1. Applying PCA on a dataset that has 8 points of a circle centered around the origin would result in a non-unique PCA. 8 such points for a circle of radius 25 are:

(25, 0), (17.6776695, 17.6776695), (~~1.5308085e~~ 0, 25), (-17.6776695, 17.6776695), (-25, 0), (-17.6776695, -17.6776695)

(0, -25), (17.6776695, -17.6776695)

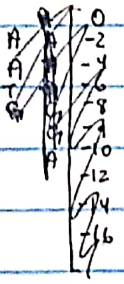
All these points are the same distance from each other on the circle, so "fitting" a line anywhere through this dataset such that the variance of the projected points is maximized would have a symmetrical solution via a diffrent line in this dataset.

2. GATTACA and AATGGACA

Essentially, I create a score matrix with padding on the sides for the gap scores. It looks like this initially:

$$\begin{array}{c|cccccccc} & & G & A & T & T & A & C & A \\ \hline & 0 & -2 & -4 & -6 & -8 & -10 & -12 & -14 \\ A & -2 & 0 & \cdots & & & & & \\ A & -4 & 0 & 0 & \ddots & & & & \\ T & -6 & \vdots & & & \ddots & & & \\ G & -8 & \vdots & & & & & & \\ G & -10 & \vdots & & & & & & \\ A & -12 & \vdots & & & & & & \\ C & -14 & & & & & & & \\ A & -16 & & & & & & & \end{array}$$

For the next step, the mismatch is a score of -1, which is lower than the other scores (-2), so that cell is filled with -1. In this fashion, the matrix is then populated with scores. The maximal score ends up at the bottom right cell of the score matrix:



$$\begin{array}{c|cccccccc} & & G & A & T & T & A & C & A \\ \hline & 0 & -2 & -4 & -6 & -8 & -10 & -12 & -14 \\ A & -2 & -1 & -1 & -3 & -5 & -7 & -9 & -11 \\ A & -4 & -3 & 0 & -2 & -4 & -4 & -6 & -8 \\ T & -6 & -5 & -2 & 1 & -1 & -3 & -5 & -7 \\ G & -8 & -5 & -4 & -1 & 0 & -2 & -4 & -6 \\ G & -10 & -7 & -6 & -3 & -2 & -1 & -3 & -5 \\ A & -12 & -9 & -6 & -5 & -4 & -1 & -2 & -2 \\ C & -14 & -11 & -8 & -7 & -6 & -3 & 0 & -2 \\ A & -16 & -13 & -10 & -9 & -8 & -5 & -2 & 1 \end{array}$$

The maximal score is 1. We can find the alignment by backtracking. Since A matches with A, we performed 0 + 1 to reach the bottom right. C also matches C, so we performed -1 + 1 = 0. A matches A, so we performed -2 + 1 = -1. At -2, we see that we performed -1 + -1 for a mismatch. ~~0+(-2) for a gap score, so there is a gap there where we advance one base for AATGGACA but not GATTACA. Thus, our backtrack looks like: ACA~~ Here, after advancing diagonally again, we see that we performed 1 + (-2) = -1 for a gap, so there is a gap where we advance one base for AATGGACA but not GATTACA. Thus, our backtrack looks like ACAT– matching with ACAGG. For the rest, we have two matches and one mismatch.

At the end, we receive this optimal alignment:

GAT-TACA

AAT GGACA . Score = 1

4. Expected number of cells uniquely barcoded in a droplet single-cell RNA seq experiment.

Probability that a barcode ends up in k different cells:

$$\mathbb{P}(N=k) = \binom{N}{k} p^{k} (1-p)^{N-k}$$

Probability that a cell is assigned any specific barcode:

$$p = \frac{1}{M}$$

Probability that a barcode is associated to <u>exactly</u> 1 cell (unique barcode)?

$k=1$

$$\binom{N}{1}\left(\frac{1}{M}\right)^{1}\left(1-\frac{1}{M}\right)^{N-1} = \mathbb{P}(N=1)$$

$$\mathbb{P}(N=1) = N\left(\frac{1}{M}\right)\left(1-\frac{1}{M}\right)^{N-1}$$

$$= \left(\frac{N}{M}\right)\left(1-\frac{1}{M}\right)^{N-1}$$

Expectation: multiply by M:

$$M\cdot\left(\frac{N}{M}\right)\left(1-\frac{1}{M}\right)^{N-1} = \boxed{N\left(1-\frac{1}{M}\right)^{N-1}}$$