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

Problem Set 3

Submit your solutions as a single PDF file via Canvas by **10am Thursday January 27th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

Problem 1 (20 points)

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 of X (up to a constant factor). What do the diagonal and off-diagonal terms of this matrix represent in terms of cells and their gene expression levels?

(b) (6 points) Outline how you would obtain the principal components of the m gene vectors of X

(c) (4 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.

(d) (4 points) In this case, what is the maximum number of principal components that can have non-zero variance and why?

Problem 2 (30 points)

When performing PCA on a data matrix X using functions from scientific computation libraries, output from two different implementations of PCA may look quite different. We will explore some of the reasons with which this may happen.

(a) (4 points) Show that any scalar multiple of an eigenvector is still an eigenvector, so that any principal component is unique only up to a linear transformation.

(b) (4 points) Suppose we ask principal components to be normalized to have unit l_2 norm, does this ensure uniqueness?

(c) i. (6 points) Suppose the matrix X has a set of eigenvectors $\{\mathbf{v}_i\}_{i=1}^n$ with identical eigenvalues λ, show that any linear combination of this set of eigenvectors is also an eigenvector.

ii. (10 points) 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.

iii. (6 points) Suppose we scale all principal components of a data matrix so their last coordinate is 1, furthermore we know that by definition principal components must be mutually orthogonal, do these conditions altogether ensure their uniqueness? Explain your answer.

Problem 3 (10 points)

Finding principle 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.

(b) (5 points) Consider the most rigorous 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 rigorous 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).

Problem 4 (40 points)

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.