Latent Variable Models for Dimensionality Reduction

Amit Kumar

Machine Learning and Computing
Department of Mathematics
Indian Institute of Space Science and Technology, Trivandrum

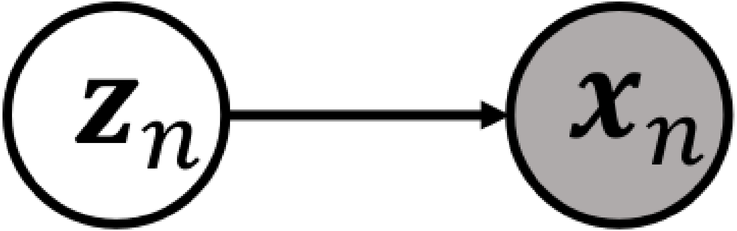
July 15, 2021

Overview

1. Latent Variable Model
2. Probabilistic PCA (PPCA) Intro
3. EM algorithm
4. PPCA

Latent Variable

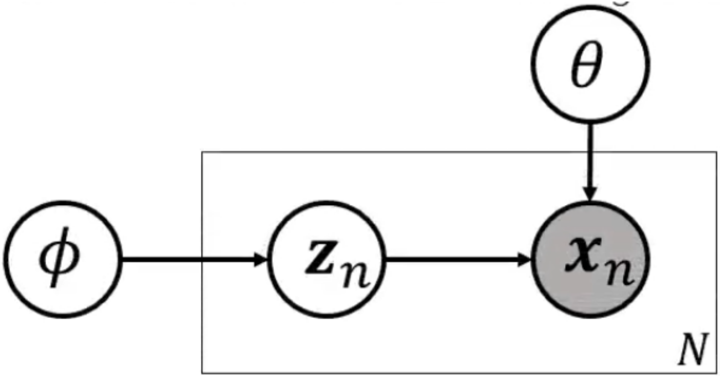
- In generative model in unsupervised learning each data point x_n is associated with a latent variable.



- Clustering: The cluster id z_n (discrete, or a K-dim one-hot rep, or a vector of cluster membership probabilities)
- Dimensionality reduction: The low-dim representation $z_n \in R^K$

Generative Models with Latent Variables

- A typical generative model with latent variables might look like this

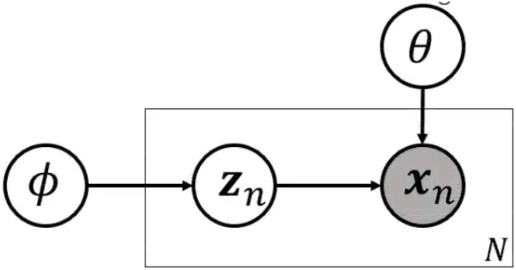


Latent Variable Model

- $p(z_n|\,\phi)$: A suitable distribution based on the nature of z_n.
- $p(x_n|z_n,\,\theta)$: A suitable distribution based on the nature of x_n.
- In this generative model, observations x_n assumed generated via latent variables z_n.
- The unknowns in such latent var models (LVMs) are of two types
  - Global variables: Shared by all data points (θ and ϕ in the previous diagram)
  - Local variables: Specific to each data point (z_n 's in the previous diagram)

Parameter Estimation for Generative LVM

- how do we estimate the parameters of a generative LVM?



- we can make a guess what the value of each z_n and then estimate θ and ϕ.
- The guess about z_n can be in one of the two forms
  - A "hard" guess – a fixed value (some "optimal" value of the random variable z_n)
  - The "expected" value $\mathbb{E}[z_n]$ of the random variable z_n.

Parameter Estimation for Generative LVM

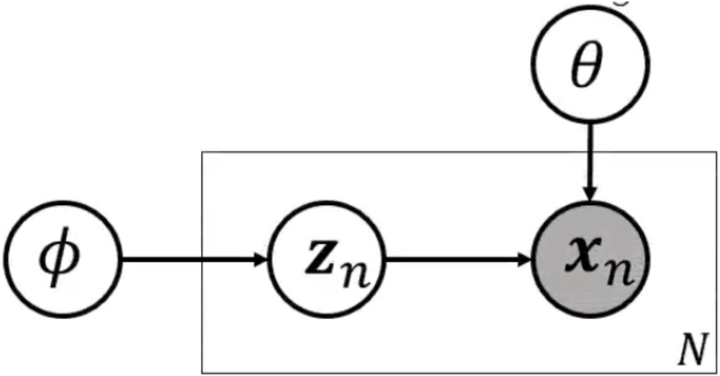
- Can we estimate parameters $(\theta\ ,\ \phi) = \Theta$ (say) of an LVM without estimating z_n ?
- In principle yes, but it is harder
- Given N observations, x_n, $n = 1,2, \ldots\ ,$ N, the MLE problem for Θ will be

$$\underset{\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\Theta) \quad = \underset{\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)$$

- $\mathrm{p}(x_n,z_n|\Theta) = \mathrm{p}(z_n|\phi)\ \mathrm{p}(x_n \mid z_n\ ,\ \theta)$
- The log of sum doesn't give us a simple expression; MLE can still be done using gradient based methods but update will be complicated. ALT-OPT or EM make it simpler by using guesses of z_n's

Probabilistic Principal Component Analysis (PPCA)

- Probabilistic PCA (PPCA) is example of a generative latent var model
- Assume a K-dim latent var z_n mapped to a D-dim observation x_n via a prob. mapping



- p($z_n|\phi$)= $\mathcal{N}(0,I_k)$

Probabilistic mapping

- Probabilistic mapping means that will be not exactly but somewhere around the mean.

$$\underset{D \times 1 \text{ mean of the mapping}}{\boldsymbol{\mu}_n} = \overset{D \times K \text{ mapping matrix}}{\mathbf{W}} \overset{K \times 1}{\mathbf{z}_n}$$

$$p(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{W}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_n, \sigma^2 \mathbf{I}_D)$$

- Instead of a linear mapping Wz_n , the z_n to x_n mapping can be defined as a nonlinear mapping (variational autoencoders, kernel based latent variable models).

PPCA over PCA

- PPCA has several benefits over PCA, some of which include
  - Can use suitable distributions for x n to better capture properties of data.
  - Parameter estimation can be done faster without eigen-decomposition (using ALT-OPT/EM algos)
  - In PCA, eigen vector are orthogonal but in ppca we don't have such requirement.
- If the z_n were known, it just becomes a probabilistic version of the multi-output regression problem.

Need for EM

- Consider an LVM with latent variables and parameters. Trying to estimate parameters without also estimating the latent variables (by marginalizing them) is difficult.
- Consider a complex prob. density (without any latent vars) for which MLE is hard.

What is EM Doing?

- The MLE problem was $\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \log p(\boldsymbol{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \log \sum_z p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)$ which is Incomplete data log likelihood
- What EM (and ALT-OPT in a crude way) did is max of CLL:

$$\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \mathbb{E}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)]$$

- But we did not solve the original problem. Is it okay?
- Assume $p_z = \mathrm{p}(Z|X, \Theta)$ and q(Z) to be some prob distribution over Z, then

$$\log p(\boldsymbol{X}|\Theta) = \mathcal{L}(q, \Theta) + KL(q||p_z)$$

EM

- $\mathcal{L}(q,\Theta) = \sum_Z q(Z)\log\left\{\frac{p(X,Z|\Theta)}{q(Z)}\right\}$ and $KL(q||p_z) = -\sum_Z q(\boldsymbol{Z})\log\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{X},\Theta)}{q(\boldsymbol{Z})}\right\}$
- Since KL is always non-negative log(p X|)$\geq$ L (q, Θ), so L is a lower-bound on ILL.
- Thus if we maximize L (q, Θ), it will also improve log(p X|)
- Let's maximize L (q, Θ) w.r.t. q with Θ fixed at Θold
- $\hat{q} = \operatorname{argmax}_q \mathcal{L}(q,\Theta^{\text{old}}) = \operatorname{argmin}_q KL(q||p_z) = p_z = p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}})$
- Now let's maximize L (q, Θ) w.r.t. Θ with q fixed.

$$
\begin{aligned}
\Theta^{\text{new}} = \operatorname{argmax}_{\Theta} \mathcal{L}(\hat{q}, \Theta) &= \operatorname{argmax}_{\Theta} \sum_{Z} p(\boldsymbol{Z}|\boldsymbol{X}, \Theta^{\text{old}}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)}{p(\boldsymbol{Z}|\boldsymbol{X}, \Theta^{\text{old}})} \right\} \\
&= \operatorname{argmax}_{\Theta} \sum_{Z} p\left(\boldsymbol{Z}\middle|\boldsymbol{X}, \Theta^{\text{old}}\right) \log p(\boldsymbol{X}, \boldsymbol{Z}|\Theta) \\
&= \operatorname{argmax}_{\Theta} \mathbb{E}_{p\left(\boldsymbol{Z}\middle|\boldsymbol{X}, \Theta^{\text{old}}\right)}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)]
\end{aligned}
$$

The EM Algorithm in its general form

- Maximization of L (q, Θ) w.r.t. q and Θ gives the EM algorithm (Dempster, Laird, Rubin, 1977) constituents.

The EM Algorithm

1. Initialize Θ as $\Theta^{(0)}$, set $t = 1$
2. Step 1: Compute posterior of latent variables given current parameters $\Theta^{(t-1)}$

$$p(\boldsymbol{z}_n^{(t)}|\boldsymbol{x}_n, \Theta^{(t-1)}) = \frac{p(\boldsymbol{z}_n^{(t)}|\Theta^{(t-1)})p(\boldsymbol{x}_n|\boldsymbol{z}_n^{(t)}, \Theta^{(t-1)})}{p(\boldsymbol{x}_n|\Theta^{(t-1)})} \propto \text{prior} \times \text{likelihood}$$

3. Step 2: Now maximize the expected complete data log-likelihood w.r.t. Θ

$$\Theta^{(t)} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{(t-1)}) = \arg\max_{\Theta} \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n^{(t)}|\boldsymbol{x}_n, \Theta^{(t-1)})}[\log p(\boldsymbol{x}_n, \boldsymbol{z}_n^{(t)}|\Theta)]$$

4. If not yet converged, set $t = t + 1$ and go to step 2.

The Expected CLL

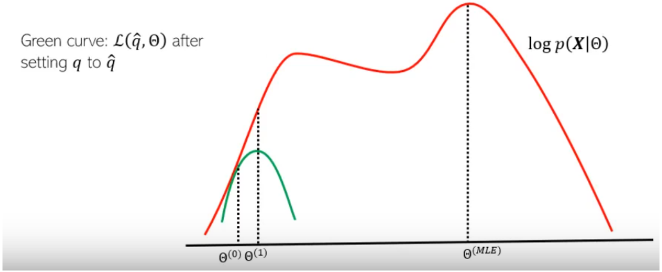
- Expected CLL in EM is given by (assume observations are i.i.d.)

$$
\begin{aligned}
\mathcal{Q}(\Theta, \Theta^{old}) &= \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n,\Theta^{old})}[\log p(x_n, z_n|\Theta)] \\
&= \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n,\Theta^{old})}[\log p(x_n|z_n, \Theta) + \log p(z_n|\Theta)]
\end{aligned}
$$

- In resulting expressions, replace terms containing z_n's by their respective expectations, e.g.,
  - $\mathbf{z}_n$ replaced by $\mathbb{E}_{p(\mathbf{z}_n|\boldsymbol{x}_n, \widehat{\Theta})}[\mathbf{z}_n]$
  - $\mathbf{z}_n\mathbf{z}_n^\top$ replaced by $\mathbb{E}_{p(\mathbf{z}_n|\boldsymbol{x}_n, \widehat{\Theta})}[\mathbf{z}_n\mathbf{z}_n^\top]$
- However, in some LVMs, these expectations are intractable to compute and need to be approximated (beyond the scope of this presentation)

EM: An Illustration

- As we saw, EM maximizes the lower bound L (q, Θ) in two steps
- Step 1 finds the optimal q setting it the posterior of Z given current Θ
- Step 2 maximizes L (q, Θ) w.r.t. Θ which gives a new Θ.



Probabilistic PCA (PPCA)

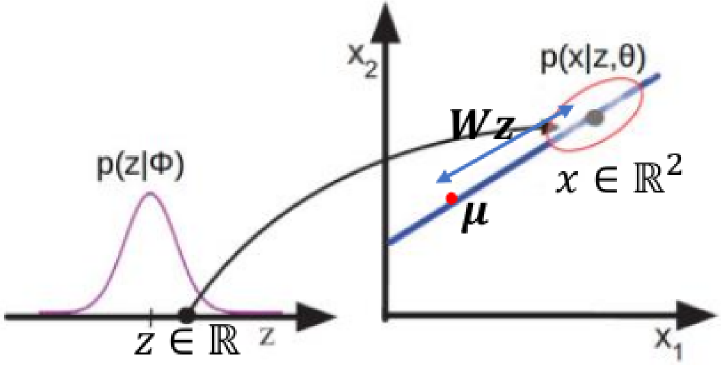
- Assume obs $x_n \in R^D$ as a linear mapping of a latent var $z_n \in R^K$ + Gaussian noise

$$\boldsymbol{x}_n = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\epsilon}_n$$

- Equivalent to saying $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\mu}, \boldsymbol{W}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z}_n, \sigma^2 I_D)$
- Assume a zero-mean Gaussian prior on z_n
- Joint distr. of x_n and $z_n$is Gaussian (since p($x_n|z_n$) and p(z_n) are individually Gaussian) and the marginal distribution of x_n will be Gaussian.

$$p(\boldsymbol{x}_n|\boldsymbol{W}, \sigma^2) = N(\boldsymbol{x}_n\ |\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 I_D)$$

Pictorial



Learning PPCA using EM

- Ignoring for notational simplicity, ILL is

$$p(\boldsymbol{x}_n|\boldsymbol{W},\sigma^2) = N(\boldsymbol{x}_n\,|\boldsymbol{0},\boldsymbol{W}\boldsymbol{W}^\top + \sigma^2 I_D)$$

- Can maximize ILL but requires solving eigen-decomposition (PRML: 12.2.1)
- EM will instead maximize expected CLL, with CLL given by

$$\log p(\mathbf{X},\mathbf{Z}|\mathbf{W},\sigma^2) = \log\prod_{n=1}^{N} p(\mathbf{x}_n,\mathbf{z}_n|\mathbf{W},\sigma^2) = \log\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n,\mathbf{W},\sigma^2)p(\mathbf{z}_n) = \sum_{n=1}^{N}\{\log p(\mathbf{x}_n|\mathbf{z}_n,\mathbf{W},\sigma^2) + \log p(\mathbf{z}_n)\}$$

Using $p(\mathbf{x}_n|\mathbf{z}_n,\mathbf{W},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{D/2}}\exp\left[-\frac{(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n)^\top(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n)}{2\sigma^2}\right]$, $p(\mathbf{z}_n)\propto\exp\left[-\frac{\mathbf{z}_n^\top\mathbf{z}_n}{2}\right]$ and simplifying

$$\text{CLL} = -\sum_{n=1}^{N}\left\{\frac{D}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||\mathbf{x}_n||^2 - \frac{1}{\sigma^2}\mathbf{z}_n^\top\mathbf{W}^\top\mathbf{x}_n + \frac{1}{2\sigma^2}\text{tr}(\mathbf{z}_n\mathbf{z}_n^\top\mathbf{W}^\top\mathbf{W}) + \frac{1}{2}\text{tr}(\mathbf{z}_n\mathbf{z}_n^\top)\right\}$$

Learning PPCA using EM

- The EM algo for PPCA alternates between two steps:
  - Compute conditional posterior of z_n given parameters Θ

$$p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{W}, \sigma^2) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top \mathbf{x}_n, \sigma^2 \mathbf{M}^{-1}) \qquad \text{(where } \mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2 \mathbf{I}_K\text{)}$$

  - Maximize the expected CLL

$$-\sum_{n=1}^{N} \left\{ \frac{D}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||\mathbf{x}_n||^2 - \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^\top \mathbf{W}^\top \mathbf{x}_n + \frac{1}{2\sigma^2}\text{tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\mathbf{W}^\top\mathbf{W}) + \frac{1}{2}\text{tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]) \right\}$$

- Taking derivative of expected CLL w.r.t. W and setting to zero gives

$$\boxed{\mathbf{W} = \left[\sum_{n=1}^{N} \mathbf{x}_n \mathbb{E}[\mathbf{z}_n]^\top\right] \left[\sum_{n=1}^{N} \mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\right]^{-1}}$$

- Required expectations can be found from the conditional posterior of z_n

$$
\begin{aligned}
p(\boldsymbol{z}_n|\boldsymbol{x}_n, \mathbf{W}) &= \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top \boldsymbol{x}_n, \sigma^2\mathbf{M}^{-1}) \qquad \text{where } \mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}_K \\
\mathbb{E}[\boldsymbol{z}_n] &= \mathbf{M}^{-1}\mathbf{W}^\top \boldsymbol{x}_n \\
\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top] &= \mathbb{E}[\boldsymbol{z}_n]\mathbb{E}[\boldsymbol{z}_n]^\top + \text{cov}(\boldsymbol{z}_n) = \mathbb{E}[\boldsymbol{z}_n]\mathbb{E}[\boldsymbol{z}_n]^\top + \sigma^2\mathbf{M}^{-1}
\end{aligned}
$$

Full EM algo for PPCA

- Specify K, initialize $\mathbf{W}$ and σ^2 randomly. Also center the data ($\boldsymbol{x}_n = \boldsymbol{x}_n - \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$)
- **E step:** For each n, compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ using current $\mathbf{W}$ and σ^2. Compute exp. for the M step

$$
\begin{aligned}
p(\boldsymbol{z}_n|\boldsymbol{x}_n, \mathbf{W}) &= \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top\boldsymbol{x}_n, \sigma^2\mathbf{M}^{-1}) \qquad \text{where } \mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}_K \\
\mathbb{E}[\boldsymbol{z}_n] &= \mathbf{M}^{-1}\mathbf{W}^\top\boldsymbol{x}_n \\
\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top] &= \text{cov}(\boldsymbol{z}_n) + \mathbb{E}[\boldsymbol{z}_n]\mathbb{E}[\boldsymbol{z}_n]^\top = \mathbb{E}[\boldsymbol{z}_n]\mathbb{E}[\boldsymbol{z}_n]^\top + \sigma^2\mathbf{M}^{-1}
\end{aligned}
$$

- **M step:** Re-estimate $\mathbf{W}$ and σ^2

$$
\begin{aligned}
\mathbf{W}_{new} &= \left[\sum_{n=1}^{N}\boldsymbol{x}_n\mathbb{E}[\boldsymbol{z}_n]^\top\right]\left[\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]\right]^{-1} \\
\sigma^2_{new} &= \frac{1}{ND}\sum_{n=1}^{N}\left\{||\boldsymbol{x}_n||^2 - 2\mathbb{E}[\boldsymbol{z}_n]^\top\mathbf{W}_{new}^\top\boldsymbol{x}_n + \text{tr}\left(\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]\mathbf{W}_{new}^\top\mathbf{W}_{new}\right)\right\}
\end{aligned}
$$

- Set $\mathbf{W} = \mathbf{W}_{new}$ and $\sigma^2 = \sigma^2_{new}$. If not converged (monitor $p(\mathbf{X}|\Theta)$), go back to E step

Thank You