Interpretable joint clustering of single-cell transcriptomes

Andreas Fønss Møller[1,2], Jesper Grud Skat Madsen[1,3,4,5]†

Affiliations:

[1]Institute of Biochemistry and Molecular Biology, University of Southern, Denmark

[2]Sino-Danish College (SDC), University of Chinese Academy of Sciences, China

[3]Institute of Mathematics and Computer Science, University of Southern Denmark

[4]Center for Functional Genomics and Tissue Plasticity (ATLAS), Odense M, 5230, Denmark

[5]The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

†Corresponding author. Email: jgsm@imada.sdu.dk

Supplementary Note 1: Consensus PCA

As a preprocessing step for the JOINTLY model, we developed a common PCA, that allows flexibility toward dataset-specific variation. For normalized and standardized gene expression values for highly variable genes, we calculate the variance-covariance matrix for each dataset:

$$V_d = \frac{\left({X_d}^T * X_d\right)}{n_d - 1}$$

For each dataset, d, calculate the variance-covariance (V_d) by matrix multiplication of the normalized and standardized gene expression values (X_d) and adjusting for the number cells (n_d) – 1. This matrix captures the variance of each feature along the diagonal and the covariance between each pair of features off the diagonal. Next, we calculate the within-group variance-covariance matrix (V) by summing the variance-covariance matrices for each dataset (V_d) multiplied by the number of cells in each dataset (n_d) and divided by the total number of cells across all datasets:

$$\mathrm{V} = \sum_{d=1}^{D} \frac{V_d * n_d}{\sum_{d=1}^{D} n_d}$$

The within-group variance-covariance matrix is the decomposed into a reduced dimensional space using randomized singular value decomposition[49]. By matrix multiplication between the left singular vectors (U) and the transposed normalized and standardized gene expression values for each dataset, we calculate a reduced dimensional space for cells.

For each dataset, we calculate the proportion of variance explained in the reduced dimensional space, and for datasets with less than 80% of variance explained, we calculate additional singular vectors. First, we calculate the residual matrix between the variance-covariance matrix (V_g) and the reduced dimensional space using the left singular vectors (U)

$$R_g = V_g - \mathrm{U} * \mathrm{U}^T * V_g$$

The residual matrices (R_g) are subject to randomized SVD as described above and dataset-specific left singular vectors are concatenated to the U matrix, such that U describes at least 80% variance in each dataset. Finally, the consensus PCA space is calculated by matrix multiplication between U and the normalized and standardized gene expression values.

Supplementary Note 2: The JOINTLY algorithm

Graph regularized kernel non-negative matrix factorization

We have implemented graph regularized kernel non-negative matrix factorization (NMF) from KOGNMF[10]. Briefly, let $X = (x_1, x_2, \dots x_n) \in \mathbb{R}^{mxn}$ as a data matrix of non-negative elements. Regular NMF factorizes X into two low-rank non-negative matrices:

$$X \approx VH$$

where H is the clustering matrix defined as $H^T = (h_1, h_2, \dots h_k) \in \mathbb{R}^{nxk}$, V is the basis matrix defined as $V = (v_1, v_2, \dots v_k) \in \mathbb{R}^{mxk}$, and k is the factorization rank, which is generally much smaller than the smallest dimension of X. This problem can be solved using gradient descent to minimize the reconstruction error under non-negative constraints[50].

To obtain a non-linear factorization, a non-linear transformation of X to a higher or infinite D dimensional space, such that $\Phi(X) = (\Phi(x_1), \Phi(x_2), \dots \Phi(x_n)) \in \mathbb{R}^{Dxn}$ is introduced. This can also factorize this non-linear space into two low-rank matrices:

$$\Phi(X) \approx WH$$

To solve this problem, the following loss function can be defined:

$$arg\min_{W,H} = \left|\left|\Phi(X) - WH\right|\right|$$

However, this problem cannot directly be minimized since $\Phi(X)$ potentially has infinite dimensions[10]. This issue can be circumvented by using the kernel trick by defining that W is a linear combination of the features in the non-linear transformed space $\Phi(X)$. The linear combinations are defined by a new matrix called F, such that:

$$W = \Phi(X)F$$

Substitution of this into the loss function obtains:

$$arg\min_{F,H} = \left|\left|\Phi(X) - \Phi(X)FH\right|\right|$$

Now, let $K \in \mathbb{R}^{nxn}$ be a kernel matrix such that $K = \Phi^T(X)\Phi(X)$. This can be exploited to derive multiplicative updating rules for updating H and F without factorizing $\Phi(X)$ with weights α and μ:

$$H \leftarrow H \odot \frac{\alpha F^T K + 2\mu H}{\alpha F^T KFH + 2\mu HH^T H}$$

$$F \leftarrow F \odot \frac{KH^T}{KFHH^T}$$

To add graph regularization, such that the factorization also reconstructs the geometric structure of the data in the non-linear feature space, the following loss function is introduced:

$$arg\min_{F,H} = \alpha\left|\left|\Phi(X) - \Phi(X)FH\right|\right| + \lambda \mathrm{Tr}(HLH^T)$$

where L is the Laplacian matrix defined as $L = D - A$ where D is the degree matrix and A is the adjacency matrix of the graph resulting in the following update rules:

$$H \leftarrow H \odot \frac{\alpha F^T K + 2\mu H + \lambda HA}{\alpha F^T KFH + 2\mu HH^T H + \lambda HD}$$

$$F \leftarrow F \odot \frac{KH^T}{KFHH^T}$$

Extending KOGNMF for joint clustering of single-cell RNA-sequence datasets

In the case of single-cell RNA-sequence datasets, this loss function optimizes two low-rank matrices per dataset; the clustering matrix, H (factor by cell) and the basis matrix in kernel space, F (cell by factor). The required inputs are kernel matrix, K (cell by cell), which we calculate based on an adaptive heat-based kernel in consensus PCA space (see Methods and Supplementary Note 1), as well as an adjacency and a degree matrix (cell by cell) both of which are calculated from a shared nearest neighbourhood graph built on consensus PCA.

The graph-regularized kernel NMF can factorize a single dataset considering non-linear similarities between data points. In the case of joint clustering of single-cell RNA-sequencing datasets, we have multiple datasets, and there are no guarantees that the same features would contribute to the same factors if the datasets were factorized independently. To solve this and learn a clustering matrix, H, where the factors are explained by similar factors across datasets, JOINTLY introduces a feature matrix $V_d = (v_1, v_2, \dots v_k) \in \mathbb{R}^{mxk}$ per dataset, d. This matrix is factorized with the clustering matrix, H, resembling regular NMF and reconstructing the original data matrix as $X_d \approx V_d H_d$. To use this for minimizing the difference between factors across datasets, JOINTLY introduce a loss based on the difference between the reconstructed gene expression space using the V matrix for the target dataset and all other datasets, minimizing the difference in reconstruction between V matrices:

$$\underset{F,H}{arg\min} = \alpha\left|\left|\Phi(X_d) - \Phi(X_d)F_d H_d\right|\right| + \lambda \mathrm{Tr}\left(H_d L_d H_d^T\right) + \beta \sum_{d=1}^{D} \sum_{j \neq d} \left|\left|V_j H_d - V_d H_d\right|\right|$$

For each dataset, this gives us the following update rules for H and F with weights α, μ, λ and β:

$$H_d \leftarrow H_d \odot \frac{\alpha F_d^T K_d + 2\mu H_d + \lambda H_d A_d + \sum_{j \neq d} \beta V_j^T X_d + \beta V_d^T V_d H_d}{\alpha F_d^T K_d F_d H_d + 2\mu H_d H_d^T H_d + \lambda H_d D_d + \sum_{j \neq d} \beta V_j^T V_j H_d + 2\beta V_d^T V_j H_d + \beta V_d^T X_d}$$

$$F_d \leftarrow F_d \odot \frac{K_d H_d^T}{K_d F_d H_d H_d^T}$$

After updating H and F for all datasets, we update V for each dataset using the new H and least squares setting negative values to 0:

$$V_d = \left(X_d X_d^T\right)^{-1} X_d^T H_d$$