Matrix factorizations for dimensionality reduction

Troy Whitfield

Whitehead Institute
Bioinformatics and Research Computing

March 18, 2021

Outline

1. **Introduction**
   - Contemporary applications in biology
2. **Eigenvalue decomposition**
   - Characterizing molecular motions
   - Important properties
3. **Singular value decomposition**
   - Gene expression processing
   - Statement of factorization
   - Data filtering and compression
4. **Non-negative matrix factorization**
   - Definition and examples
5. **References**

Brief historical background

- Elements of eigenvalue problems and their solution date to the 18th and 19th centuries in the work of Euler, Lagrange, Cauchy and others.
- Analysis leading to the singular value theorem dates from the late 19th and early 20th centuries.
- Efficient numerical methods for computing the eigenvalue and singular value decompositions (SVD) date to the mid-20th century.
- Non-negative matrix factorization (NMF) is a more recent technique that can be usefully applied in a variety of contexts. Lee and Seung's seminal paper on this method [Lee and Seung, 1999] had some late 20th century antecedents.
- All of these linear methods can be used to discover patterns in high dimensional data-sets like those from high-throughput biological experiments.

Biological applications

- SVD ($\mathbf{A} = \mathbf{U\Sigma V}^T$) is often used to study genome-wide expression data [Alter *et al.*, 2000] including those from single cell experiments.
- Eigenvalue decomposition ($\mathbf{A} = \mathbf{Q\Lambda Q}^{-1}$) can be used to understand biomolecular motions and solve problems in ecology and evolution, among others.
- Non-negative matrix factorization ($\mathbf{A} \approx \mathbf{WH}$, subject to positivity constraints) can provide features (e.g. combinations of genes) that may be more readily interpretable compared with those of SVD [Brunet *et al.*, 2004] and has been adapted for analysis of multi-omics data [Zhang *et al.*, 2012, Yang and Michailidis, 2016].

Application of eigenvalue decomposition: NMA

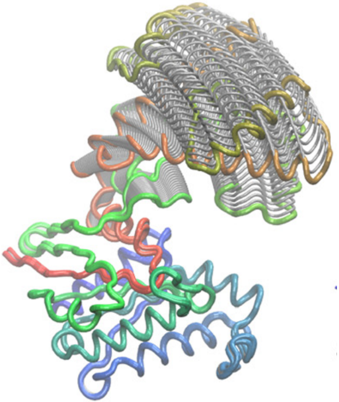

Fig. 1: *Normal mode analysis of a predicted open form of Gα, a heterotrimeric G protein [Skjaerven* et al.*, 2014]. The normal modes capture large scale domain motion.*

Eigenvalue decomposition for a matrix A

Given a diagonalizable square $n \times n$ matrix $\mathbf{A}$, its eigenvalue decomposition is

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1},$$

where the columns of $\mathbf{Q}$ and the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvectors and eigenvalues of $\mathbf{A}$ (i.e. $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$).

- Not the same as SVD, but related.
- $\mathbf{Q}$ is orthonormal.
- Directly relevant to a subset of biological problems, but also indirectly relevant for $n \times p$ rectangular data matrices.

Motivating example: analysis of gene expression

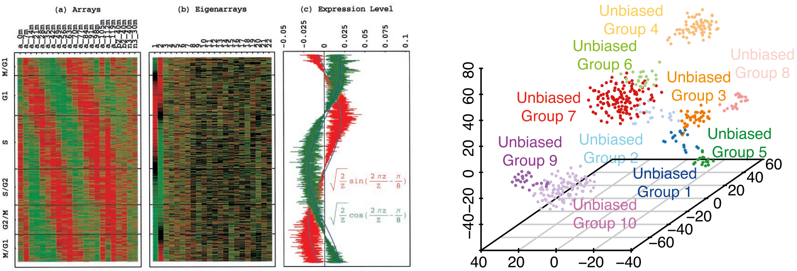

Fig. 2: *On the left are traveling waves in yeast gene expression, shown by correlation of individual gene expression with that of the two leading eigengenes [Alter* et al., *2000]. Right, clusters of cells appear in the principal component space of single cell RNA-seq data from primary cortex samples [Darmanis* et al., *2015].*

PCA for a matrix A

For a rectangular $n \times p$ matrix $\mathbf{A}$, the singular value decomposition is

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T,$$

where the non-zero elements of the diagonal matrix $\boldsymbol{\Sigma}$ are the singular values, $\mathbf{U}$ is a $n \times n$ unitary matrix in the column space of $\mathbf{A}$ and $\mathbf{V}$ is a $p \times p$ unitary matrix in the row space of $\mathbf{A}$. Since $\mathbf{U}$ and $\mathbf{V}$ are unitary (or orthonormal for a real matrix $\mathbf{U}$, i.e. $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = I$), it follows that

$$\begin{aligned} \mathbf{A}\mathbf{A}^T &= \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^T, \\ \mathbf{A}^T\mathbf{A} &= \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T. \end{aligned}$$

If the matrix $\mathbf{A}$ is *centered*, the singular value decomposition can be used to compute the *principal components* for a covariance matrix of the data.

Some properties and observations

- Formally equivalent to eigenvalue decomposition on covariance matrices.
- Orthonormal singular vectors.
- The sequence of singular values is unique. If these values are all distinct, the singular vectors are also unique except for a phase factor of ± 1.
- For gene expression data, this is a linear transformation from genes × sample space to a reduced "eigengene" × "eigensample" space [Alter *et al.*, 2000].

Graphical representation of the decomposition

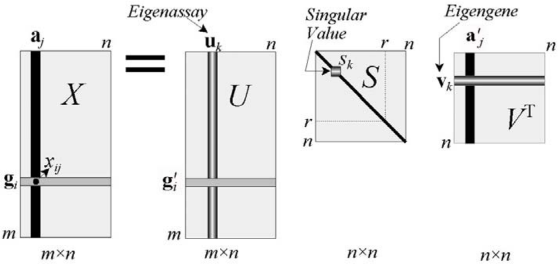

Fig. 3: *Graphical depiction of SVD in the context of gene expression data [Wall* et al.*, 2003].*

Explained variance by each singular value

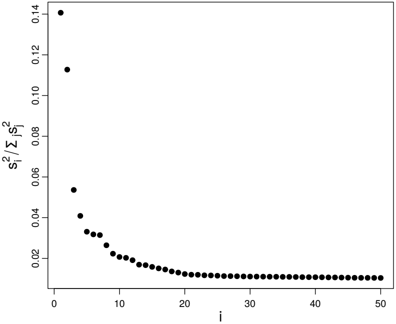

Fig. 4: *By convention, the singular values are ordered. Above, the explained variance for SVD/PCA on a large integrated data-set from single cell RNA-seq experiments is shown. How can this spectrum guide our choices in reducing the dimensionality of the data-set?*

SVD/PCA can aid in classification

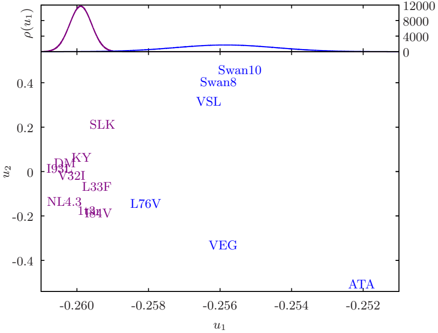

Fig. 5: *HIV-1 protease variants projected on top two principal components of the correlation matrix. The upper margin shows that the distribution of u_1 is well described by a Gaussian mixture model with two components of variable width. Colors reflect membership in the mixture.*

SVD yields optimal lower rank approximations

The rank of a $n \times p$ matrix $\mathbf{A}$ is at most $\min(n, p)$, but we can use the SVD to get an even lower rank approximation. Rewriting the decomposition as the sum of rank-1 matrices

$$\mathbf{A} = \sum_{i=1}^{\min(n,p)} s_i \mathbf{u}_i \mathbf{v}_i^T,$$

we can terminate the sum for $k < \min(n, p)$. An important property of the SVD is that this represents the closest rank-k approximation to the original matrix (Eckart–Young theorem). This approximation can be useful for removing noise and compressing the data.

Matrix norms and reconstruction error

We'd like a way to measure how accurately a rank $k < r$ approximation is able to recover the original data matrix $\mathbf{A}$. A matrix norm, which is a way to measure the size of a matrix that's analogous to the dot product for vectors, serves this need. One such norm is the Frobenius norm

$$\|\mathbf{A}\|_F^2 = |a_{11}|^2 + |a_{12}|^2 + \cdots + |a_{np}|^2 .$$

The reconstruction error for an approximation, $\mathbf{B}$, of $\mathbf{A}$ can then be defined as $\|\mathbf{A} - \mathbf{B}\|_F$.

Data compression example

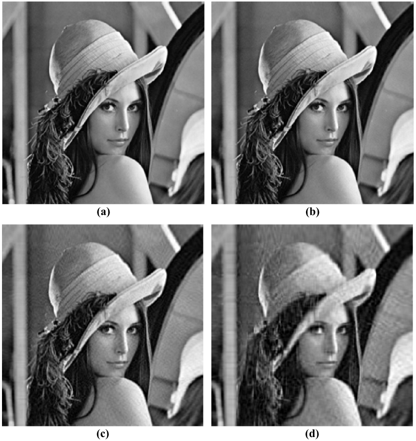

Fig. 6: *The original image* **(a)** *is 256* × *256 pixels. Reconstructions are based on 128* **(b)**, *64* **(c)** *and 32* **(d)** *singular values [Rufai* et al.*, 2014].*

Sample application: detection of tumor cellular programs

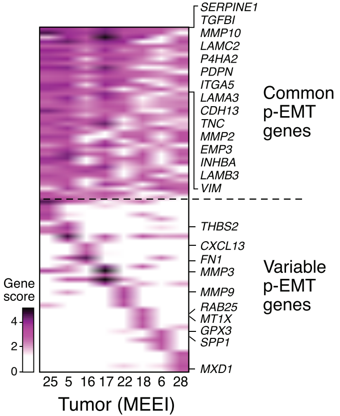

Fig. 7: *Heatmap of NMF scores for common (top rows) and tumor-specific (bottom rows) genes within the p-EMT program by tumor [Puram* et al.*, 2017].*

NMF for a data matrix A

Given a non-negative matrix **A**, factorize it as $\mathbf{A} \sim \mathbf{WH}$ by minimizing an error (here Frobenius norm):

$$\min_{W,H} ||\mathbf{A} - \mathbf{WH}||_F \text{ such that } \mathbf{W} > 0, \mathbf{H} > 0$$

- Defines an optimal rank k approximation to **A**.
- The non-negativity constraint replaces the orthogonality of SVD.
- Optimal factorizations are insensitive to scaling and rotation (i.e. not unique).
- Non-negativity constraint should lead to “parts-based”, interpretable factorizations.
- NMF yields clustering similar to k-means.

Graphical representation of the factorization

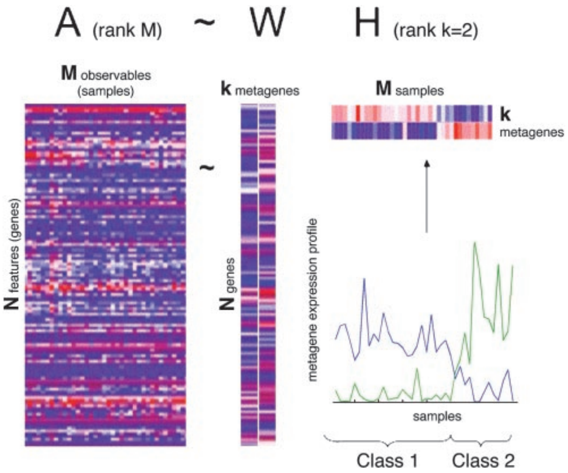

Fig. 8: *Graphical depiction of NMF in the context of gene expression data [Brunet* et al., *2004].*

Choice of rank in NMF

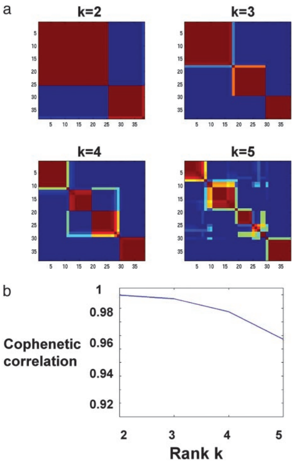

Fig. 9: *Behavior of cophenetic correlation can guide the selection of rank for NMF. Here, 38 bone marrow samples from AML, ALL T and ALL B subjects are analyzed [Brunet* et al., *2004].*

“Parts-based” representation of data in NMF

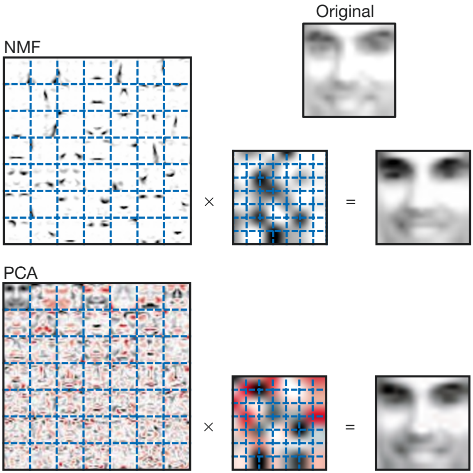

Fig. 10: *Unlike PCA (bottom), the NMF features (top) are interpretable as a set of component parts for facial reconstruction [Lee and Seung, 1999].*

Summary

- Matrix factorizations can provide good low-rank approximations to higher-dimensional data-sets.
- These factorizations can be readily interpretable.
- Can interpretability be further extended? Imposing sparsity constraints can help.
- For pattern discovery with high dimensional data, however, non-linear methods for dimensionality reduction (e.g. t-SNE, UMAP) may offer advantages.

References I

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling.
Proc. Natl. Acad. Sci. USA, **97** (18), 10101–10106.

Brunet,J.P., Tamayo,P., Golub,T.R. and Mesirov,J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization.
Proc. Natl. Acad. Sci. USA, **101** (12), 4164–4169.

Darmanis,S., Sloan,S.A., Zhang,Y., Enge,M., Caneda,C., Shuer,L.M., Hayden Gephart,M.G., Barres,B.A. and Quake,S.R. (2015) A survey of human brain transcriptome diversity at the single cell level.
Proc. Natl. Acad. Sci. USA, **112** (23), 7285–7290.

References II

- Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization.
  Nature, **401** (6755), 788–791.

- Puram,S.V., Tirosh,I., Parikh,A.S., Patel,A.P., Yizhak,K., Gillespie,S., Rodman,C., Luo,C.L., Mroz,E.A., Emerick,K.S., Deschler,D.G., Varvares,M.A., Mylvaganam,R., Rozenblatt-Rosen,O., Rocco,J.W., Faquin,W.C., Lin,D.T., Regev,A. and Bernstein,B.E. (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer.
  Cell, **171** (7), 1611.e1–1611.e24.

References III

- Rufai,A.M., Anbarjafari,G. and Demirel,H. (2014) Lossy image compression using singular value decomposition and wavelet difference reduction.
  Digital Signal Processing, **24**, 117–123.

- Skjaerven,L., Yao,X.Q., Scarabelli,G. and Grant,B.J. (2014) Integrating protein structural dynamics and evolutionary analysis with Bio3D.
  BMC Bioinformatics, **15** (1), 399.

- Wall,M.E., Rechtsteiner,A. and Rocha,L.M. (2003) Singular value decomposition and principal component analysis.
  In *A Practical Approach to Microarray Data Analysis*, (Berrar,D.P., Dubitzky,W. and Granzow,M., eds),. Kluwer Norwell, MA pp. 91–109.

References IV

Yang,Z. and Michailidis,G. (2016) A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data.
Bioinformatics, **32** (1), 1–8.

Zhang,S., Liu,C.C., Li,W., Shen,H., Laird,P.W. and Zhou,X.J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data.
Nucleic Acids Res., **40** (19), 9379–9391.