1a) $D_{KL}(q(z|x) \| p(z))$

given dimension of z is 1 & $p(z) \sim N(0,1)$

$$\therefore p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

$$q(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

$$D_{KL}(q(z|x) \| p(z)) = -\int q(z|x) \log\left(\frac{p(z)}{q(z|x)}\right) dz$$

$$= -\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(z-\mu)^2}{2\sigma^2}\right) \log\left(\frac{\sigma \exp(-z^2/2)}{\exp\left(\frac{-(z-\mu)^2}{2\sigma^2}\right)}\right)$$

$$= -E\left[\log\sigma + \frac{(z-\mu)^2}{2\sigma^2} - \frac{z^2}{2}\right] \quad E \to \text{entropy}$$

$$= -\left[\log\sigma + \frac{1}{2\sigma^2} E((z-\mu)^2) - E\left(\frac{z^2}{2}\right)\right]$$

$$= -\left[\log\sigma + \frac{1}{2\sigma^2}\sigma^2 - \frac{\sigma^2+\mu^2}{2}\right]$$

$$= \log\frac{1}{\sigma} + \frac{\sigma^2}{2} + \frac{\mu^2}{2} - \frac{1}{2}$$

1b) If α is too large then the reconstruction loss term will not be given enough weightage and so the network will focus mainly on keeping the $q(z|x)$ as close to $p(z)$ as possible.

This will result in often incorrect reconstructions as z will look very similar for all x inputs. Hence the decoder will not be able to correctly generate the reconstructions

1c) 1. PCA without complex kernels is a linear operation while VAEs are not. Hence VAEs can compute more complex data

2. PCA features are linearly uncorrelated by design while VAE features may not be. This is because VAE features are tuned for accurate reconstructions only.

2a) False

We need a well trained discriminator to force the generator to work harder. Hence we typically train the discriminator more to force the generator to work harder. If we train generator to saturation for a poor discriminator then the generator will have no incentive to improve once it becomes better than the discriminator

b) Value of $D(G(z))$ will be 0

This is because the generator output will be drastically different from the ground truth making it easy for the generator to distinguish

c) We will use non-saturating loss

This is because early in training the non saturating loss will have much larger values as $D(g(z))$ is closer to 0 and hence will provide stronger gradients for faster training. (Goodfellow et. al. 2014)

d) True.

when $D(G(z))$ is close to 1 then the discriminator thinks that the input comes from real data. Hence this means that for a well trained discriminator the generator outputs mimic the real data very well. Moreover the expectations of the lossfunction form a convex set and that implies that the generator output approaches real data output. as proved in the original GAN paper by Goodfellow et. al. 2014