Matrix factorization

Why, when and how do we use them?

Ha Vu

07/17/2023

Contents

1 Singular value decomposition

1.1 How does it work?

SVD proposes to factorize $A_{n\cdot m} = U_{n\cdot n}\Sigma_{n\cdot m}V^T_{m\cdot m}$, such that:

- Columns of U are eigen vectors of AA^T
- Σ is a diagonal matrix with the diagonal entry being the squareroot of eigen values of A^TA, in decesding order
- Columns of V are eigen vectors of A^TA

1.2 Proof the existence of SVD for any matrix $A_{n\cdot m}$

I have based my proof on the blog post of Gregory Gundersen, who is also a student at Prof. Barbara Engelhardt's lab.

First, A^TA is a positive definite matrix, because $\mathbf{x}^TA^TA\mathbf{x} > 0\forall\mathbf{x} \neq \mathbf{0}$. (**Note**: if you wonder why with M symmetric and $\mathbf{x}^TM_{n\cdot n}\mathbf{x} > 0\forall\mathbf{x} \neq \mathbf{0} \Leftrightarrow M$ is positive definite. **Proof**: $\Leftarrow$, because M is symmetric by definition of positive definite, and hence $M = QSQ^T$, by existence of eigen value decomposition of symmetric matrix, and S is diagonal with all positive diagonal (also eigen) values, by definition of positive definite. Hence, $M = H^TH$, where $H = Q^T \cdot \sqrt{S}$, hence $\mathbf{x}^TM\mathbf{x} = \mathbf{x}^TH^TH\mathbf{x} > 0\forall\mathbf{x} \neq \mathbf{0}$. **Proof**: $\Rightarrow$, similarly, starting with M being a symmetric matrix, we can write $M = QSQ^T$, and because $\mathbf{x}^TM\mathbf{x} > 0\forall\mathbf{x} \neq \mathbf{0}$, we can can then conclude that S needs positive diagonal values).

Second, beacause A^TA is positive definite, $(A^TA)_{m\cdot m} = V\Lambda V^T$, where $\Lambda = \begin{pmatrix} \sigma_1^2 & ... & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & ... & \sigma_m^2 \end{pmatrix}$,

and $VV^T = V^TV = I$.

We can also write $A^TA\mathbf{v}_i = \sigma_i\mathbf{v}_i$, because $\mathbf{v}_i$ is eigen vector of A^TA.

Now, write $\mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i}$. By construction, $\mathbf{u}_i$ is a unit eigen vector of AA^T. Because:

- $AA^T\mathbf{u}_i = AA^T \cdot \frac{A\mathbf{v}_i}{\sigma_i} = A\sigma_i\mathbf{v_i} = \sigma_i^2\mathbf{u}_i$, second = is because $A^TA\mathbf{v}_i = \sigma_i\mathbf{v}_i$, third = is by definition of $\mathbf{u}_i$.
- $\mathbf{u}_i^T\mathbf{u}_i = \frac{\mathbf{v}_i^TA^TA\mathbf{v}_i}{\sigma_i^2} = 1$, because $\mathbf{v}_i$ is unit eigen vector of A^TA, with corresponding eigen value σ_i^2.

Since we defined $\mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i}$, with A_{n*m}, $(A^TA)_{m*m}$ and hence $\mathbf{v}_i$ a m-dim vector, $\mathbf{u}_i$ is n-dim.

Hence $U_{n*n} = A_{n*m}V_{m*m}\Sigma^{-1}_{m*n}$, where $\Sigma^{-1}_{m*n} = \begin{pmatrix} \sigma_1^{-1} & ... & 0 & ... & 0 \\ 0 & \sigma_2^{-1} & 0 & ... & 0 \\ 0 & ... & \sigma_m^{-1} & ... & 0 \end{pmatrix}$.

Hence, $A_{n*m} = U_{n*n}\Sigma_{n*m}V_{m*m}$. (**Note:** The dimension of Σ can be changed depending on the matrix multiplication it is involved in, simply by adding rows/columns of $\mathbf{0}$).

1.3 How is it applied in biological research

- **scRNA-seq data analysis**: A method can try to model the scRNA-seq input **count** data (cells*genes) in its raw form (*no normalization before visualization, clustering, differential analysis, trajectory analysis, etc.*). A general method will try to put assumptions on the count data distribution, and assumption on how the parameters for the distributions are constructed. For example, the mean gene expression parameter in each cell, for each gene, can be generated from generalized linear combinations of cells' variates (batch, etc.) and genes' variates (length, GC content). Then, in order to infer parameters and coefficients for the framework, SVD was applied. [**ZINB-WaVE**]. At

the same time, in general, SVD can be applied in getting lower-dim representation of cells, such as those used in the Seurat and scanpy pipeline. **How can we extend this?**: Develop methods that can apply this principles to scATAC-seq data ⇒ there can be drop out, but there can also *a lot more* spots of actual zero read count. Right now, ArchR tries to put reads into 500-bp bins, and then apply peak calling on top of that before doing downstream analysis ⇒ there must be a better way. Secondly, can we apply this type of model on differential gene expression analysis based on scRNA-seq data?

- **Population genetics**: Population Structure Analysis and Principal Component Analysis (PCA) ⟵ the famous John Novembre paper. Ancestry inference. Association Studies (SVD is used to control for population structure before association testing). Ask ChatGPT for some pointer papers.