Summary and discussion of: Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse

ZHUANG, Zeyan

1 Summary

1.1 Background

Variational autoencoders(VAEs) is a powerful method to model the high dimensional data by low dimensional latent vatiables. Suppose we have a data set $\{x_1, x_2, ..., x_N\}$, VAEs assume that each data point x_i is generated by a latent variable z_i [2]:

$$\begin{aligned} z_i &\sim p(\cdot) \\ x_i|z_i &\sim p_\theta(x_i|z_i) \end{aligned} \tag{1}$$

An important goal of VAE is to maximise the marginal likelihood $p_\theta(x) = p(z)p_\theta(x|z)$. once optimal θ is found, We can approximate the exact posterior $p_\theta(z|x)$ to encode data. However, $p_\theta(x)$ is often not intractable for complex models. These are done by the variational approximation. If we rewrite the log-likelihood:

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}\big[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\big] + D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{2}$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big]}_{\text{(A)}} \underbrace{-D_{KL}(q_\phi(z|x)||p(z))}_{\text{(B)}} + D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{3}$$

The (A) and (B) of the right hand side(RHS) of (8) represent evidence lower bound(ELBO), which respectively represent the reconstruction error and the KL divergence between the encoder and the prior $p(z)$.$q_\phi(z|x)$ is variational approximation of exact posterior $p_\theta(z|x)$. ELBO becomes a proxy objective.

However VAEs always suffer from posterior collapse(PC):the variational posterior distribution of a latent variable is equal to its prior:

$$q_\phi(z|x) \approx p(z) \tag{4}$$

That is, the variatonal postrior distribution is the same regardless of the data points, and the encoder cannot provide meaningful representations. Meanwhile PC makes it impossible for the decoder to make use of the information content of all of the latent dimensions.

Many studies have given different explanations for this phenomenon, for example:

1. Early stopping of training causes the variational distribution $q_\phi(z|x)$ to fail to match the exact posterior distribution $q_\theta(z|x)$.

2. When the model is flexible, the local optimum of ELBO includes the PC points(KL divergence becomes 0).

In this Pater, the authors campared Linear VAE and pPCA and discusses the relations between observation noise, local maxima, and PC. Then explain that PC is not exclusively due to ELBO. ELBO will not introduce any additional spurious local maxima in Linear VAE but the likelihood $p_\theta(x)$ and the model itself(observation noise) cause local optimums.

1.2 Theoretical results

1.2.1 Probabilistic PCA

pPCA model assumes that the data $x \in \mathbb{R}^d$ is generated linearly from the latent space $z \in \mathbb{R}^q$:

$$x = Wz + \mu + \epsilon$$

Stand Gaussian prior is used for z,

$$P(z) = \mathcal{N}(0, I) \tag{5}$$

$$P(x|z) = \mathcal{N}(Wz + \mu, \sigma^2 I) \tag{6}$$

For fixed σ^2, the stationary points of MLE is [1]:

$$W = U_q(\Lambda_q - \sigma^2 I)^{1/2} R \tag{7}$$

Where $U_q \in \mathbb{R}^{d\times q}$ and its columns $\mathbf{col_j} U_q$ are eigenvectors of sample covariance matrix S. $R \in \mathbb{R}^{q\times q}$ is a arbitrary orthogonal matrix. $\Lambda_q \in \mathbb{R}^{q\times q}$ is diagonal.There are two cases:

$$[\Lambda_q]_{jj} = \begin{cases} \lambda_j(S), & or \\ \sigma^2, \end{cases} \tag{8}$$

when $[\Lambda_q]_{jj} = \lambda_j(S)$, it requires that $(\lambda_j(S), \mathbf{col_j} U_q)$ are eigenvalue-vector pairs of S and $\lambda_j(S) > \sigma^2$. If $[\Lambda_q]_{jj} = 0$, $\mathbf{col_j} U_q$ is a arbitrary vector.

Assume the eigenvalues are in decreasing order, $\lambda_i(S) > \lambda_{i+1}(S)$. By further analysis reveals that the MLE of W, σ are:

$$\sigma^2_{MLE} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j(S), \tag{9}$$

$$W_{MLE} = U_q^*(\Lambda_q^* - \sigma^2_{MLE} I)^{1/2} R. \tag{10}$$

that there is no 0 in diag of Λ_q^*. i.e. no columns of U_q is 0.

1.3 Linear VAE

Linear VAE is a special case of VAEs as we discussed before. The model is defined as:

$$\begin{aligned} P(x|z) &= \mathcal{N}(Wz + \mu, \sigma^2 I) \\ p(z|x) &= \mathcal{N}(V(x - \mu), D) \end{aligned} \tag{11}$$

Where D is diagnoal convariance matrix.The author says it is a significant restriction.

The author proved that the global maximum of Linear VAE matchs with pPCA. i.e.

$$\max \mathbf{ELBO} = \log p(x|\sigma^2_{MLE}, W_{MLE}) \tag{12}$$

At the same time There are some properties at the optimal of ELBO. The authoer shows that the optimal W of Linear VAE must orthogonal in column space. i.e. $(W^*_{VAE})^T W^*_{VAE}$ is a diagonal matrix. In these cases the variation distribution recover the exact postrior. In the equivalence part of pPCA is that $R = I$. A natural corollary is that at the global optimum of ELBO, Linear VAE identify the principal components of S.

Athother important thing is that, Can VAE reach the global optimal to recover the pPCA? The futher results showed that ELBO will not introduce extra local maxima. i.e. any additional stationary points of ELBO must be saddle points.

Combining the two points above, An important conclusion is that the PC will occur in the stationary solutions of $\log p_\theta(x)$ not in local maxima of ELBO.

Then, we can turn back to PC situation. Under what circumstances does PC occur in Linear model? An intuitive view is that the model is define as $x = Wz + \mu + \epsilon$. If the ground truth is that one colunm of W is equal to zero,

$$\mathbf{col_i} W = \mathbf{0} \tag{13}$$

then z_i and x are independent, so $p(z_i|x) = p(z_i)$. Then the PC happens in i^{th} latent dimension. It is same in the learning process, if $\mathbf{col_i} W_{VAE} = 0$, $D_{KL}(p(z_i|x)||p(z_i)) = 0$. According to pPCA, different level of σ^2 have different stationary points and the larger σ^2 there are more 0 columns in W, so learning σ^2 is important(Unlike the general case, setting $\sigma^2 = 1$ fixed).

1.4 Non-linear case

For the nonlinear case, the authors apply the conclusions of Linear VAE to Deep Gaussian VAE and analyze it by experiments. The objective ELBO of Deep Gaussian VAE:

$$L = -D_{KL}(q_\phi(z|x)||p(z)) - 0.5\frac{1}{\sigma^2}\mathbb{E}_{q_\phi(z|x)}[||D_\theta(z) - x||^2] - 0.5\log(2\pi\sigma^2) \tag{14}$$

In this case, σ^2 plays the similar role in Linear VAE.

2 Result and Discussion

The experimental setup is the same as in the original paper. The Dataset contains 1000 randomly chosen MNIST images. The code are availiable at github code.The reference of code is explained in readme.

2.1 Match of pPCA and Linear VAE

This part mainly verifies the first lemma proposed in the paper(12).In this experiment, we need to observe that for different latent space dimensions, whether pPCA and inear VAE can match each other.

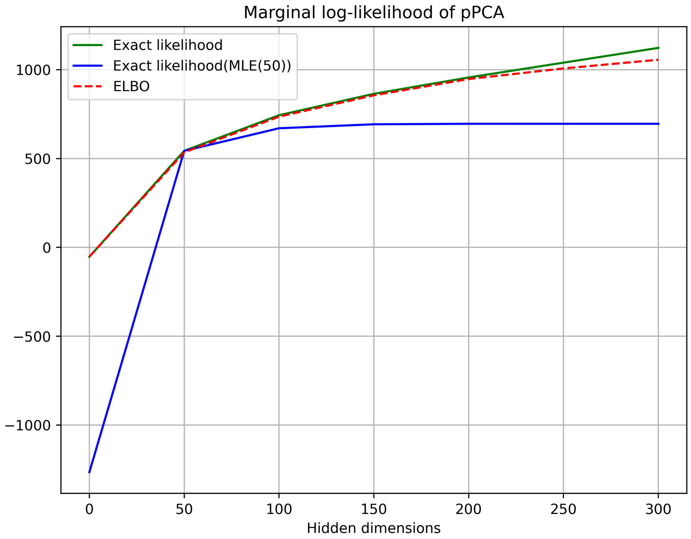
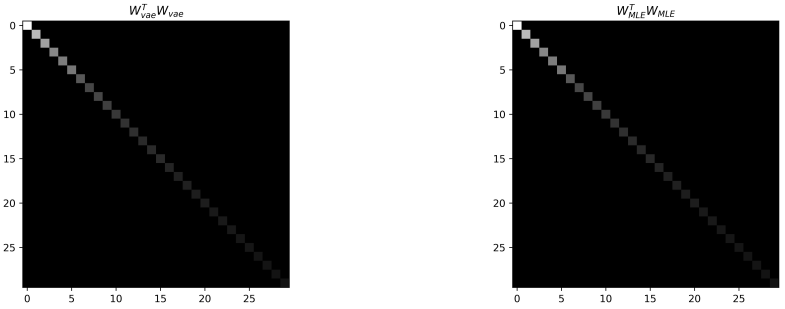


Figure 1: Marginal log-likelihood of pPCA

We first compare the green and red lines. The red line indicates the final ELBO of Linear VAE and the green line denote the $\log p(x|MLE)$ of pPCA. Among the different range of latent space dimension, they are very close to each other. This result match with the original paper and (12).

the blue line denote that we use fixed $\hat{\sigma}^2 = \frac{1}{784-50}\sum_{j=50}^{784}\lambda_j(S)$ to calculate the optimal W_{MLE}. Except for 50, the final log-likelihood function is lower than global optimum with $\hat{\sigma}^2 = \sigma^2_{MLE}$. This result is also expected and conformed to theory. In this condition, PC will happen.



(a) Gram VAE

(b) Gram MLE

Figure 2: Gram matrix of W of Linear VAE and pPCA

We can see that the gram matrices of the two W matrices are very similar.

2.2 Effect of stochastic ELBO

Since the linear model allows the analytical form for the reconstruction error:

$$\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] = f(W, D, V, \mu, \sigma^2) \tag{15}$$

but in the real case, the model is more complex and can only be estimated using Monte Carlo estimation and calculated with the reparameterization trick. This may produces large variance and even leads to non-convergence of the results.

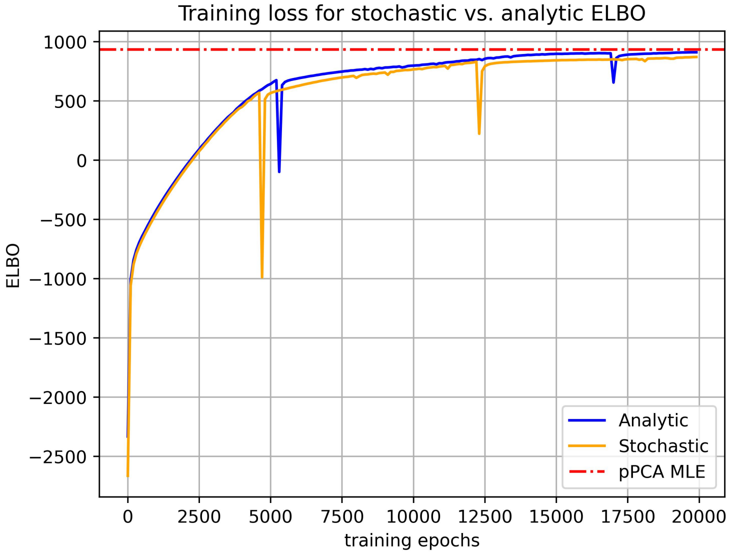


Figure 3: Stochastic vs analytic ELBO training

I compared the effect of analytic solution and random sampling. The experimental results indicate that the two methods are almost identical and the convergence rate is almost the same witch is contrast to the paper.I think there are some reasons: True distribution of data $p(x)$: If the authors have added some noise or used a different method of sampling MNIST when doing the experiment, this leads directly to the estimation of the log-likelihood $\log p_\theta(x)$. It is possible that the variance of the original data is small, so the noise in the gradient is small enough. Different calculation frameworks (Pytorch,TensorFlow) may also introduce biase.

But the results of the experiment are meaningful: Random sampling to estimate the reconstruction error is efficient.

2.3 Effect of encoders

Assume that there is no changes in the assumptions of the generative model(or marginal likelihood):

$$p_\theta(x) = \int p_\theta(x,z)dz \tag{16}$$

$$= \int p(z)p_\theta(x|z)dz \tag{17}$$

$$= \int p(z)\mathcal{N}(Wz+\mu,\sigma^2 I)dz \tag{18}$$

But we can use a stronger model to get the variational distribution $q_\phi(z|x)$. Theoretical results show that the optimal ELBO meet with the pPCA again (12)(I think here because the marginal likelihood don't change) However, the stationary points are more complex. So the author used different encoders.

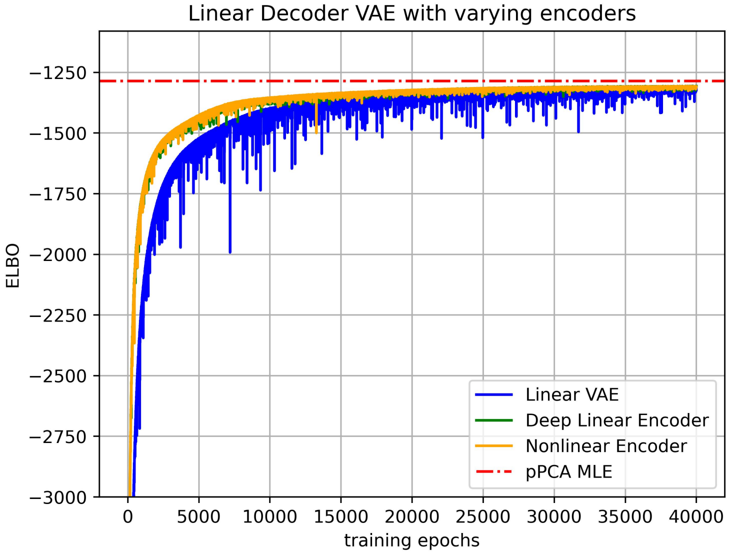


Figure 4: VAEs with linear decoders and varying encoders

In my setting, there are three kinds of decoder: Linear, nonLinear and Deep Linear. Linear and non-Linear are same with the paper. In Deep Linear case, I use one linear layers to produce variational mean and variance. (the different between Linear and Deep Linear is that in Linear the variance D changes to $D(x)$).Also apply augmentation to the MNIST.

The experimental results are consistent with the previous discussion that ELBO matchs with the pPCA log-likelihood (12). However results are a little different from the original paper: In the original paper, the linear result was the best but my results showed that the nonlinear is better. This is in line with my expectations, nonlinear can be more efficient approximation.

2.4 Impact of σ^2 on PC in nonlinear case

The author extend the result to the non-linear case to some extent. The authors define a measurement for a PC for the i^{th} latent dimension:

$$\mathbb{P}[D_{KL}(q(z_i|x)||p(z_i)) < \epsilon] \leq 1 - \delta \tag{19}$$

if we turn back to (7), larger observation noise will introduce more zero columns in W. The author first fix σ^2 and train the VAE. Secondly, initialize the σ^2 in decoder and then train it.The results of my implementation are shown below:

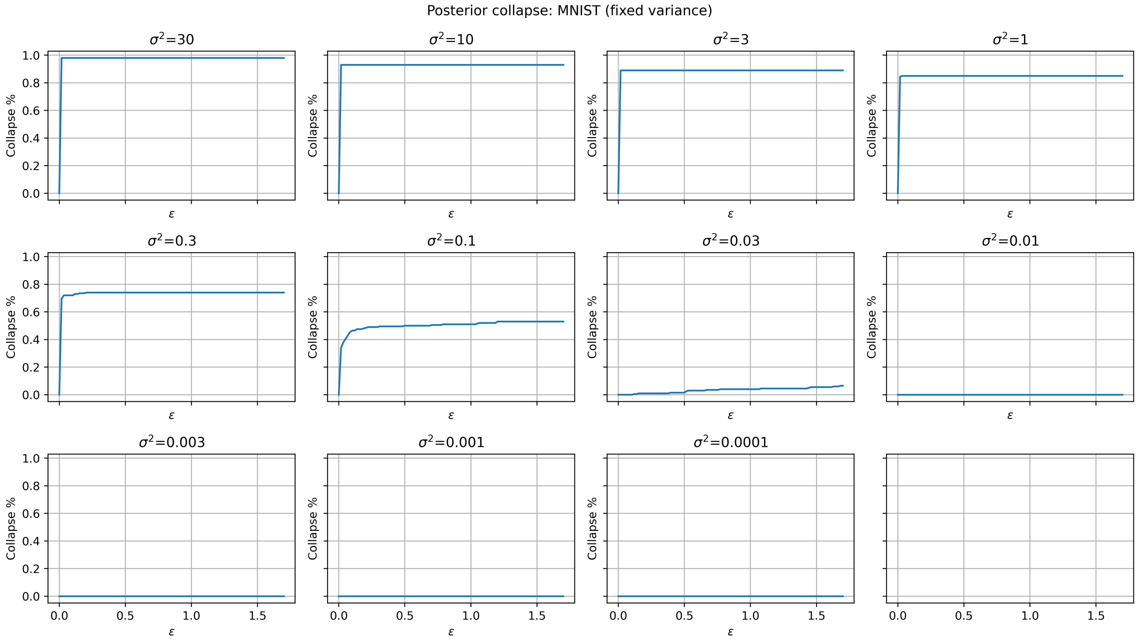


Figure 5: PC percentage of latent dims with fixed σ^2

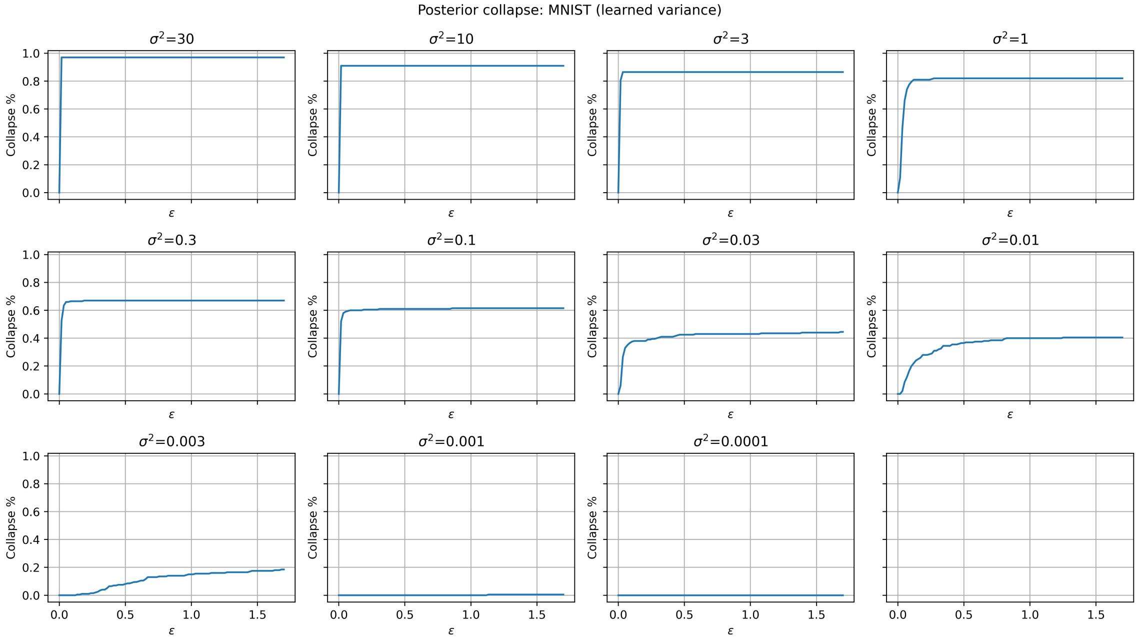


Figure 6: PC percentage of latent dims with trained σ^2

The experimental results show that the smaller the σ^2(Here the real observation noise may be relatively small, The previous result indicates that if the small σ^2 that don't match the σ_{MLE} will cause PC) smaller the PC percentage . This result matchs with the Linear VAE. For learnable parameters, the initialized value still has an effect on the PC percentage . I think this is due to the the loss surface of the deep model is very complex. For different initial value of σ^2, there exists local optimal of the loss surface near the σ^2 so these sub-optimal points will be reached and cause PC.

3 Conclusion

By comparing Linear VAE and pPCA the authors proposed a new understanding of the PC problem. Different from the classical point of view, the marginal likelihood will introduce PC, i.e. the exact posterior is collapsed. Then the approximate posterior will never reach the uncollapsed distribution. To some extend, the theoretical results can be generalized to the non-linear case. In this paper, the PC problem is found in high observation noise case. I think it may be that for different generative models(non-Gaussian), or even different priors, PC can be found in different levels. It also can be analyzed in the same framework.

References

[1] Tipping, Michael E., and Christopher M. Bishop. "Probabilistic principal component analysis." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61.3 (1999): 611-622.

[2] Wang, Yixin, David Blei, and John P. Cunningham. ”Posterior collapse and latent variable non-identifiability.” Advances in Neural Information Processing Systems 34 (2021): 5443-5455.