Bi/BE/CS 183 2021-2022
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 3 Solution

Problem 1

Consider X as a $n \times m$ gene expression matrix, where n is the number of cells and m is the number of genes. Normally we treat the genes as features and cells as observations, and compute the *principal components* of the collection of n cell vectors (row vectors) by finding the eigenvectors of the covariance matrix of X^T (which is proportional to $X^T X$ assuming columns of X have zero mean). Suppose now you treat cells as features, so you want to find the principal components of the m gene vectors (column vectors) of X.

(a) (6 points) Give an expression for the covariance matrix Σ (up to a constant factor), where Σ_{ij} is the covariance between cell i and cell j. Note that X may not be mean-centered.

A: The covariance matrix of X is $\frac{1}{m-1}\bar{X}\bar{X}^T$ where the rows of X are subtracted by the row mean to obtain $\bar{X}$, i.e. $\bar{X}^T = (I_m - \frac{1}{m}J_m)X^T$, where I_m is the identity matrix and J_m is a matrix of all ones. Thus, $C = \frac{1}{m-1}X(I_m - \frac{1}{m}J_m)X^T$.

(b) (10 points) Outline how you would obtain the principal components of the m gene vectors of X

A: We can obtain the principal components by finding the eigenvectors of the covariance matrix of X which is $\bar{X}\bar{X}^T$ as in part (a). Specifically, first center X such that each row has a zero mean. Then, perform singular value decomposition of $\bar{X}^T$: $X^T = U\Sigma V^T$. The matrix V consists of the eigenvectors which diagonalize the covariance matrix C.

(c) (6 points) Provide an interpretation for the principal components you would obtain in this case, as well as the measured variance along them, please interpret in terms of cells and their gene expression levels.

A: The principal components (eigenvectors) of XX^T are directions of variance between the cells (based on their gene measurements, across all genes) and the measured (projected) variance along it denotes the variance of different observations (genes).

(d) (8 points) In this case, what is the maximum number of principal components that can have non-zero variance and why?

A: The maximum number of principal components is min(m-1,n) because that is the maximum number of independent components (rank) of the gene expression matrix. If $n > m-1$, by centering the matrix, we further reduce the rank by one, because row means of X are zero and it means the m columns of X are not independent. Thus, the rank is at most m-1. If $n <= m-1$, then the rank is n.

Problem 2

Construct 8 unique data points in $\mathbb{R}^2$, such that the eigenvectors of their covariance matrix have equal eigenvalues, provide their coordinates and explain why they have the desired property.

A: We can choose 8 points uniformly on the unit cycle, then the nonzero eigenvalues are equal due to symmetry. Note that in R^2, the angle between two eigenvectors is 90 degree. We can rotate the 8 data points by 90 degree and they still look the same. Therefore, the eigenvalues which can be interpreted as the projected variance onto the eigenvectors are the same.

Their coordinates are $(cos(\frac{2\pi}{8}k + 0.1\pi), sin(\frac{2\pi}{8}k + 0.1\pi))$, for k =1,...,8.

Problem 3

Finding principal components is equivalent to finding a matrix transformation that decorrelates a set of random variable. For example, given a vector $\mathbf{x} \in \mathbb{R}^m$ where the components are random variables representing the expression level of different genes, the corresponding $n \times m$ gene expression matrix can be viewed as n samples/observations of the random vector $\mathbf{x}$. When performing PCA on the gene expression matrix, we are trying in finding a projection matrix P such that $\mathbf{y} = P\mathbf{x}$ has uncorrelated components, i.e. the covariance matrix $\Sigma_{\mathbf{y}}$ is diagonal. In doing so, we are making the implicit assumption that all pairwise correlations $\rho(y_i, y_j)$ together captures most of the statistical dependencies in our measurements.

(a) (5 points) Explain why PCA may not remove all statistical dependencies, i.e. why components of $\mathbf{y}$ may still be statistically dependent.

A: Statistically independence means that the distribution of y_i is the same no matter what y_j is. $\mathbf{y}$ having uncorrelated components doesn't necessarily means that they are statistically dependent. For a counterexample, we can consider the variable X and Y in Homework 2 problem 3b. They are not independent but they have zero correlation.

(b) (5 points) Consider the more stringent form of removing redundancy which is statistical independence.

$$P(y_i, y_j) = P(y_i)P(y_j), \tag{1}$$

for all $i \neq j$, where $P(\cdot)$ denotes the probability density. The class of algorithm that attempts to satisfy this much more stringent constraint is known as Independent Component Analysis (ICA). Show that PCA actually accomplishes statistical independence (Equation 1) when $\mathbf{x}$ is (multivariate) Gaussian distributed (Hint: uncorrelated, jointly Gaussian random variables are independent).

A: Since $\mathbf{x}$ is (multivariate) Gaussian distributed and P denotes a linear projection, $\mathbf{y}$ is still Gaussian distributed. Note that $\mathbf{y}$ has uncorrelated components after projection. Therefore, as uncorrelated, jointly Gaussian random variables, components of $\mathbf{y}$ are independent, which proves the statement.

Problem 4

In this problem you will compare the results of PCA and SVD, common procedures for dimensionality reduction of a single-cell dataset. Using the eigenvectors (components) of these factorization procedures we will see how relevant "directions" in biological data can be extracted, such as components which distinguish the various cell types in the data.

The link to the Problem 4 notebook is here. Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.

The answer to the problem 4 is here