Variational Autoencoder

11-785 Introduction to Deep Learning

- lecture 22 -

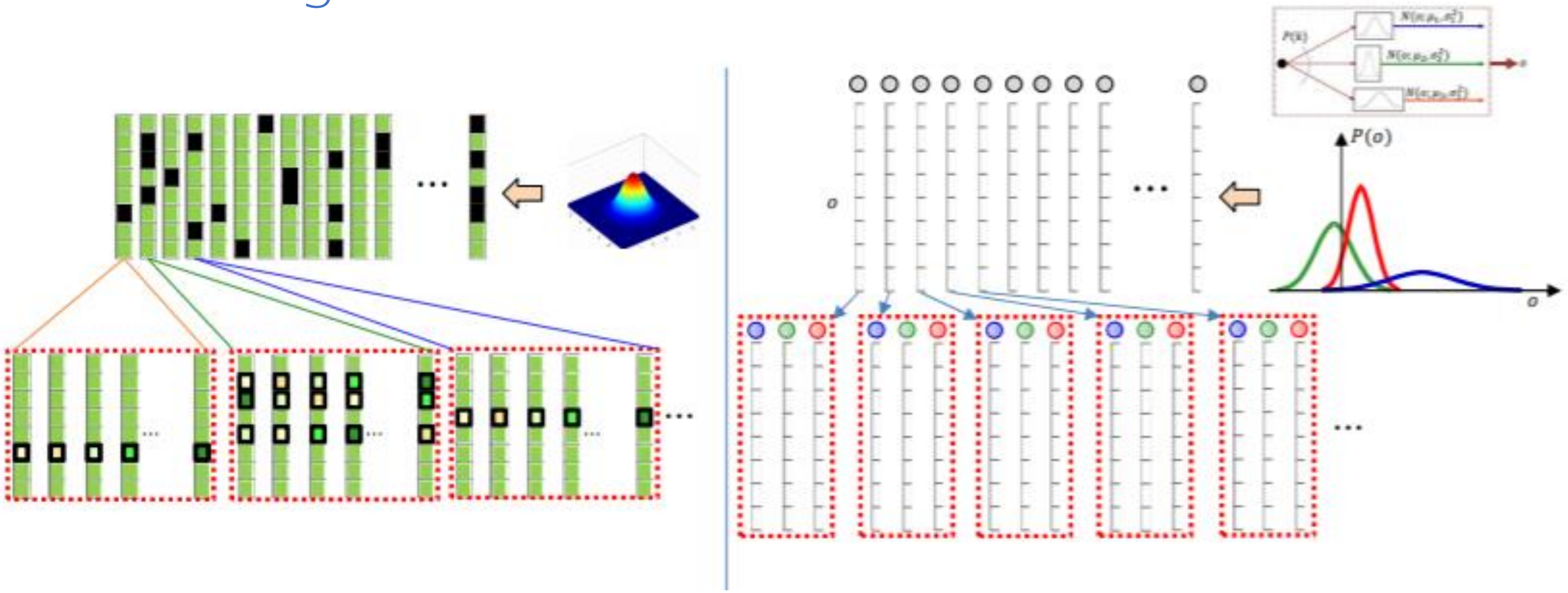
TAVE Research DL001
Heeji Won

Contents

Contents

0. Recap

EM algorithm for missing data



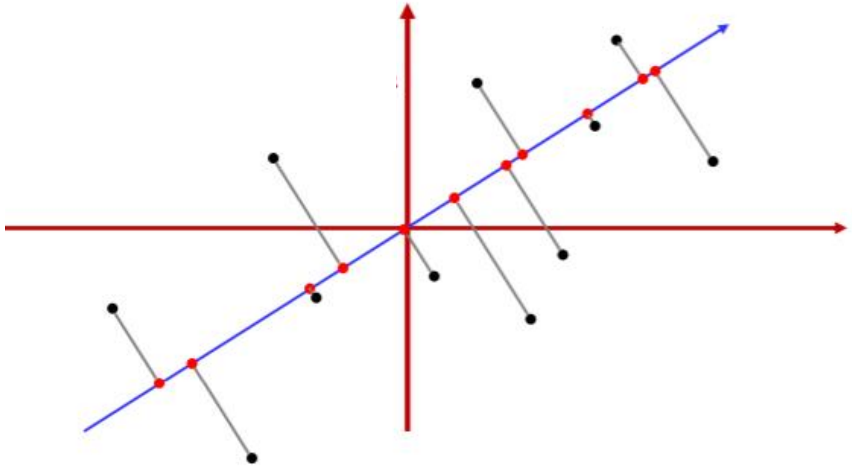
- ✓ Complete the data according to the posterior probabilities $P(h|o)$ computed by the current model
- ✓ Re−estimate the model from completed data

$$\underset{\theta}{\operatorname{argmax}} \sum_{o \in O} \log P(o;\theta)$$

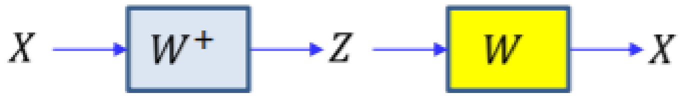
$$\log P(o;\theta) \geq \sum_{h} P(h|o;\theta^k)\ \log P(h,o;\theta) - \sum_h P(h|o;\theta^k) \log P(h|o;\theta^k)$$

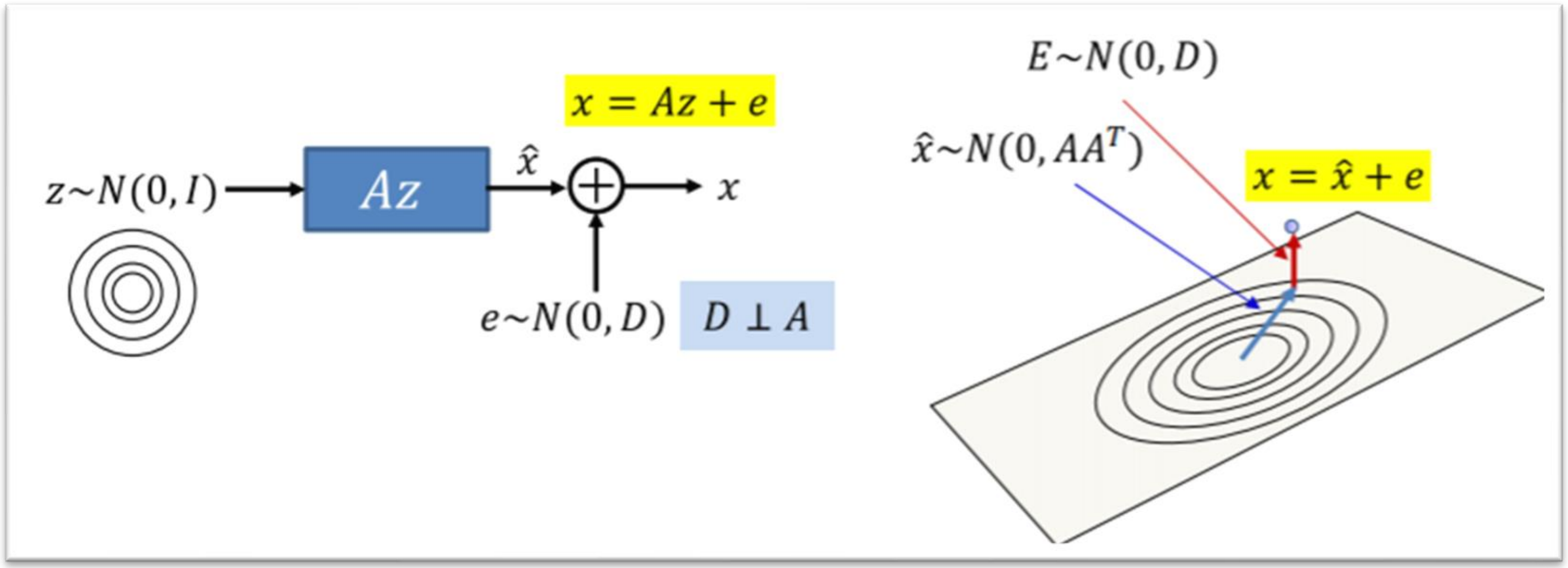
0. Recap

The generative story behind PCA



Find the principal subspace that can be projected which minimize the sum of the squared lengths

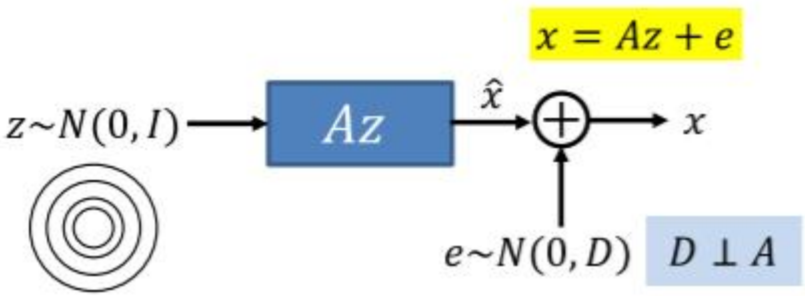




Contents

1. PCA

The probability modelled by PCA



$$\hat{x} = Az \Rightarrow \qquad P(\hat{x}) = N(0, AA^T)$$

$$x = \hat{x} + E \Rightarrow \qquad P(x) = N(0, AA^T + D)$$

ML estimation :

$$\underset{A,D}{\operatorname{argmax}} \sum_x \log \frac{1}{\sqrt{(2\pi)^d |AA^T + D|}} \exp(-0.5 x^T (AA^T + D)^{-1} x)$$

- ✓ PCA models a Gaussian distribution!
- ✓ But, we don' t know z

1. PCA

Missing information for PCA

- If we have complete information

$$x = Az + E$$
$$P(x|z) = N(Az, D)$$

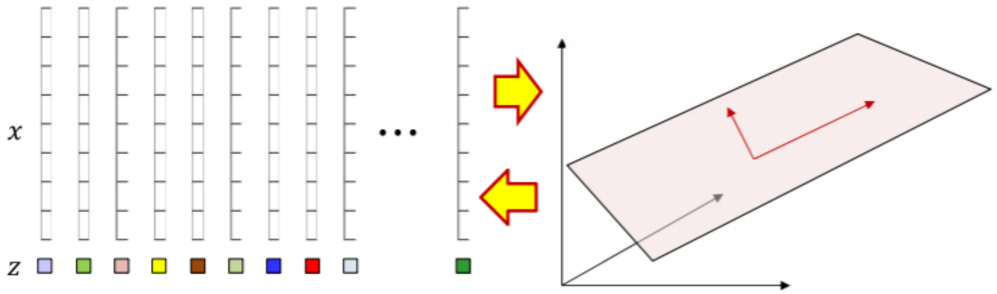
- Given complete information $(x_1,\ z_1), (x_2, z_2), \ldots,$ MSE is

$$\underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x,z) = \underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x|z)$$
$$= \underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} \log \frac{1}{\sqrt{(2\pi)^d |D|}} \exp(-0.5(x - Az)^T D^{-1}(x - Az))$$

=> $A = XZ^+$

But, we don't have z

- EM for PCA



- Initialize the plane
- 'Complete' the data by computing posterior prob.
- Re-estimate the plane

iterate

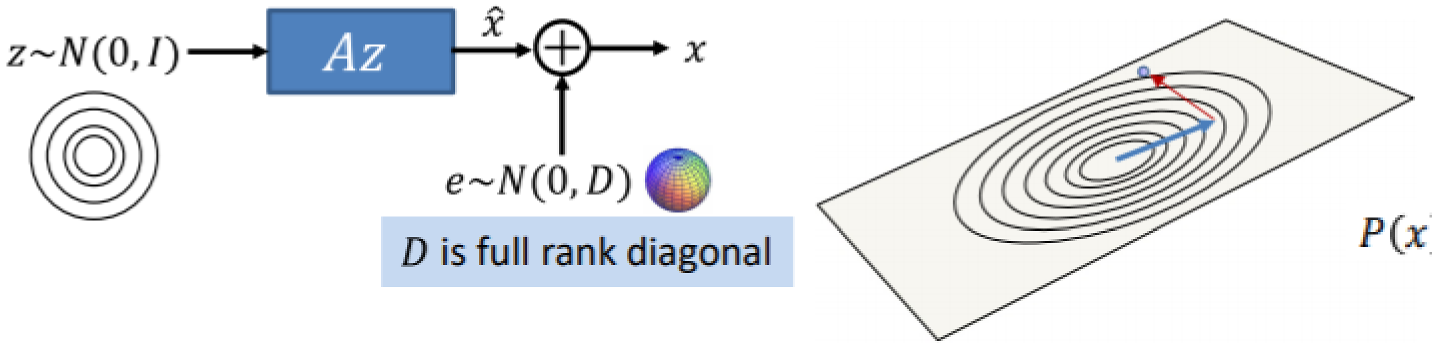
✓ PCA assumes the noise is orthogonal to the data

✓ Let's us generalize the model to permit non-orthogonal noise

Contents

2. The Linear Gaussian model

$x = Az + e$



$$P(x) = N(0, AA^T + D)$$

"Add full-rank Gaussian noise that is independent of the position on the hyperplane"

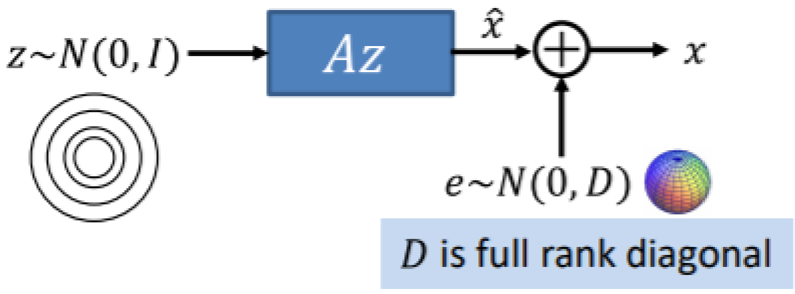
$$\underset{A,D}{\operatorname{argmax}} \sum_x \log \frac{1}{\sqrt{(2\pi)^d |AA^T + D|}} \exp(-0.5 x^T (AA^T + D)^{-1} x)$$

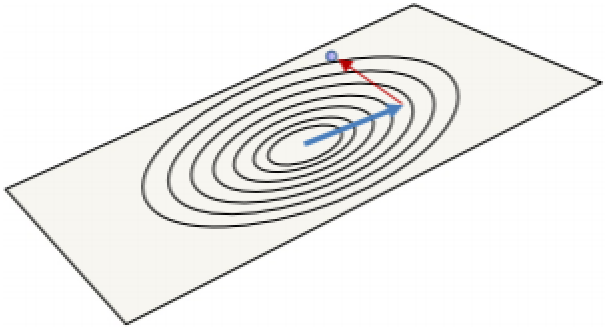
This doesn't have a nice closed form solution

2. The Linear Gaussian model

Missing information for LGMs

$x = Az + e$





We don't know z

- LGM with complete information

$$x = Az + e$$
$$P(x|z) = N(Az, D)$$

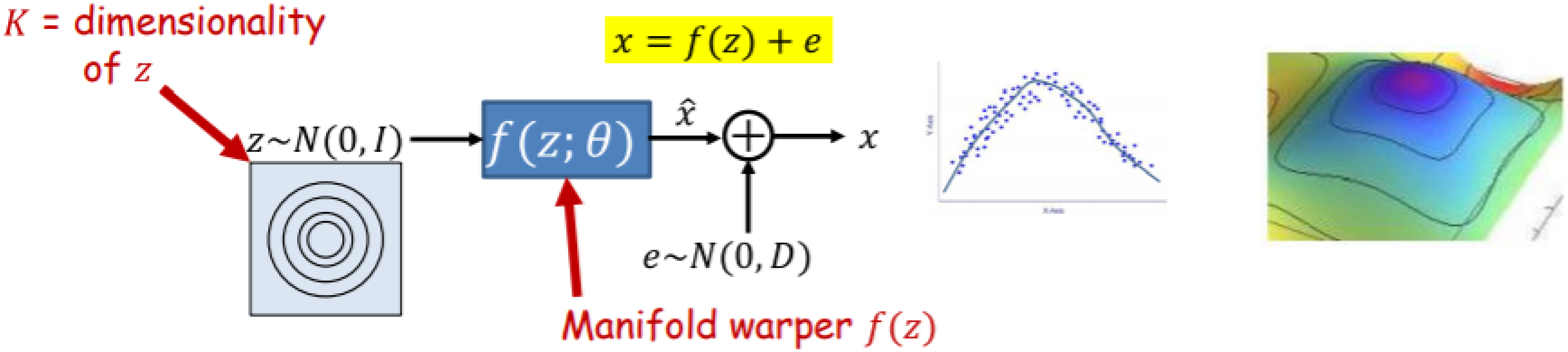
Given complete information X, Z

$$\underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x, z) = \underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x|z)$$

$$= \underset{A,D}{\operatorname{argmax}} \sum_{(x,z)} -\frac{1}{2} \log|D| - 0.5(x - Az)^T D^{-1} (x - Az)$$

Contents

3. The Non-linear Gaussian model

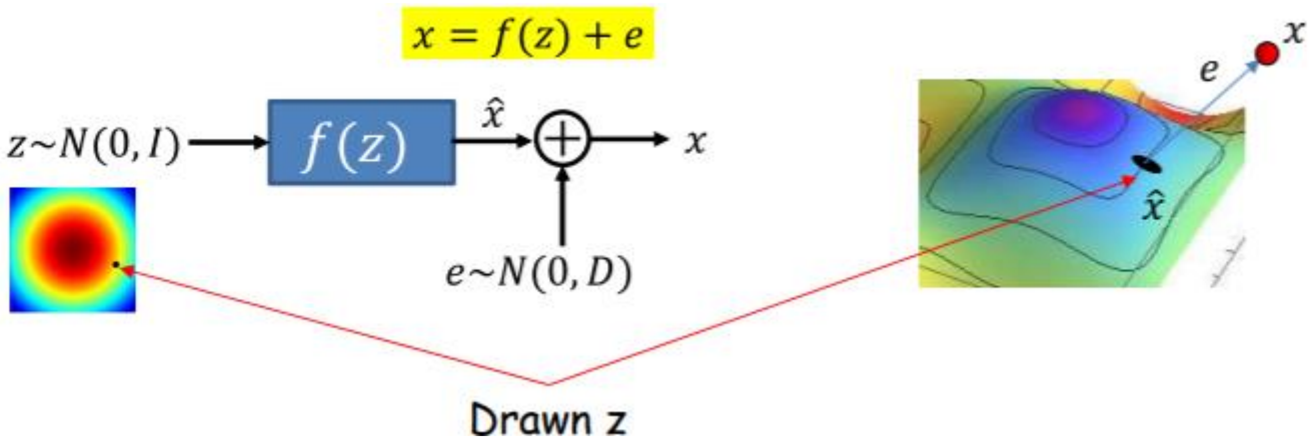


- $f(z)$ is a non-linear function that produces a curved manifold

- Key design issues
  - ✓ Select the dimensionality of the manifold
  - ✓ Choosing the right function $f(z)$ that is capable of learning the shape of the manifold

3. The Non-linear Gaussian model

Generating Process



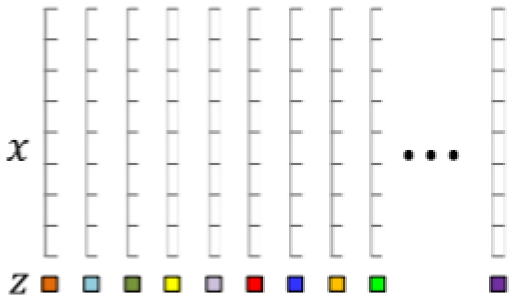
$$P(x|z) = N(x;\ f(z;\theta), D)$$

$$P(x) = \int_{-\infty}^{\infty} P(x|z)P(z)dz = \int_{-\infty}^{\infty} N(x;\ f(z;\theta), D)\ N(z; 0, D)\ dz$$

✓ $f(z;\theta)$ is not tractable, and cannot get a closed form for $P(x)$

3. The Non-linear Gaussian model

- MSE with complete information



$$x = f(z;\theta) + e$$
$$P(x|z) = N(f(z;\theta), D)$$

$$\theta^*, D^* = \underset{\theta,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x,z) \; = \underset{\theta,D}{\operatorname{argmax}} \sum_{(x,z)} \log P(x|z)$$

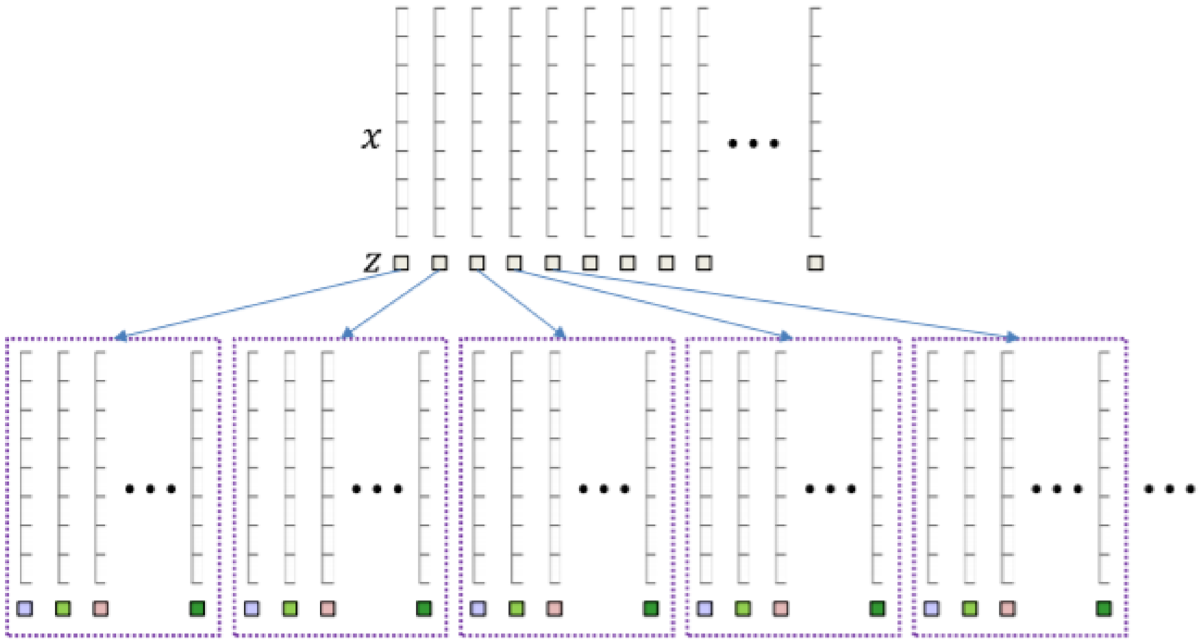
$$= \underset{\theta,D}{\operatorname{argmax}} \sum_{(x,z)} \underbrace{-\frac{1}{2}\log|D| - 0.5(x - f(z;\theta))^T D^{-1}(x - f(z;\theta))}_{L(\theta,D)}$$

$$\theta^*, D^* = \underset{\theta,D}{\operatorname{argmin}}\, L(\theta, D)$$

✓ But we don' t know z => EM algorithms!

3. The Non-linear Gaussian model

EM for NLGM



- Complete the data

Sol 1) In every possible way proportional to $P(z|x)$

Sol 2) By drawing samples from $P(z|x)$

$$P(x) = \int_{-\infty}^{\infty} P(x|z)P(z)dz = \int_{-\infty}^{\infty} N(x;\ f(z;\theta), D)\ N(z; 0, D)\ dz$$

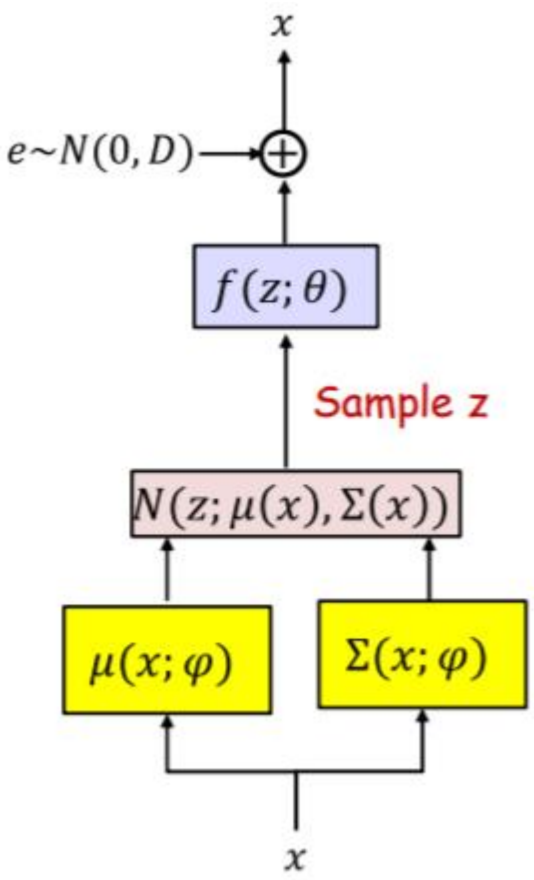
intractable!

$\Rightarrow$ $P(z|x)$ is intractable as a closed form solution

$\Rightarrow$ have to approximate $P(z|x)$

3. The Non-linear Gaussian model

Approximating $P(z|x)$



Approximate $P(z|x)$ as

$$P(z|x) \approx Q(z,x) = Gaussian\ N(z;\mu(x),\Sigma(x))$$

- initialize θ and φ

- Iterate:
  - ✓ Sample z from $N(z;\mu(x),\Sigma(x))$
  - ✓ Re-estimate θ from the entire
  - ✓ Estimate φ using the entire

3. The Non-linear Gaussian model

Re-estimate θ

$$L(\theta, D) = \sum_{(x,z)} \log|D| + (x - f(z;\theta))^T D^{-1} (x - f(z;\theta))$$

$$\theta^*, D^* = \underset{\theta, D}{\operatorname{argmin}}\, L(\theta, D)$$

$$L(\theta, \sigma^2) = d \log \sigma^2 + \sum_{(x,z)} \frac{1}{\sigma^2} \|x - f(z;\theta)\|^2$$

Estimate φ

"Estimate φ to minimize the error between $Q(z,x)$ and $P(z|x)$"

$$KL(Q(z,x)P(z|x)) = E_{z\sim Q} \log \frac{Q(z,x)}{P(z|x)}$$

$$= E_{z\sim Q} \log Q(z,x) - E_{z\sim Q} \log P(z) - E_{z\sim Q} \log P(x|z) + E_{z\sim Q} \log P(x)$$

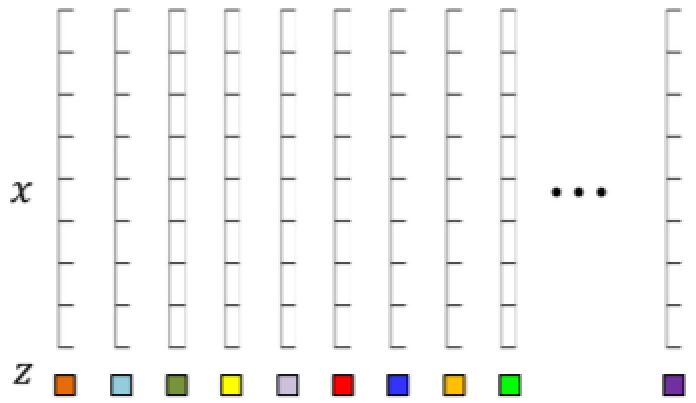
$$= KL\big(Q(z,x), P(z)\big) - E_{z\sim Q} \log P(x|z) + E_{z\sim Q} \log P(x)$$

$$\varphi^* = \underset{\varphi}{\operatorname{argmin}}\, KL(Q(z,x)P(z|x))$$

$$= \underset{\varphi}{\operatorname{argmin}}\, KL\big(Q(z,x), P(z)\big) - E_{z\sim Q} \log P(x|z)$$

3. The Non-linear Gaussian model

NLGM with complete data



– Given the completed data as $[X, Z] = \{(x, z)\}$

$$P(Z|X;\theta) = \prod_{(x,z)\in[X,Z]} P(z|x;\theta) \qquad \log P(Z|X;\theta) = \sum_{(x,z)\in[X,Z]} \log P(z|x;\theta)$$

$$Q(Z,X;\varphi) = \prod_{(x,z)\in[X,Z]} Q(z,x;\varphi) \qquad \log Q(Z,X;\varphi) = \sum_{(x,z)\in[X,Z]} \log Q(z,x;\varphi)$$

minimize the discrepancy

$$\log Q(Z,X;\varphi) - \log P(Z|X;\theta)$$

$$= \sum_{(x,z)\in[X,Z]} \log Q(z,x;\varphi) - \log P(z|x;\theta)$$

$$= \sum_{(x,z)\in[X,Z]} \log Q(z,x;\varphi) - \log P(z) - \log P(x|z;\theta) + \log P(x;\theta)$$

$$\Rightarrow \quad L_Q(\varphi) = \sum_{(x,z)\in[X,Z]} \log Q(z,x;\varphi) - \log P(z) - \log P(x|z;\theta)$$

How do we complete data?

3. The Non-linear Gaussian model

Complete the data

Sol1) Simply choose the samples you have already drawn by sampling

Sol2) Consider every possible value of z (to be more precise) -> can be computed in closed form

$$L_Q(\varphi) = \sum_{(x,z)\in[X,Z]} \underbrace{\log Q(z,x;\varphi) - \log P(z)} - \log P(x|z;\theta)$$

$$L_Q(\varphi) = \sum_{(x)\in[X]} \underbrace{\int_{-\infty}^{\infty} Q(z,x;\varphi)(\log Q(z,x;\varphi) - \log P(z))\, dz}_{KL(Q(z,x;\varphi),P(z))} - \sum_{(x,z)\in[X,Z]} \log P(x|z;\theta)$$

- We have

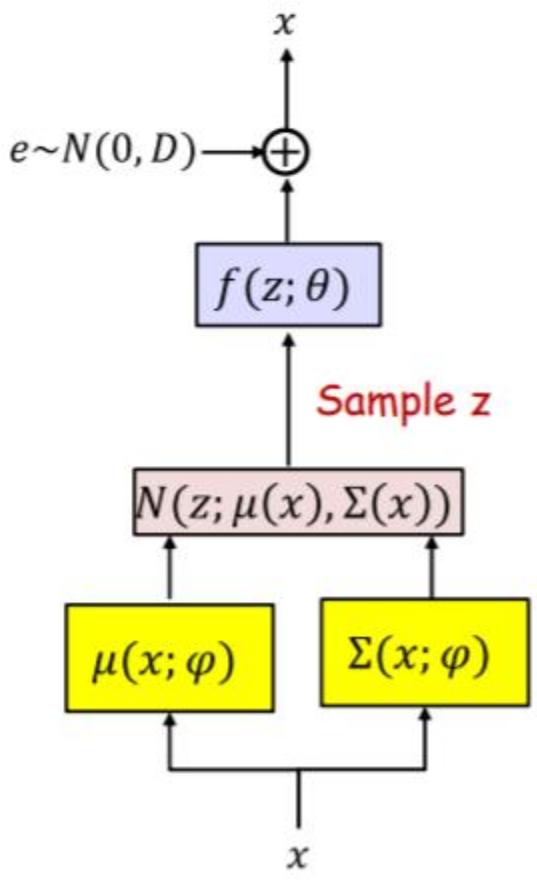
$$KL(Q(z,x;\varphi),P(z)) = \frac{1}{2}\Big(tr\big(\Sigma(x;\varphi)\big) + \mu(x;\varphi)^T(\mu(x;\varphi) - d - \log|\Sigma(x;\varphi)|)\Big)$$

$$\log P(x|z;\theta) = \sum_{(x,z)} -\frac{1}{2}\log|D| - 0.5(x - f(z;\theta))^T D^{-1}(x - f(z;\theta))$$

$$\boxed{\begin{aligned} L_Q(\varphi) &= \sum_{x\in X} \frac{1}{2}\Big(tr\big(\Sigma(x;\varphi)\big) + \mu(x;\varphi)^T(\mu(x;\varphi) - d - \log|\Sigma(x;\varphi)|)\Big) + \sum_{(x,z)\in[X,Z]} \frac{1}{2}\log|D| + 0.5(x - f(z;\theta))^T D^{-1}(x - f(z;\theta)) \\ &= \sum_{x\in X} \Big(tr\big(\Sigma(x;\varphi)\big) + \mu(x;\varphi)^T(\mu(x;\varphi) - d - \log|\Sigma(x;\varphi)|)\Big) + \frac{1}{\sigma^2} \sum_{(x,z)\in[X,Z]} \|(x - f(z;\theta))\|^2 \end{aligned}}$$

3. The Non-linear Gaussian model

The complete training pipeline



- initialize θ and φ
- Iterate:
  - Sample z from $N\big(z;\mu(x),\Sigma(x)\big)$
  - Re-estimate θ from the entire data
  - Estimate φ

$$L(\theta,\sigma^2) = d\log\sigma^2 + \frac{1}{\sigma^2}\sum_{(x,z)} \|x - f(z;\theta)\|^2$$

$$L_Q(\varphi) = \sum_{x\in X}\Big(tr\big(\Sigma(x;\varphi)\big) + \mu(x;\varphi)^T\big(\mu(x;\varphi) - d - \log|\Sigma(x;\varphi)|\big)\Big) + \frac{1}{\sigma^2}\sum_{(x,z)\in[X,Z]} \|(x - f(z;\theta))\|^2$$

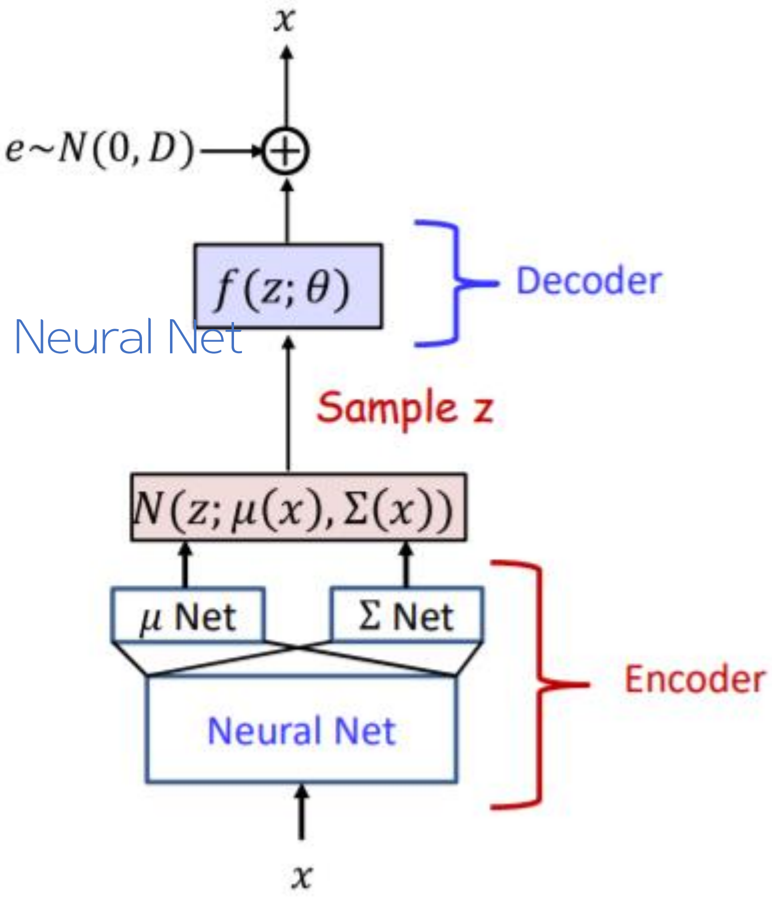
Single Step

- initialize θ and φ
- Iterate:
  - Sample z from $N\big(z;\mu(x),\Sigma(x)\big)$
  - Re-estimate θ and φ

$$L(\theta,\sigma^2,\varphi) = \sum_{x\in X}\Big(tr\big(\Sigma(x;\varphi)\big) + \mu(x;\varphi)^T\big(\mu(x;\varphi) - d - \log|\Sigma(x;\varphi)|\big)\Big) + \frac{1}{\sigma^2}\sum_{(x,z)\in[X,Z]} \|(x - f(z;\theta))\|^2 + d\log\sigma^2$$

Contents

4. The Variational Autoencoder



- The decoder is the actual generative model
- The encoder is primarily needed for training
- z is a latent-space representation of the data which captures underlying structure in the data x

- VAEs are strictly generative models
- But, they cannot be used to compute the likelihood of data
- Nevertheless, they are highly effective as generators

4. The Variational Autoencoder

Conclusions

- ✓ A simple non−linear extensions of linear Gaussian models
- ✓ Excellent generative models for the distribution of data $P(x)$
- ✓ Have also been successfully embedded into dynamical system models
  - P(z) now becomes a mixture, or a Markov model instead of N(0, I)

Thank you