Student number:	2467273
Course title:	COMPSCI4073 Data Fundamentals (H)
Questions answered:	1 and 2

1.

a)

i) Raul suggests using direct estimation. Pro: extremely efficient and fast, power iteration can be done quickly. Con: the eigenvalue might not be the best estimate of the size (it might not work for this project).

Hugh suggests using Maximum Likelihood Estimation. Pro: this model would definitely work to find some estimate (even if it is in a local minimum). Con: requires an optimisation process implementation with an objective function, parameters, maybe constraints, etc.

Clara suggests using Bayesian (probabilistic) estimation. Pro: a very easily understandable approach for humans. Con: sometimes can be very hard to compute because representation of these beliefs is difficult.

ii) Model A is most appropriate because it represents “points” as an observed variable (in a box), “size” as an unknown variable (in a circle), on which “points” depends and which depends on both “epsilon” and “spread” (both latent variables in circles)

b)

i) When comparing each sample type *j*, variance should be measure of the other indices in that size category (*i* and *k*), first, by calculating the mean of those indices and, second, by taking the sum of the squared differences of each element from the mean of that size category:

$$\sigma^2 = \frac{1}{N-1}$$

$$\sum_{i=0}^{N-1} (x_i - \mu_i)^2$$

ii) A problem is that the approach does not show any kind of variance/standard deviation for each size type. A much better visualisation approach would be to use box plots, which would be able to demonstrate both the mean/median count of the size type as a line near the middle of the box and the “diversity”/deviation with interquartile ranges as the edges of the boxes, extrema as whiskers and also outlying data points outside of the whiskers.

c)

i) Because these two frames could be very precise, division of them could result in an inexact computation that would round off some digits incorrectly. Secondly, if the system uses unsigned floating point representation, division could result in an underflow where the result might be smaller than the smallest representable number in that system.

ii) (100, 35, 2, 60)

iii)

```
def expected_locations(guess_array, img):
    last_frame = guess_array[:, :, :, -1]
    indices = np.arange(0, 35)
    log_liks = lik_cell(last_frame[
    return np.stack(last_frame[:, 0], log_liks)
expected = log_likes *
```

2.

a)

i) This is a faceted plot since there are multiple coords (of the same scale) next to each other, not overlaid.

ii) Linear transforms could be A, C, D and E since they are all just rotations and/or scaling of the original dataset. No information is lost.

iii) This would be transform D, which is very singular and unstable because of a large condition number.

iv) All singular values being close to one would imply the matrix is well-conditioned (with a small condition number), which would be transforms A, C and E.

v) A determinant close to zero would be for singular matrices, like transform D. A determinant of zero would mean that the matric cannot be inverted.

b)

i) Transform removal would be achievable by inverting the transform matrix.

ii) Using the SVD function, matrix A is split into three matrices:

$$A = U\Sigma V^T.$$

- a) The removal could then be calculated as V @ \Sigma^{-1} @ U^T, which can then be applied to the transform to acquire the original input.
- b) \Sigma contains in its diagonal the singular values of A, the biggest and smallest values of which can be used to calculate the condition number. If this number is small, the transform is stable, but if the number is large, it is numerically unstable.
- c) If U and V are identity matrices, then there will be no rotation since they are responsible for rotation. Likewise, if \Sigma is an identity matrix, then there will be no scaling.

c)

i)

```
def distances(pts, observed):

        return np.linalg.norm(observed – pts, ord=1) ** 2
```

ii) Problems and solutions:

- Title is extremely vague. Should be “Distance between calibration points and observed points by microphone”
- There is no label on x axis. Should be “distance”
- There are no units on the labels. Should be added on axis labels.

- There is no legend distinguishing between microphones. Should be added in the right corner.
- There are no point geoms to signify observations. Should be added in every corner of each line (12 for each)
- Wrong order of x/y, axes should be swapped because focal level is the independent variable and distance the dependent.

iii) If the focal level can be changed/split into more than just 12 divisions (i.e., there are focal levels in between those 12), the current line plot makes most sense because focal level values can be changed to between the 12 measured. A step/staircase plot would assume the points jump from one observation to the next and that there are no possible different values in between, which would be incorrect because there are unmeasured values between them, which could be measured and made more precise.