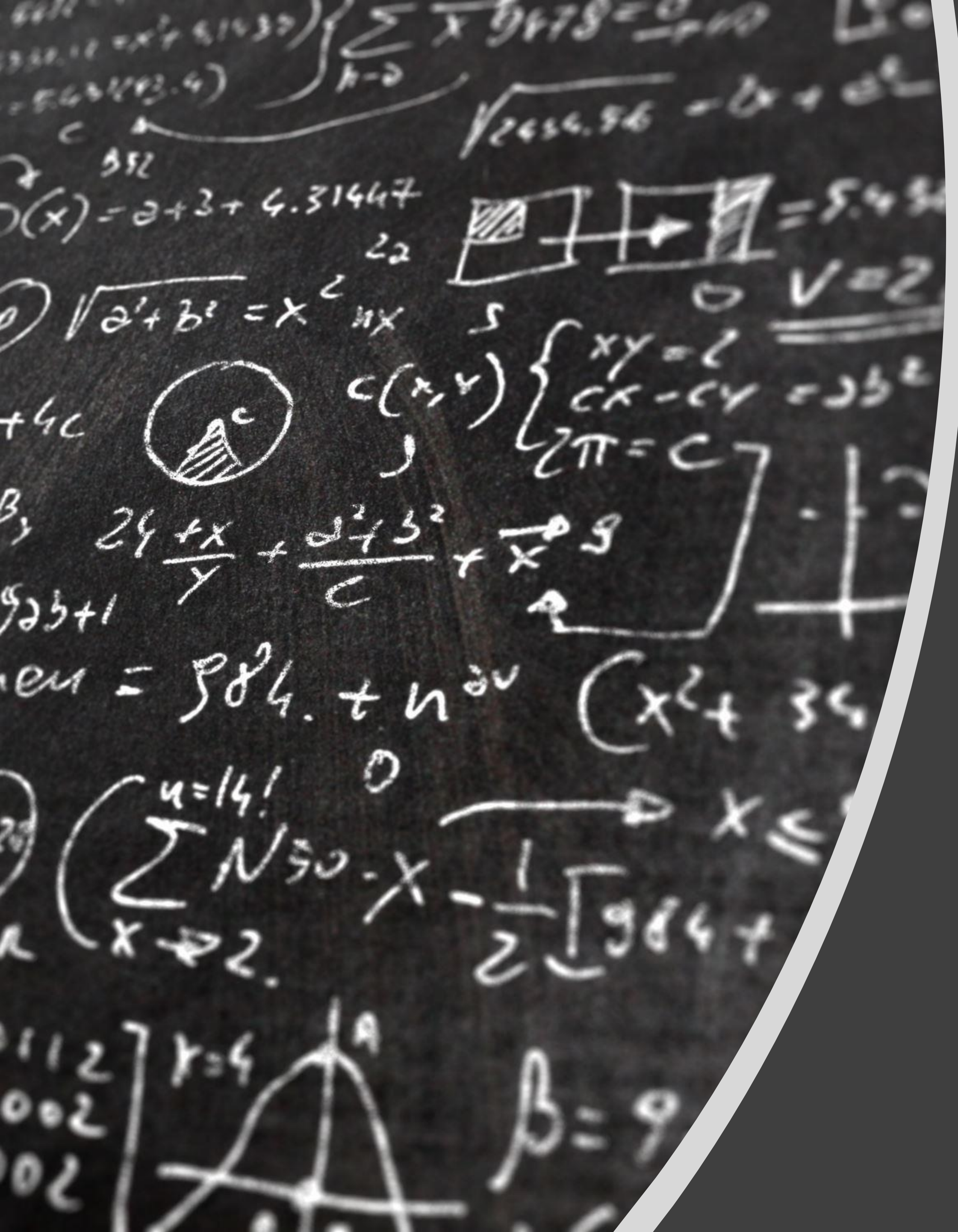


CS5340: Tutorial 6

Asst. Prof. Harold Soh

TA: Eugene Lim

Questions?

https://pollev.com/elim360

SCAN ME

The General EM Algorithm

Goal: $\max_{\theta}$ maximize $\log p(D|\theta)$

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(Z|X, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

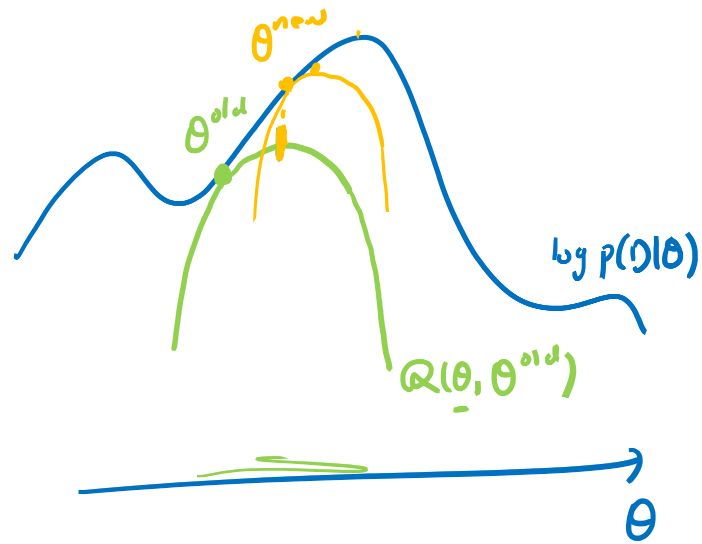
$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

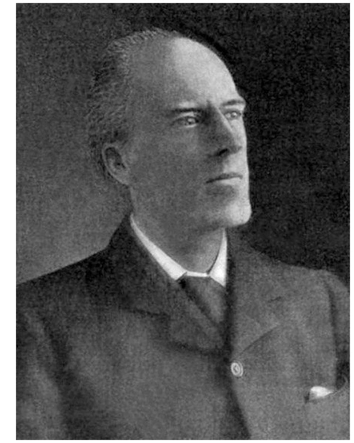
4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$



Principal Components Analysis (PCA)

- Invented in 1901 by Karl Pearson
  - Independently by Hoteling in 1930s.
- Unsupervised Learning method
- Useful for:
  - Representation learning
  - Dimensionality reduction
  - Compression
  - Data-preprocessing
  - Visualization



Karl Pearson, 1912
(image credit: Wikipedia)

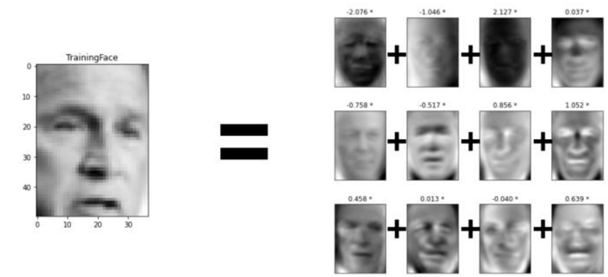


Image Credit: https://www.geeksforgeeks.org/ml-face-recognition-using-eigenfaces-pca-algorithm/

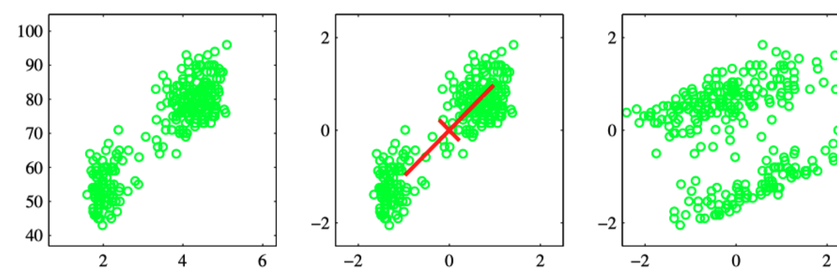
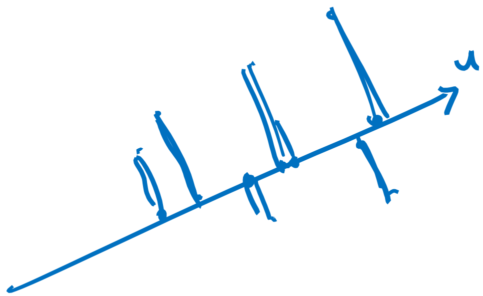


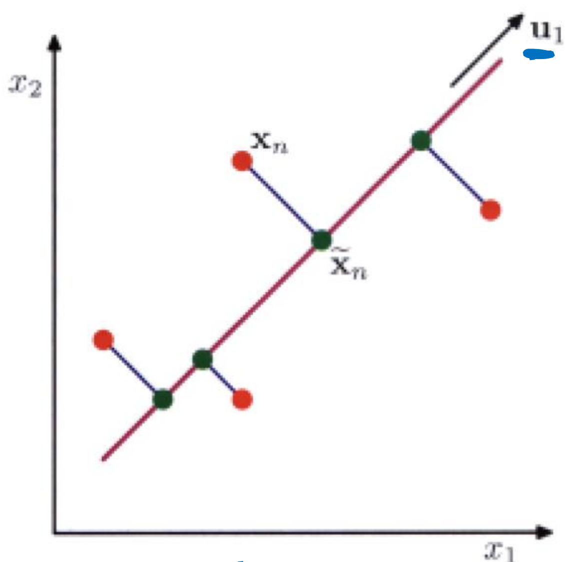
Image Credit: PRML Chp 12

PCA Setup and Intuition

- Dataset of D-dimensional data points $\mathbf{x}_i$
- Want to associate each data point $\mathbf{x}_i$ with a corresponding M-dimensional point $\mathbf{z}_i$
  - where $M < D$
- 2 approaches to derivation. Project to:
  - Maximize variance
  - Minimize distortion
- In practice, we compute $\mathbf{X}\mathbf{X}^\top$ and find the M largest eigenvectors and eigenvalues



Maximizing Variance



$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

$u^T\bar{x} = \frac{1}{N}\sum_{n=1}^{N} u^T x_n$

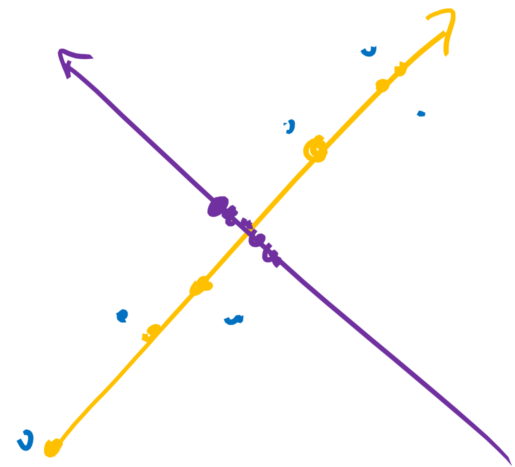
$\frac{1}{N}\sum_{n=1}^{N} \left(u^T x_n - u^T \bar{x} \right)^2$

$= u^T S u$

$S = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$

$\frac{d}{du}\left[u^T S u - \lambda(u^T u - 1)\right] = 2Su - 2\lambda u \overset{set}{:=} 0$

$\Rightarrow Su = \lambda u$

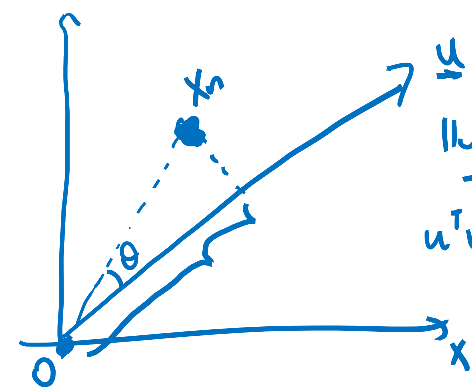


$u^T u = 1$

$u^T u - 1 = 0$

$g(x) = 0$

$f(x) + \lambda g(x)$



$\|u\| = 1$

$u^T u = 1$

$u^T x_n = |u||x_n|\cos\theta$

$= |x_n|\cos\theta$

Image Credit: PRML Chp 12

$$\frac{1}{N}\sum_{n=1}^{N}\left(u^T x_n - u^T \bar{x}\right)^2 = \frac{1}{N}\sum_{n=1}^{N}\left[u^T(x_n-\bar{x})\right]^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left[(x_n-\bar{x})^T u\right]^2$$

$$= \frac{1}{N}\sum_{n=1}^{N} u^T(x_n-\bar{x})(x_n-\bar{x})^T u$$

$$= u^T \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(x_n-\bar{x})^T\right)}_{S} u$$

$$X = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix}$$

Probabilistic PCA (PPCA)

- Derive Probabilistic variant
- Learn via EM
- Advantages:
  - Efficient EM algorithm (avoids computing $\mathbf{X}\mathbf{X}^\top$)
  - Naturally deal with missing data
  - Can be extended to include class labels, factor analysis, kernel variants ...

PPCA – Generative View

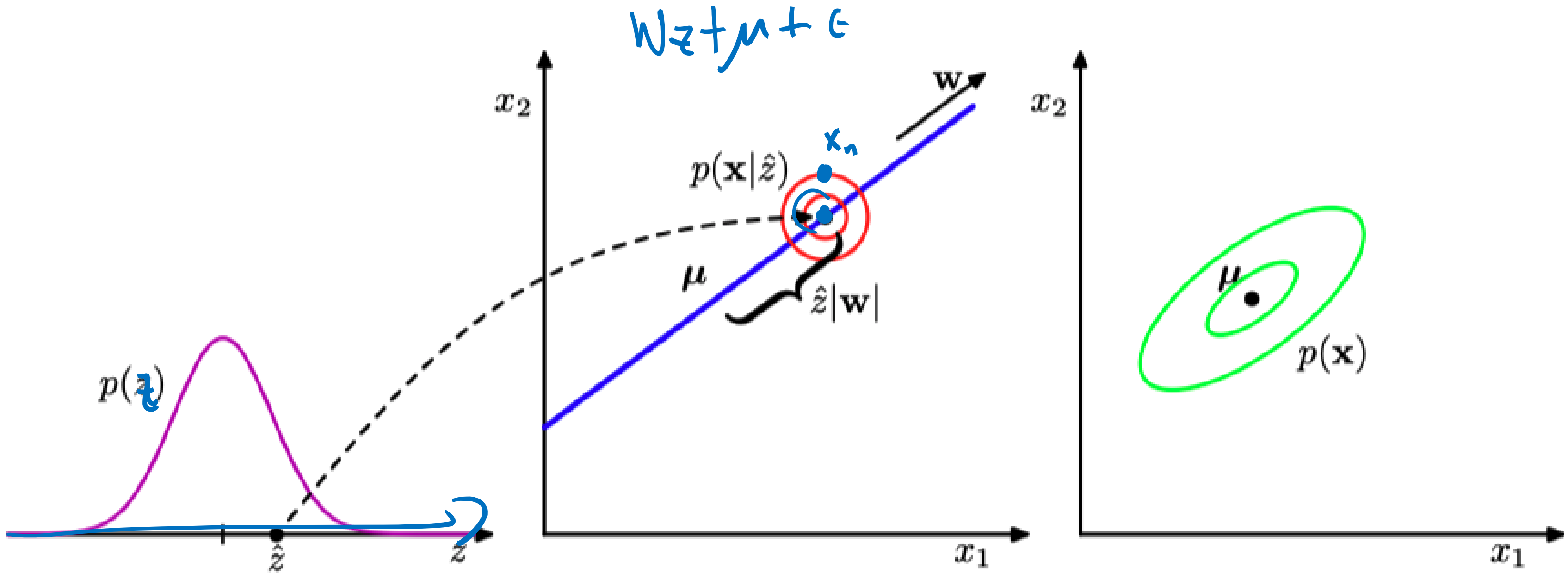


Figure 12.9 An illustration of the generative view of the probabilistic PCA model for a two-dimensional data space and a one-dimensional latent space. An observed data point $\mathbf{x}$ is generated by first drawing a value $\hat{z}$ for the latent variable from its prior distribution $p(z)$ and then drawing a value for $\mathbf{x}$ from an isotropic Gaussian distribution (illustrated by the red circles) having mean $\mathbf{w}\hat{z} + \boldsymbol{\mu}$ and covariance $\sigma^2\mathbf{I}$. The green ellipses show the density contours for the marginal distribution $p(\mathbf{x})$.

Image Credit: PRML Chp 12

Probabilistic PCA

For the probabilistic PCA model, we have D-dimensional data points $\mathbf{x}_i$ for $i = 1, 2, \ldots, N$ and we aim to find some reduced structure for the data. For each data point, we associate a M-dimensional latent variable (where often $M < D$) $\mathbf{z}_i$ that has prior distribution,

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

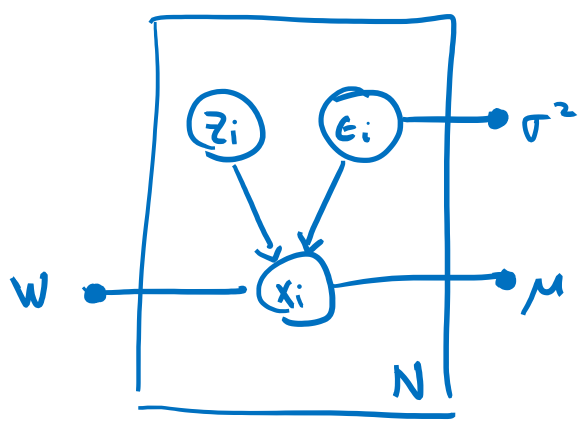
We define each observed variable $\mathbf{x}$ as,

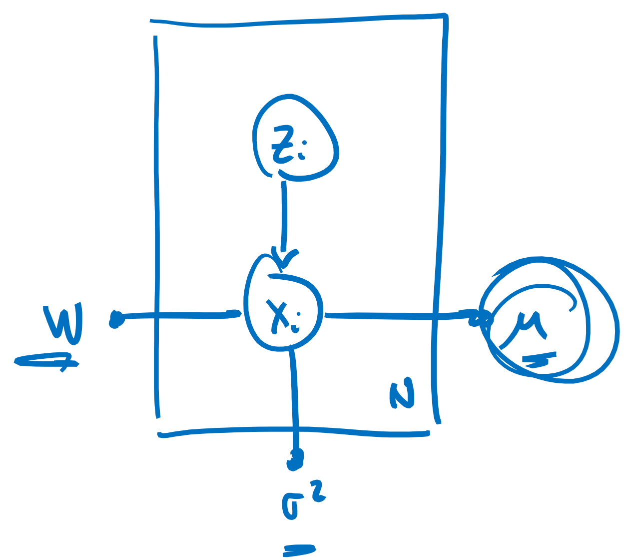
$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We can imagine that each data point is obtained by first sampling from the prior $p(\mathbf{z}_i)$ followed by an affine transformation and additive Gaussian noise.

Problem 1.a. Draw the DGM corresponding to the model above. *Hint:* use plate notation for the different data points.

Solution:

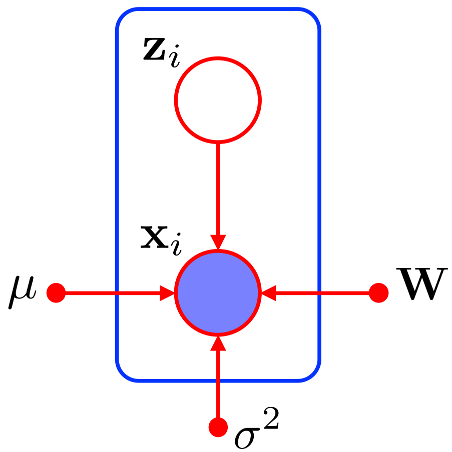




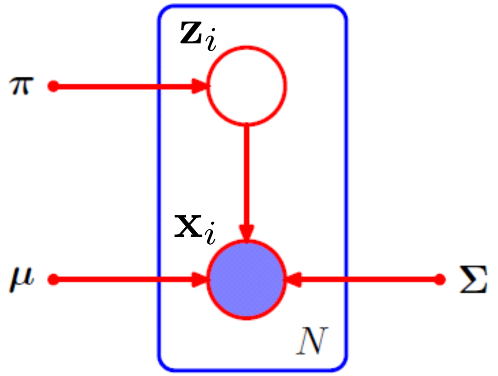
DGM for PPCA

Relationship to GMMs?

PPCA



GMM



Problem 1.b. Show that the conditional distribution for each observed variable $\mathbf{x}_i$ is given by:

$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\underline{\boldsymbol{\mu} + \mathbf{W}}\mathbf{z}_i, \sigma^2\mathbf{I})$$

Solution:

$p(x_i | z_i, \theta)$

$\underline{\epsilon_i} \sim N(\underline{0}, \underline{\sigma^2 I})$

$x_i = \underline{W z_i + \mu + \epsilon_i}$

$p(x_i | z_i) = \mathcal{N}(x_i | W z_i + \mu, \sigma^2 \underline{I})$

$$\begin{aligned} \mathbb{E}[x_i | z_i] &= \mathbb{E}[W z_i + \mu + \epsilon_i | z_i] \\ &= W z_i + \mu + \underbrace{\mathbb{E}[\epsilon_i | z_i]}_{=0} \\ &= W z_i + \mu \end{aligned}$$

$$\begin{aligned} \operatorname{Cov}(x_i | z_i) &= \operatorname{Cov}(\epsilon_i | z_i) \\ &= \sigma^2 I \end{aligned}$$

The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(\mathrm{Z}|\mathrm{X}, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

Problem 1.c. To find the MLE values for the model parameters $\mathbf{W}, \boldsymbol{\mu}$, and σ^2, we would need the marginal distribution $p(\mathbf{X}) = \prod_i^N p(\mathbf{x}_i)$ (assuming i.i.d. data). Due to the latent variables, we will use the EM algorithm[2]. This requires us to marginalize out the latent $\mathbf{z}$'s. To help us along,

1. First, show that the marginal distribution of each data point is again a Gaussian given by

$$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$

   where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

2. Then, show that the posterior distribution is also normally distributed,

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$$

   where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Hint: Given random variables $\mathbf{x}$ and variable $\mathbf{y}$ where:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x\right) \tag{4}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x}\right) \tag{5}$$

The marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T\right) \tag{6}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right) \tag{7}$$

where

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

1. First, show that the marginal distribution of each data point is again a Gaussian given by

$$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.

$p(x_i|z_i) = \mathcal{N}(x_i | W z_i + \mu, \sigma^2 I)$

If:

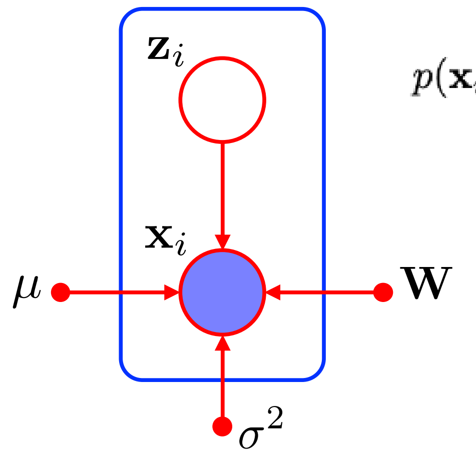
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x})$$

then:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right)$$
$$\boxed{\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}}$$

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

Solution:



$$p(\mathbf{x}_i) = \mathcal{N}(\mathbf{0} + \boldsymbol{\mu} + \mathbf{0}, \mathbf{W}\mathbf{I}\mathbf{W}^\top + \sigma^2\mathbf{I})$$
$$= \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})$$

$$\left.\begin{array}{l} x \Rightarrow z_i \\ \mu \Rightarrow 0 \\ \Sigma_x \Rightarrow I \\ y \Rightarrow x_i \\ A \Rightarrow W \\ b \Rightarrow \mu \\ \Sigma_{y|x} \Rightarrow \sigma^2 I \end{array}\right\}$$

$$p(x_i) = \mathcal{N}(x_i | W \cdot 0 + \mu, \sigma^2 I + W I W^T)$$
$$= \mathcal{N}(x_i | \mu, \underbrace{\sigma^2 I + W W^T}_{C})$$

2. Then, show that the posterior distribution is also normally distributed,

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$$

where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Solution:

$$\Sigma_{x|y} = \left(I + \frac{W^T I W}{\sigma^2} \right)^{-1}$$

$$= \left(I + \frac{1}{\sigma^2} W^T W \right)^{-1}$$

$$= \left(\frac{\sigma^2}{\sigma^2} I + \frac{1}{\sigma^2} W^T W \right)^{-1}$$

$$= \left(\frac{1}{\sigma^2} (\sigma^2 I + W^T W) \right)^{-1}$$

$$= \sigma^2 \left(\sigma^2 I + W^T W \right)^{-1} = \sigma^2 M^{-1}$$

$x \Rightarrow z_i$

$\mu \Rightarrow 0$

$\Sigma_x \Rightarrow I$

$y \Rightarrow x_i$

$A \Rightarrow W$

$b \Rightarrow \mu$

$\Sigma_{y|x} \Rightarrow \sigma^2 I$

If:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x\right)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x}\right)$$

then:

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T\right)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \overbrace{\boldsymbol{\Sigma}_x^{-1}\underbrace{\boldsymbol{\mu}}_{=0}}^{=0}\right), \boldsymbol{\Sigma}_{x|y}\right)$$
$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

$$p(z_i|x_i) = \mathcal{N}\left(z_i \,\middle|\, \cancel{\sigma^2} M^{-1} \left(\frac{W^T I (x_i - \mu)}{\cancel{\sigma^2}} \right), \sigma^2 M^{-1} \right)$$

$$= \mathcal{N}\left(z_i \,\middle|\, M^{-1} W^T (x_i - \mu), \sigma^2 (M^{-1} \right)$$

$$\Sigma_{y|x} = \sigma^2 I$$

$$\Sigma_{y|x}^{-1} = \frac{I}{\sigma^2}$$

$$(cA)^{-1} = \frac{A^{-1}}{c}$$

2. Then, show that the posterior distribution is also normally distributed,

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\underline{\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu})}, \sigma^2\underline{\mathbf{M}^{-1}})$$

where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

If:

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x\right)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x}\right)$$

then:

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T\right)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

Solution:

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}\left(\mathbf{z}_i|\boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i}\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{x}_i|\mathbf{z}_i}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \boldsymbol{\Sigma}_{\mathbf{z}_i}^{-1}\mathbf{0}\right), \boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i}\right)$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{x}_i|\mathbf{z}_i} &= \sigma^2\mathbf{I} \\
\boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i} &= \left(\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1} + \mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i}^{-1}\mathbf{W}\right)^{-1} \\
&= \left(\mathbf{I}^{-1} + \mathbf{W}^T(\sigma^2\mathbf{I})^{-1}\mathbf{W}\right)^{-1} \\
&= \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1} = \sigma^2\mathbf{M}^{-1}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{z}_i|\mathbf{x}_i) &= \mathcal{N}\left(\mathbf{z}_i|\sigma^2\mathbf{M}^{-1}\left(\mathbf{W}^T(\sigma^2\mathbf{I})^{-1}\right)(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right) \\
&= \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})
\end{aligned}$$

Problem 1.d. Finally, derive the E-step and the M-step for the EM algorithm applied to probabilistic PCA. *Hint:* If you are really stuck, refer to Chapter 12.2.2. of Bishop's Pattern Recognition and Machine Learning. This portion is not especially difficult but is notationally heavy and requires algebraic manipulation.

Objective : Find μ_{ML}.

$$\frac{d}{dx}(x^T A x) = 2Ax$$

$$A = A^T$$

$$\underset{\mu}{\text{maximize}} \log p(D|\mu) = \underset{\mu}{\text{maximize}} \log \prod_{i=1}^{N} p(x_i|\mu)$$

$$\Rightarrow \sum_{i=1}^{N} \log p(x_i|\mu) = \sum_{i=1}^{N} \log \mathcal{N}(x_i \,|\, \underline{\mu, WW^T+\sigma^2 I})$$

$$= \sum_{i=1}^{N} \text{const} + \cancel{\log}\exp\Big(-\frac{1}{2}(x_i-\mu)^T \overbrace{(WW^T+\sigma^2 I)^{-1}}^{M^{-1}}(x_i-\mu)\Big)$$

$$\frac{d}{d\mu} \Rightarrow \sum_{i=1}^{N} \frac{1}{2}\left(2M^{-1}(x_i-\mu)\right) \overset{\text{set}}{=} 0 \quad \Rightarrow \quad M^{-1}\sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{N}(x_i-\mu) = 0 \quad \Rightarrow \quad \sum_{i=1}^{N} x_i = N\mu \quad \Rightarrow \quad \mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

The General EM Algorithm

$E[z_i z_i^T] = Cov(z_i)$
$- E[z_i]E[z_i]^T$

$Var(X)$
$= E[X^2] - E[X]^2$
$XX^T \quad E[X]E[X]^T$

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(\mathrm{Z}|\mathrm{X}, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$E[z_i]$
$E[z_i z_i^T]$

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

$E_z[\ln p(z_i) + \ln p(x_i|z_i)]$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

$$" = E_{Z \sim p(Z|X, \theta^{old})}[\ln p(X, Z|\theta)] "$$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

Problem 1.d. Finally, derive the E-step and the M-step for the EM algorithm applied to probabilistic PCA. *Hint:* If you are really stuck, refer to Chapter 12.2.2. of Bishop's Pattern Recognition and Machine Learning. This portion is not especially difficult but is notationally heavy and requires algebraic manipulation.

The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(\mathrm{Z}|\mathrm{X}, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

1. Parameters: $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}$
2. Expectation: Evaluate $p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$
3. Maximization:
   - Obtain the Q function
   - We need:

$$\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = \sum_{n=1}^{N} \{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\}$$

   - Which will lead to the expectation over the posterior
   - That we then maximize in the usual way.

M-Step

- First compute:

$$\ln p\left(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = \sum_{n=1}^{N} \{\ln p(\mathbf{x}_n | \mathbf{z}_n) + \ln p(\mathbf{z}_n)\}$$

M-Step

$\mathrm{Tr}\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\right) = 1 + 5 + 9$

$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$

$\mathrm{Tr}([10]) = 10$

$A = A^T$

$(a+b)^T A (a+b) = a^T A a + b^T A b + 2a^T A b.$

- First compute:

$\mathbb{E}[z]$

$\mathbb{E}[zz^T]$

$$\mathbb{E}_{z \sim p(z|x,\theta^{old})}\left[\ln p\left(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2\right)\right] = \sum_{n=1}^{N}\{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\}$$

$$= \prod_{n=1}^{N} p(x_i, z_i|\theta)$$

$$\ln\ p(\mathbf{x}_n|\mathbf{z}_n,\theta) = -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_n)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{x}_n - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_n)$$

$$= -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{x}_n - \boldsymbol{\mu})^T(\mathbf{x}_n - \boldsymbol{\mu}) - \underbrace{\frac{1}{2\sigma^2}(\mathbf{W}\mathbf{z}_n)^T(\mathbf{W}\mathbf{z}_n)}_{(*)} + \frac{1}{\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu})$$

$$(*)\quad = -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\mathbf{x}_n - \boldsymbol{\mu}||^2 - \frac{1}{2\sigma^2}\mathrm{trace}(\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}) + \frac{1}{\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu})$$

$$\ln\ p(\mathbf{z}_n|\theta) = -\frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n$$

$$(*) = \mathrm{Tr}\left(\underbrace{z_n^T W^T W}_{A}\underbrace{z_n}_{B}\right) = \mathrm{Tr}\left(\underbrace{z_n}_{A}\underbrace{z_n^T W^T W}_{B}\right)$$

$$\mathbb{E}[\ln p\left(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2\right)] = -\sum_{n=1}^{N}\left\{\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right)\right.$$
$$+\frac{1}{2\sigma^2}\|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n - \boldsymbol{\mu})$$
$$\left.+\frac{1}{2\sigma^2}\mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)\right\}. \tag{12.53}$$

M-Step

Matrix Cookbook

$$
\begin{aligned}
\mathbb{E}[\mathbf{z}_n] &= \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n - \overline{\mathbf{x}}) \\
\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}] &= \sigma^2\mathbf{M}^{-1} + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}
\end{aligned}
$$

- Take derivative

$Q(\theta, \theta^{old})$

$p(z|x, \theta^{old})$

$$
\begin{aligned}
\mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}|\underbrace{\boldsymbol{\mu}, \mathbf{W}, \sigma^2}_{\theta})] = -\sum_{n=1}^{N}\Big\{ &\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right) \\
&+ \frac{1}{2\sigma^2}\|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}(\mathbf{x}_n - \boldsymbol{\mu}) \\
&+ \frac{1}{2\sigma^2}\mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)\Big\}. \qquad (12.53)
\end{aligned}
$$

$$
\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T
$$

$$
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{B}\mathbf{X}^T\mathbf{X}) = \mathbf{X}\mathbf{B}^T + \mathbf{X}\mathbf{B}
$$

$\frac{dQ}{dW} = 0$

$$
\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right]^{-1}
$$

$\frac{dQ}{d\sigma^2} = 0$

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

Questions?

$$\mathbb{E}(\mathrm{Tr}(zz^T X)) = \mathrm{Tr}[\,\mathbb{E}(zz^T]\, X\,]$$

https://pollev.com/elim360

SCAN ME