Ex 2.1

a) A linear dataset can be simplified with a simple linear equation where it has a polynomial degree of 1. This means two variables can be linearly seperated with a line unlike non-linear data which has many dimensions and is difficult to visualize.

PCA can easily comprehend this linear relationship in the data however if the data has nonlinear pattern, PCA might reverse the analysis and point to the exact opposite direction. This can also result to information loss along with biased output. Therefore, it is important to make sure the data is correct.

b) PCA essentially maximizes the variance to identify which components contribute how much in our dataset. If we do not normalize and data points are very far away from each other it is likely that PCA will take the largest variance as the most important component and provide a biased output. For example, in this case we might not able to see the contri-

- bution of other [illegible]
Only one feature may get highlighted. This is why normalizing the data is important where all features are taken into account.

c) i) PCA doesnot take account of label information. Therefore, always reducing dimensions may lead to losing important features. For example, when distinguishing between cars and buses, taking how many miles they have gone factor won't work even though it may have large variance. Height might be a good factor to consider. If the dataset is arranged in a way that PCA is biased towards wheels discarding height as a factor would be wrong.

ii) When the data is in different scales or notations, PCA should not be applied. Such as if we have categorical data or numerical information that are in different scales (i.e currency), PCA would result in wrong components and [illegible] features

we would lose important features by reducing dimensions.

iii) Another case would be health related data where data might not fluctuate as much thus resulting in low variance but could be an important factor. Blindly reducing dimension here would cause information loss and ineffective classification.

Ex. 2.4

a) Given,

$$y = X^T . w + \epsilon$$

Assuming ϵ as normally distributed,

$$f(y) = X^T . w$$

conditional mean of y given X,

$$P(y|x) = X^T . w$$

and the conditional variance of y given X is

$$Var(y) = \sigma^2$$

probability function of Y given x,

Conditional probability

$$P(y|x) = \frac{P(\{x^T.w\} \cap y)}{P(x^T.w)}$$

b) Since $\{x_n, y_n\}_{n=1}^{N}$; $x_n \in R^D$

hence the mode becomes:

$$y_i = X_i^T.w + \epsilon \quad ; \text{where } i = 1, 2 \ldots n$$

So the likelihood function is:

$$L(x_i, y_i; w, \sigma^2, \epsilon) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[-\frac{1}{2\sigma^2} \left(y_i - X_i^T w - \epsilon \right)^2 \right]$$

d) To calculate MLE for w:

Removing the constants and taking Log of the function -

$$L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - X_i^T w - \epsilon \right)^2$$

To find the MLE's for w, we have to find LSE of w.

$$\frac{\partial L}{\partial \epsilon} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(y_i - X_i^T w - \epsilon\right)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\epsilon - w\sum_{i=1}^{n} X_i^T = 0$$

$$\Rightarrow \epsilon = \bar{y} - w\bar{X}^T$$

$$\frac{\partial L}{\partial w} = \frac{1}{-2\sigma^2}\sum_{i=1}^{n} 2\left(y_i - X_i^T w - \epsilon\right)(-X_i)$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^T - \epsilon\sum_{i=1}^{n} x_i^T - w\sum_{i=1}^{n} x_i^T 2 = \vec{0}^T$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^T - \left(\bar{y} - w\bar{x}^T\right)\sum_{i=1}^{n} x_i^T - w\sum_{i=1}^{n} x_i^T 2 = 0$$

$$\Rightarrow w = \frac{\sum_{i=1}^{n}\left(x_i^T - \overline{x^T}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i^T - \overline{X^T}\right)^2}$$

This is the maximum likelihood estimator for w.