Don't blame the ELBO! A Linear VAE Perspective on Posterior Collapse

James Lucas, George Tucker, Roger Grosse, Mohammad Norouzi

NeurIPS: 2019

Presented by: Achint Kumar

Duke University

June 27, 2023

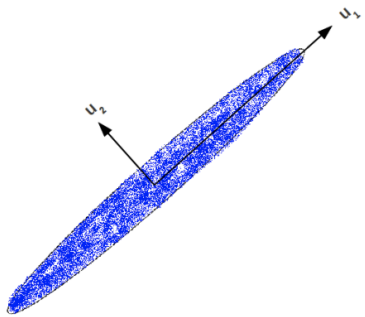
Principal Component Analysis

Given n-dimensional data, reduce it to k dimensions along directions of maximum variance. Steps:

1. Calculate covariance matrix Σ
2. Calculate eigenvalues, vectors of covariance matrix. The equation is given by,

$$\Sigma U = \Lambda U$$

Here, U_k correspond to the first k principal components of data with the corresponding eigenvalues, $\lambda_1, \ldots, \lambda_k$



Problems with PCA

1. Cannot handle missing data
2. Computing eigen-decomposition of covariance matrix is computationally expensive
3. Inferring optimal number of dimension is not possible.

Probabilistic PCA (pPCA)

In pPCA, data x is assumed to be generated by an affine transformation of a low dimensional latent vector, z. Mathematically we have

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

We assume latent vector follows standard normal distribution, $N(0, \mathbb{I})$ (same as VAEs) and noise ϵ obeys isotropic normal distribution, $\mathcal{N}(0, \sigma^2\mathbb{I})$. So, our starting point is:

	Mean, μ	Covariance, Σ
Prior, p(z)	0	$\mathbb{I}$
Likelihood, $p(x\|z)$	$Wz + \mu$	$\sigma^2\mathbb{I}$

From this we can calculate,

	Mean, μ	Covariance, Σ
Marginal likelihood, p(x)	μ	C
Posterior, $p(z\|x)$	$M^{-1}W^T(x - \mu)$	$\sigma^2 M^{-1}$

where $M = W^T W + \sigma^2\mathbb{I}$ and $C = WW^T + \sigma^2\mathbb{I}$.
To derive the above expressions use,

$$p(x) = \int p(x|z)p(z)dz$$

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)}.$$

Notice, everything obeys normal distribution. Next, we ask what is the optimal (MLE) values of W, μ and σ.

Maximum Likelihood Estimate of Parameters: μ, σ^2, W

The marginal likelihood is given by

$$p(x) = \prod_{i=1}^{N} \frac{1}{(\sqrt{(2\pi)^n |C|}} e^{-\frac{(x_i-\mu)^2}{2C}}$$

To find maximum likelihood estimate of parameters, μ, σ^2, W we optimize $\log p(x)$ as follows:

$$\frac{\partial \log p(x)}{\partial \mu} = 0 \Rightarrow \mu_{MLE} = \frac{1}{N} \sum_{i}^{N} x_i = \langle x \rangle$$

This makes intuitive sense in that the latent vector is being translated to the mean value of data, x.

$$\frac{\partial \log p(x)}{\partial \sigma^2} = 0 \Rightarrow \sigma^2_{MLE} = \frac{1}{n-k} \sum_{j=k+1}^{n} \lambda_j$$

This makes intuitive sense in that σ^2_{MLE} represents average variance lost in projection. The least noise variance possible is the variance of the last PC λ_n.

$$\frac{\partial \log p(x)}{\partial W} = 0 \Rightarrow W_{MLE} = U_k(\Lambda_k - \sigma^2_{MLE}\mathbb{I})^{1/2} R$$

U_k matrix contains the first k PC of data. This equation is saying, the optimal weights are the principal components scaled by standard deviation in that direction beyond the intrinsic noise. R is arbitrary rotation matrix. Let's think some more...

Thinking about W_{MLE}

We have,

$$W_{MLE} = U_k(\Lambda_k - \sigma^2_{MLE}\mathbb{I})^{1/2}R$$

Notice the following properties:

1. If $R = \mathbb{I}$, then $M = W^T_{MLE}W_{MLE} + \sigma^2_{MLE}\mathbb{I} = \Lambda_k$ (useful later)
2. If $\lambda_j = \sigma^2_{MLE}$ then from j^{th} to k dimension, W_{MLE} corresponding to those directions are zero. This is called posterior collapse. Corresponding z's don't contribute to the generative process.

Next we find the stability of W_{MLE} by figuring out whether it is a minima, maxima or saddle point.

Stability of W_{MLE}

Replace

$$W_{MLE} = U_k(\Lambda_k - \sigma^2_{MLE}\mathbb{I})^{1/2}$$

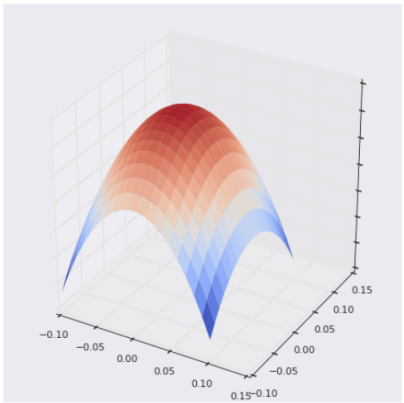
by

$$W = U_k(K_k - \sigma^2\mathbb{I})^{1/2}$$

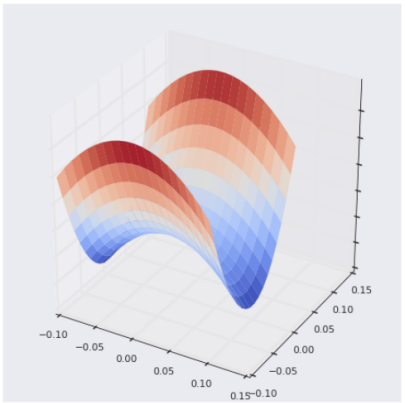
Here, K_k matrix consists of singular values of W. We will now vary σ^2 and test the stability of the 5^{th} and 7^{th} principal components. We find,

1. If $\sigma^2 = \lambda_4$ (large), then both PC directions are unstable (figure a)
2. If $\sigma^2 = \lambda_6$ (intermediate), then one PC direction(5^{th}) is stable, other one(7^{th}) is unstable (figure b)
3. If $\sigma^2 = \lambda_8$ (small), then both the PC directions are stable (figure c)

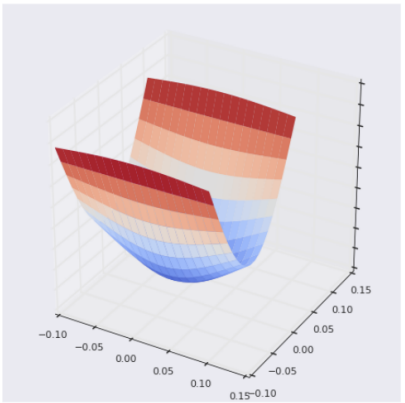
Lesson: Learning σ^2 is necessary for gaining full latent representation.



a) $\sigma^2 = \lambda_4$



b) $\sigma^2 = \lambda_6$



c) $\sigma^2 = \lambda_8$

Some unclear bits:

Two intriguing points:

1. Even though W_{MLE} is unique upto a rotation, pPCA is considered unidentifiable. Why?
2. $M^{-1} = (W^T W + \sigma^2 \mathbb{I})^{-1}$ is the pseudo inverse of $W^T W$ matrix. Is there a way to interpret its presence in the posterior, $p(z|x)$?
3. The mathematical analysis of stability of W is quite involved. Can we use the fact that $(\Lambda_k - \sigma^2 \mathbb{I})^{1/2}$ becomes imaginary to claim that those principal components are unstable?

Linear VAEs

For a generic VAE,

$$p(z) = \mathcal{N}(0, \mathbb{I})$$
$$p(x|z) = \mathcal{N}(\mu, \sigma^2)$$
$$q(z|x) = \mathcal{N}(\mu(x), \sigma^2(x))$$

For linear VAE,

$$p(z) = \mathcal{N}(0, \mathbb{I})$$
$$p(x|z) = \mathcal{N}(Wz + \mu, \sigma^2\mathbb{I})$$
$$q(z|x) = \mathcal{N}(V(x - \mu), D)$$

Here D is a diagonal covariance matrix, used globally for all datapoints.
The output of VAE, $\tilde{x}$ is distributed as,

$$\tilde{x}|x \sim \mathcal{N}(WV(x - \mu) + \mu, WDW^T)$$

The output of linear VAE is invariant under the following transformation,

$$W \leftarrow WA$$
$$V \leftarrow A^{-1}V$$
$$D \leftarrow A^{-1}DA^{-1}$$

where A is a diagonal matrix.

Linear VAEs: Lemma 1

Lemma 1

The global maximum of ELBO objective for the linear VAE, is identical to the global log marginal likelihood of pPCA

Proof: Recall, in pPCA posterior under MLE condition is given by,

$$p(z|x) = \mathcal{N}(M_{MLE}^{-1} W_{MLE}^T (x - \mu_{MLE}), \sigma_{MLE}^2 M_{MLE}^{-1}).$$

For linear VAE, the posterior is $q(z|x) = \mathcal{N}(V(x-\mu), D)$. If we set

$$V = M^{-1} W_{MLE}^T \quad D = \sigma_{MLE}^2 M^{-1} = \sigma_{MLE}^2 \Lambda_k^{-1} \quad \mu = \mu_{MLE}$$

Then, $q(z|x) = p(z|x)$. In addition if we make decoder weights, $W = W_{MLE}$, then we find $ELBO = \log p(x)$. ■
This result makes sense since everything in pPCA and linear VAE is Gaussian.

Linear VAEs: Corollary 1

Corollary 1

The global maximum of ELBO objective for the linear VAE has the scaled principal components as the columns of the decoder network.

Proof follows directly from lemma 1 where we found that that $W = W_{MLE} = U_k(\Lambda_k - \sigma^2_{MLE}\mathbb{I})$ ■

VAEs is trained by maximizing ELBO and not $p(x)$, so we next ask if ELBO objective introduces additional local maxima

(Answer: No)

Linear VAEs: Theorem 1

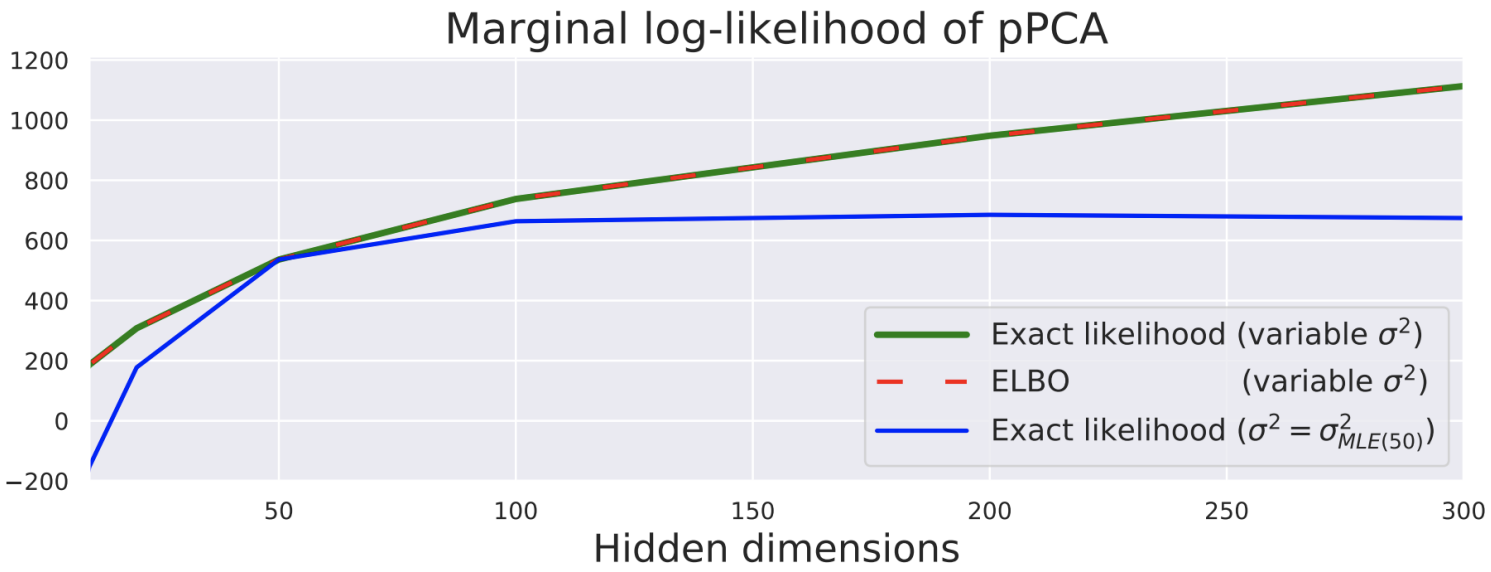
Theorem 1

The ELBO objective for a linear VAE does not introduce any additional local maxima to the pPCA model.

Experiment 1: Comparing ELBO and log-likelihood

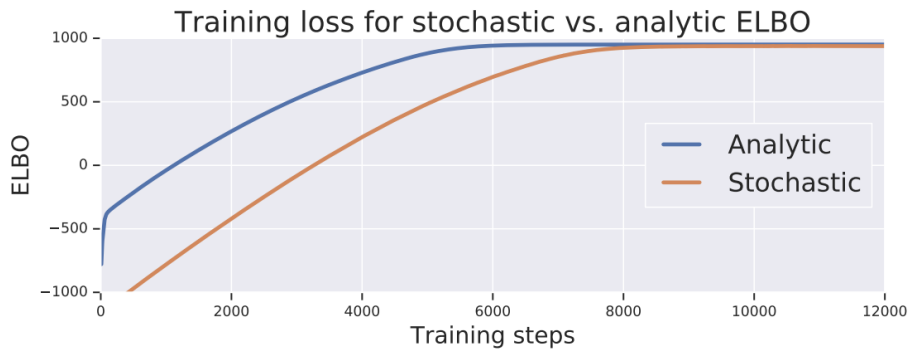
Train linear VAE on MNIST dataset for different number of latent dimensions and compute the ELBO. We observe the following:

1. ELBO(red)= p(x) when σ^2 is also trained. As expected from Lemma 1.
2. If σ^2 is fixed then $p(x)$ is smaller.



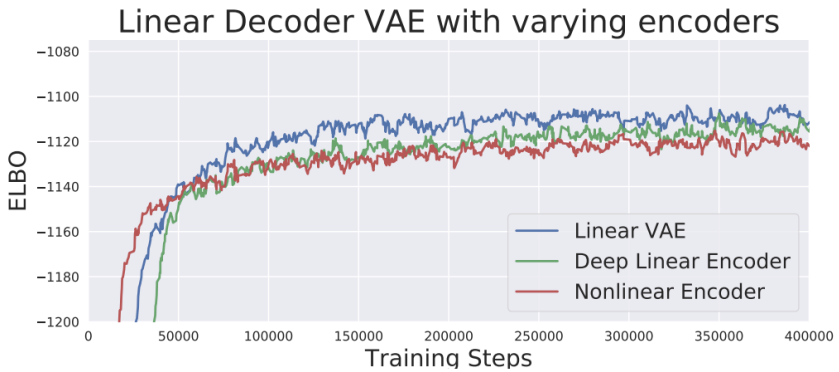
Experiment 2: Comparing analytic ELBO with stochastic ELBO

We can calculate ELBO exactly for linear VAE. They compare analytic ELBO expression with estimated ELBO and find that using analytic ELBO makes training faster.



Experiment 3: Effect of non-linear encoder and linear decoder

Deep linear encoder and non-linear encoders don't outperform linear encoder. This is expected since linear VAE is able to learn optimal posterior.



Experiment 4: Analysis of non-linear VAEs

Evaluation of deep Gaussian VAEs (averaged over 5 trials) on real-valued MNIST. We report the ELBO on the training set in all cases. Collapse percent gives the percentage of latent dimensions which are within 0.01 KL of the prior for at least 99

	Model		ELBO	σ^2-tuned ELBO	Tuned σ^2	Posterior collapse (%)	KL Divergence
	Init σ^2	Final σ^2					
MNIST	10.0		-1450.3 ± 4.2	-1098.2 ± 28.3	1.797	89.88	28.8 ± 1.4
	1.0		-1022.1 ± 5.4	-1018.3 ± 5.3	1.145	27.38	125.4 ± 4.2
	0.1		-3697.3 ± 493.3	-1190.8 ± 37.4	0.968	3.25	368.7 ± 94.6
	0.01		-38612.5 ± 1189.8	-2090.8 ± 975.1	0.877	0.00	695.9 ± 118.1
	0.001		-504259.1 ± 49149.8	-1744.7 ± 48.4	0.810	0.00	756.2 ± 12.6
	10.0	1.320	-1022.2 ± 4.5	-1022.3 ± 4.6	1.318	73.75	73.8 ± 9.8
	1.0	1.183	-1011.1 ± 2.7	-1011.1 ± 2.8	1.182	47.88	106.3 ± 2.5
	0.1	1.194	-1025.4 ± 8.6	-1025.4 ± 8.6	1.195	29.25	116.1 ± 11.4
	0.01	1.194	-1030.6 ± 3.5	-1030.5 ± 3.5	1.191	23.00	121.9 ± 7.7
	0.001	1.208	-1038.7 ± 5.6	-1038.8 ± 5.6	1.209	27.00	124.9 ± 1.6