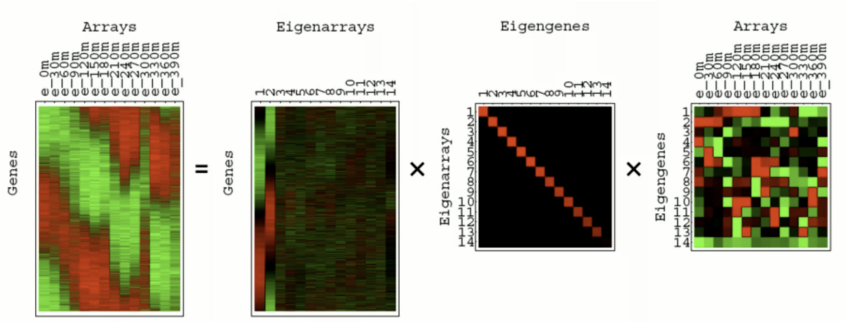

Figure 1: Visualization of SVD applied to gene expression data. The original matrix $\mathbf{A}$ is decomposed into the product of three matrices: $\mathbf{U}$ (eigenarrays), $\mathbf{S}$ (singular values), and $\mathbf{V}^{\mathsf{T}}$ (eigengenes). The heatmap patterns illustrate how the decomposition captures the fundamental structure of the data.

where:

1. **U (Eigenarrays)** The columns of $\mathbf{U}$ form an orthonormal basis for the *column space* of $\mathbf{A}$. In the context of gene expression, each column captures a "characteristic" pattern of how *all genes* behave together across the experimental conditions; these are often termed *eigenarrays* because they reflect common modes of variation in the columns of $\mathbf{A}$.
2. **S (Singular Values)** This diagonal matrix encodes the strength (magnitude) of each pattern uncovered by SVD. The largest singular value σ_1 corresponds to the strongest or most dominant axis of variation in the dataset, followed by σ_2, and so on.
3. **V (Eigengenes)** The columns of $\mathbf{V}$ form an orthonormal basis for the *row space* of $\mathbf{A}$ (when transposed, $\mathbf{V}^{\mathsf{T}}$ maps genes to these "principal directions"). Here, each *right singular vector* can be interpreted as a meta-gene or *eigengene* a weighted combination of the original genes that captures a distinct expression pattern across samples.

- **Eigenarrays** group arrays or experimental conditions that share similar global expression responses.
- **Eigengenes** cluster genes that share similar response profiles across those conditions.
- **Singular values** quantify the importance of each pair of eigenarray/eigengene components.

Because $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, SVD guarantees these derived patterns are non-redundant and capture uncorrelated sources of variation. In biological analyses, focusing on the top eigengenes (and eigenarrays) often reveals major biological signals (e.g., cell-cycle phases, disease states, treatment effects), while lower-rank components may correspond to noise or minor effects. This decomposition thereby acts as a powerful tool for **data reduction**, **noise filtering**, and **interpretation** of large-scale genomic experiments.

1.6 Examples: Yeast Cell-Cycle Analysis

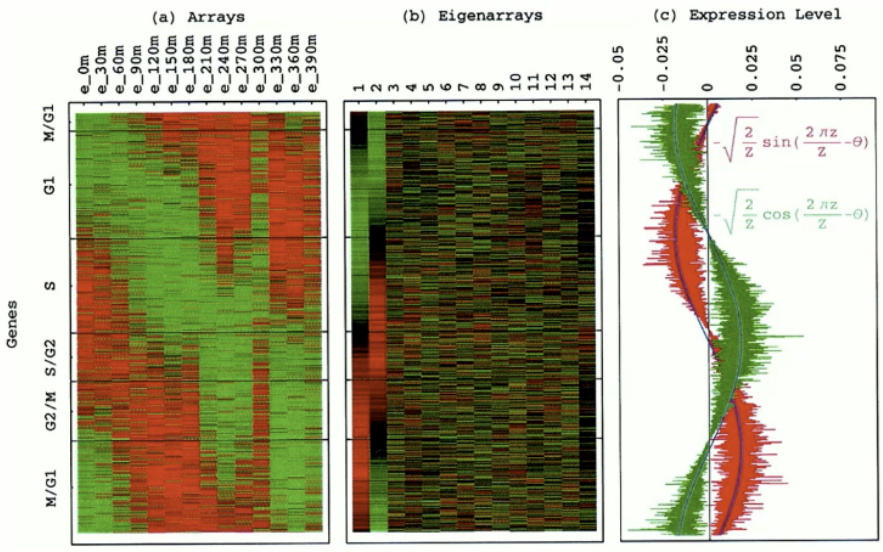

Figure 2: (a) Heatmap shows normalized gene expression data, reordered by each gene's correlation with the top two oscillatory **eigengenes** after removing the steady-state component. These two eigengenes capture the progression through the cell cycle (M/G1 $\rightarrow$ G1 $\rightarrow$ S $\rightarrow$ S/G2 $\rightarrow$ G2/M). (b) The corresponding **eigenarrays**, which indicate how strongly each array (time point) projects onto these oscillatory modes. (c) The sine/cosine-like expression patterns have period $Z = N - 1 = 5,980$ and phase $\theta \approx 2\pi/13$, confirming the periodic nature of the underlying biological process.

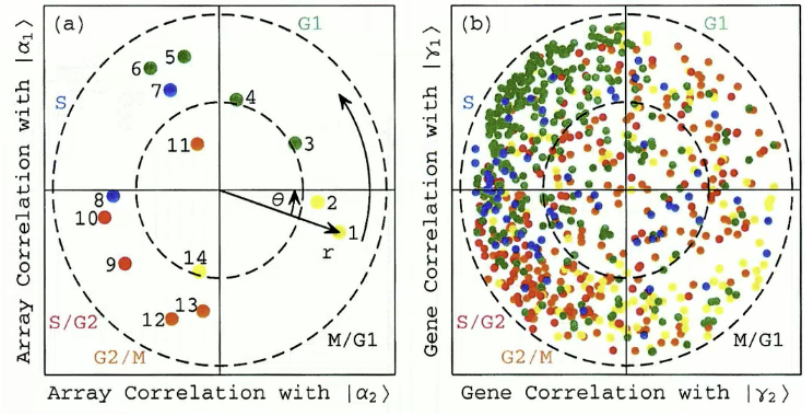

Figure 3: **(a)** Each array (time point) is plotted by projecting its expression vector onto the top two **eigenarrays**, α_1 and α_2. The horizontal and vertical axes show the correlations (dot products) with α_2 and α_1, respectively. The radial distance from the origin (dashed circles) indicates how much of the array's expression is captured by these two principal SVD modes, while the angular coordinate corresponds to its "phase" in the cell-cycle progression. The numeric labels are the time points, and the colors denote the assigned cell-cycle stages. **(b)** Each gene is likewise placed according to its correlation with the top two **eigengenes**, γ_1 and γ_2, color-coded by the cell-cycle phase in which that gene's expression peaks. Together, these panels show that the first two oscillatory modes (after removing the steady-state component) neatly capture the cyclical behavior of both arrays (time points) and genes, tracing a circular pattern that reveals the progression through cell phases

- **Data**: Matrix $\mathbf{A}$ with genes (rows) measured across cell cycle timepoints (columns)
- **SVD**: $\mathbf{A} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$ reveals temporal expression patterns
- **Results**: Top components typically reveal circular trajectory matching cyclical cell stages, with co-regulated genes clustering together

Applying SVD

- The decomposition $\mathbf{A} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$ separates out core patterns in how genes are expressed over time. For instance, certain sets of genes "activate" earlier or later in the cycle, forming distinct clusters in the projected space.

Visualization

- Plotting the first two singular vectors (or the principal components) often reveals a circular trajectory in the data, matching how the cell cycle proceeds in cyclical stages.
- Genes that are co-regulated cluster together, showing up in the same quadrant or region when colored by functional annotation.

This example illustrates how SVD can capture the cyclic nature of biological processes and help researchers identify groups of genes with similar temporal expression profiles.

Discussion/Comments

- The slides explains how valuable SVD is at uncovering underlying correlations and patterns in intricate biological datasets.
- By studying their relationship to the eigenarrays and eigengenes, SVD can be used to understand genes that are co-regulated or involved in related biological processes.
- It can be especially helpful for researching biological events that occur on a regular basis.
- By reducing the dimensionality of gene expression data, SVD can facilitate analysis and interpretation.
- Here, we use naming convention (i.e., U = eigenarrays, V= eigengenes), so readers should be aware that different sources may flip this terminology depending on how they label the rows and columns of the data matrix.
- Leveraging SVD with appropriate visualization methods highlights patterns that are difficult to observe directly in high-dimensional data.
- The orthonormal structure of the singular vectors ensures each extracted pattern is independent, making it easier to isolate and interpret specific sources of biological variation.
- Overall, SVD is a robust unsupervised method for denoising, identifying key expression trends, and guiding hypothesis generation in large-scale biological studies.

References ISL/ESL

This lecture dives deep into the statistical portion of singular value decomposition (SVD) for genome-wide expression. Here are the following suggested readings that can help solidify the foundational framework of SVD:

ISL: Chapter 6.3.1 [Up to section on "The principal Components Regression Approach"] provides a comprehensive overview of principal components analysis (PCA) with specific examples and supporting figures to understand the concepts. Chapter 12.2 [gives background on PCA/SVD] also provides an informative overview of PCA with focus on the principal components and its interpretations. I would suggest looking into chapter 12.3 [missing values and matrix completion] which provide additional

details on utilizing principal components to impute missing values in the dataset or matrix which is common in real-world settings.

Another important material is a paper called "Singular value decomposition for genome-wide expression data processing and modeling" is where most of the materials within the lecture are referenced from. The paper contains additional information and details within the SVD process for analyzing genome-wide expression data. Feel free to analyze the paper to gain further insight if any portion of the lecture is confusing.

Overall, these are great materials to review in support of SVD that was covered in class and may help in digesting the mathematical framework.

Other suggested references

- For more details of SVD: Mathematical Modeling of Biological Systems, Volume 1: Modeling and Simulation in Science, Engineering, and Technology portion: `https://link.springer.com/chapter/10.1007/978-0-8176-4558-8_32`
- Example within the application of SVD: Pathway level analysis of gene expression using singular Value decomposition: `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-225`
- Recent area for SVD: Federated singular value decomposition for high-dimensional data: `https://link.springer.com/article/10.1007/s10618-023-00983-z`