UNSUPERVISED LEARNING 2011

LECTURE :FACTOR ANALYSIS

Rita Osadchy

Based on Lecture Notes by A. Ng

Motivation

- Distribution comes from MoG
  - Have sufficient amount of data: m>>n

    dimension

    num. of training points

  - Use EM to fit Mixture of Gaussians

- If m<<n
  - difficult to model a single Gaussian
  - much less a mixture of Gaussian

Motivation

- m data points span only a low-dimensional subspace of $\Re^n$
- ML estimator of Gaussian parameters:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)(x_i - \mu)^T$$

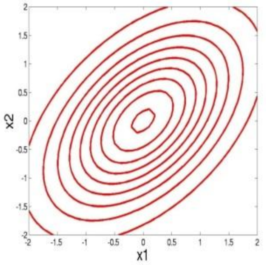
Singular → Can't compute Gaussian Density

- More generally, unless m exceeds n by some reasonable amount, the maximum likelihood estimates of the mean and covariance may be quite poor.

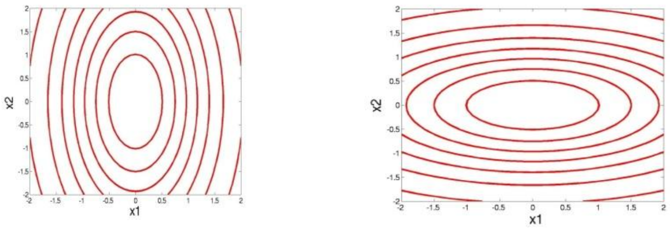
Restriction on Σ

- Goal: Fit a reasonable Gaussian model to the data when m<<n.
- Possible solutions:
  - Limit the number of parameters, assume Σ is diagonal.
  - Limit $\Sigma = \sigma^2 I$, where σ^2 is the parameter under our control.

Contours of a Gaussian Density

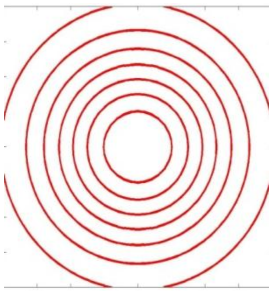


General ∑



Diagonal ∑

Contours are axis aligned



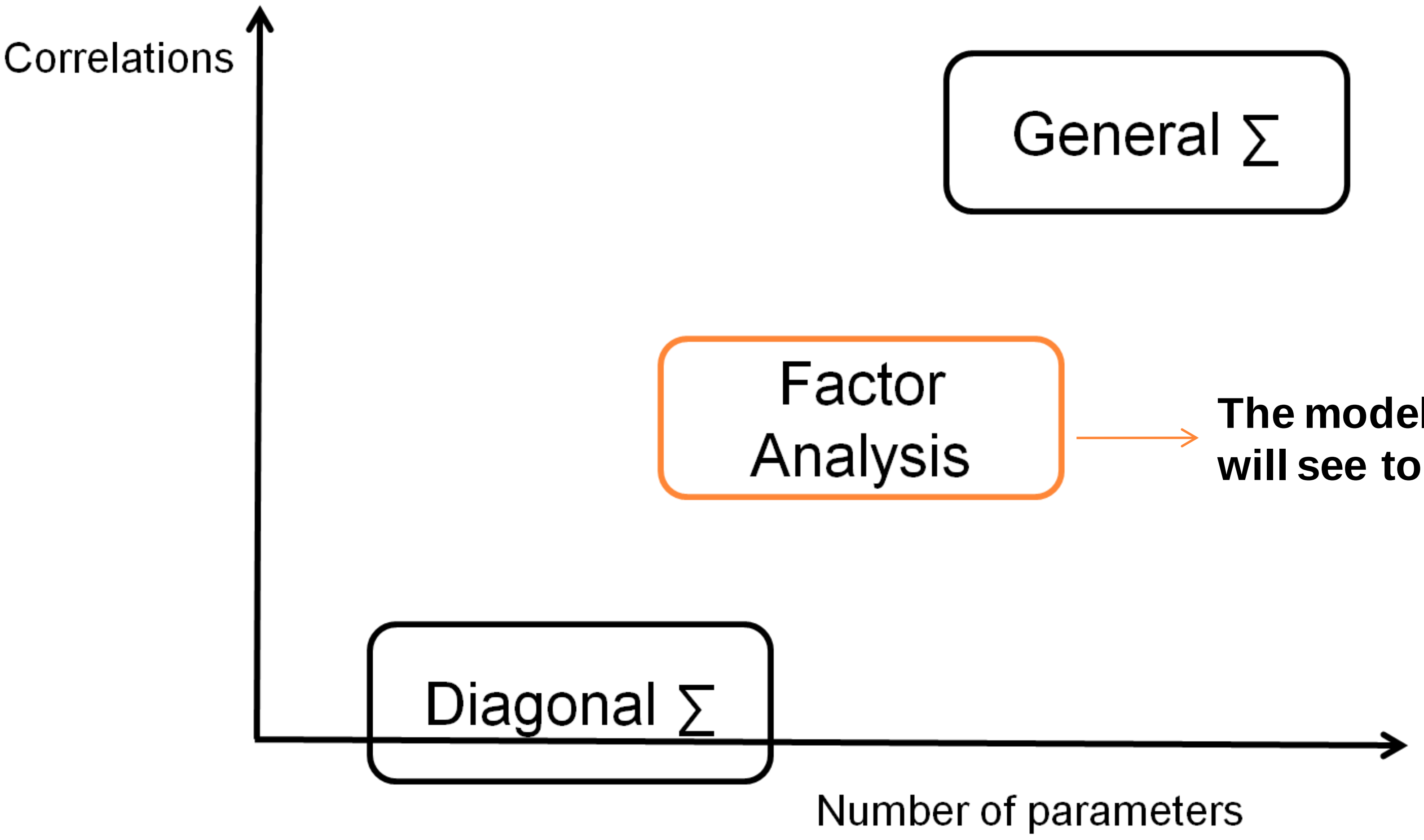
$$\Sigma = \sigma^2 I,$$

Correlation in the data

- Restricting $\sum$ to be diagonal means modelling the different coordinates of the data as being uncorrelated and independent.

- Often, we would like to capture some interesting correlation structure in the data.

Modeling Correlation



Factor Analysis Model

Assume a latent random variable $z \in \Re^k \ (k < n), \quad z \sim N(0, I)$

The parameters of the model

$$x \mid z \sim N(\mu + \Lambda z, \Psi)$$

$\mu \in \Re^n$

$\Lambda \in \Re^{n \times k}$

$\Psi \in \Re^{n \times n}$ is diagonal

Equivalently,

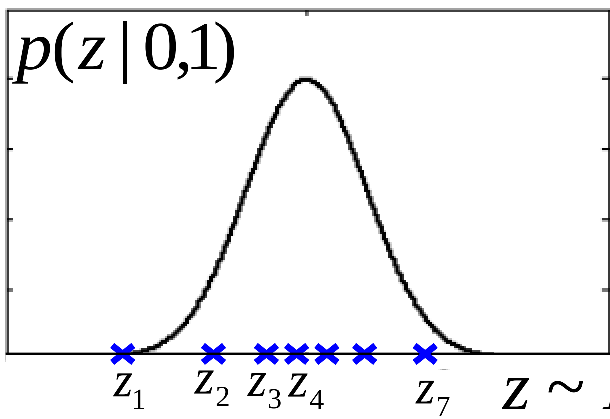
$$x = \mu + \Lambda z + \varepsilon$$

$\varepsilon \sim N(0, \Psi)$

z and ε are independent.

Example of the generative model of x

$z \in \Re^1,\ x \in \Re^2$

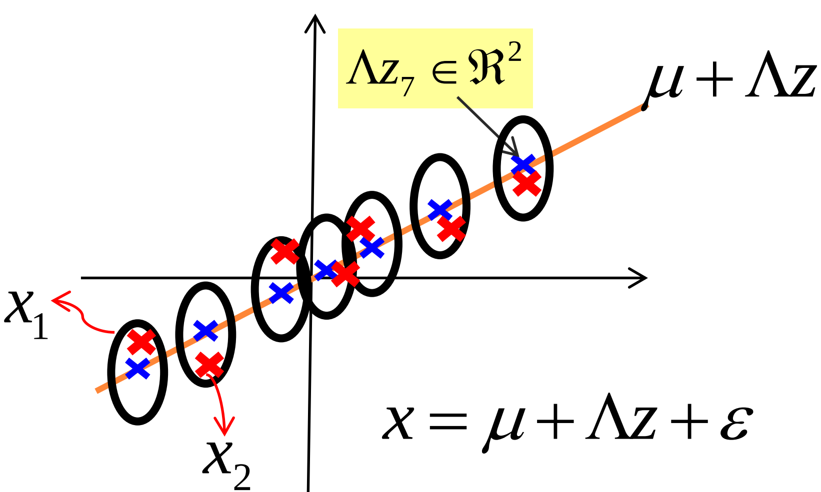


$z \sim N(0,1)$

$$\Lambda = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Psi = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



$$x = \mu + \Lambda z + \varepsilon \qquad \varepsilon \sim N(0, \Psi)$$

Generative process in higher dimensions

- We assume that each data point is generated by sampling a k-dimension multivariate Gaussian z_i .
- Then, it is mapped to a k-dimensional affine space of $\Re^n$ by computing $\mu + \Lambda z_i$
- Lastly, x_i is generated by adding covariance Ψ noise to $\mu + \Lambda z_i$.

Definitions

Partitioned vector

- Suppose $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is r.v., where $x_1 \in \Re^r$, $x_2 \in \Re^s$, $x \in \Re^{r+s}$
- Suppose $x \sim N(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Here $\mu_1 \in \Re^r, \mu_2 \in \Re^s, \Sigma_{11} \in \Re^{r \times r}, \Sigma_{12} \in \Re^{r \times s}, \ldots$ and $\Sigma_{12} = \Sigma_{21}^T$

- Under our assumptions, x_1 and x_2 are jointly multivariate Gaussian.

Marginal distribution of x_1

$$p(x_1) = \int_{x_2} p(x_1, x_2) dx_2$$

Marginal distributions of Gaussians are themselves Gaussian, hence $x_1 \sim N(\mu_1, \Sigma_{11})$

By definition of the joint covariance of x_1 and x_2

$$Cov(x) = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = E\left[(x-\mu)(x-\mu)^T\right] = E\left[\begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T\right]$$

$$= E\begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1)^T & (x_1-\mu_1)(x_2-\mu_2)^T \\ (x_2-\mu_2)(x_1-\mu_1)^T & (x_2-\mu_2)(x_2-\mu_2)^T \end{bmatrix}.$$

$$Cov(x_1) = E[(x_1-\mu_1)(x_1-\mu_1)^T] = \Sigma_{11}$$

Conditional distribution of x_1 given x_2

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \quad \begin{matrix} \leftarrow N(\mu, \Sigma) \\ \leftarrow N(\mu_2, \Sigma_{22}) \end{matrix}$$

Referring to the definition of the multivariate Gaussian distribution, it can be shown that $x_1 \mid x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$, where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Finding the Parameters of FA model

- Assume z and x have a joint Gaussian distribution:

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim N(\mu_{zx}, \Sigma)$$

- We want to find μ_{zx} and Σ

$$E[z] = 0 \quad \text{(since } z \sim N(0, I))$$

$$E[x] = E[\mu + \Lambda z + \varepsilon] = \mu + \Lambda E[z] + E[\varepsilon] = \mu.$$

$$\mu_{zx} = E\begin{bmatrix} z \\ x \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix} \begin{matrix} \updownarrow k \\ \updownarrow n \end{matrix}$$

Finding Σ

- We need to calculate
  - upper left block

$$\Sigma_{zz} = E[(z - E[z])(z - E[z])^T]$$

$$\Sigma_{zz} = Cov(z) = I$$

$z \sim N(0, I)$

  - upper-right block

$$\Sigma_{zx} = E[(z - E[z])(x - E[x])^T]$$

  - lower-right block

$$\Sigma_{xx} = E[(x - E[x])(x - E[x])^T]$$

Finding Σ_{zx}

$$E[(z - E[z])(x - E[x])^T] = E[z(\mu + \Lambda z + \varepsilon - \mu)^T]$$

=0

$$= E[zz^T]\Lambda + E[z\varepsilon^T]$$

$E[zz^T] = Cov(z)$

$E[z\varepsilon^T] =$ ← independent

$$E[z]E[\varepsilon] = 0$$

$$= \Lambda^T$$

Finding Σ_{xx}

Similarly,

$$\Sigma_{xx} = E[(x - E[x])(x - E[x])^T]$$

$$= E[(\mu + \Lambda z + \varepsilon - \mu)(\mu + \Lambda z + \varepsilon - \mu)^T]$$

$$= E[\Lambda z z^T \Lambda^T + \varepsilon z^T \Lambda^T - \Lambda z \varepsilon^T + \varepsilon \varepsilon^T]$$

$$= \Lambda E[z z^T]\Lambda^T + E[\varepsilon \varepsilon^T] = \Lambda \Lambda^T + \Psi$$

Finding the parameters (cont.)

Putting everything together, we have that,

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix} \right)$$

We also see that the marginal distribution of x is given by

$$x \sim N(\mu, \Lambda\Lambda^T + \Psi)$$

Thus, given a training set $\{x_i\}_{i=1}^m$ log likelihood of the parameters is:

$$l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^m \frac{1}{(2\pi)^{n/2} \left| \Lambda\Lambda^T + \Psi \right|} \exp\left(-\frac{1}{2} (x_i - \mu)^T \left(\Lambda\Lambda^T + \Psi \right) (x_i - \mu) \right)$$

Finding the parameters (cont.)

$$l(\mu, \Lambda, \Psi) = \log \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2} \left|\Lambda\Lambda^T + \Psi\right|} \exp\left(-\frac{1}{2} (x_i - \mu)^T \left(\Lambda\Lambda^T + \Psi\right)(x_i - \mu) \right)$$

- To perform maximum likelihood estimation, we would like to maximize this quantity with respect to the parameters.
- But maximizing this formula explicitly is hard, and we are aware of no algorithm that does so in closed-form.
- So, we will instead use the EM algorithm.

EM for Factor Analysis

- E-step:

$$Q_i(z_i) = p(z_i \mid x_i, \theta)$$

- M-step:

$$\theta = \arg\max_{\theta} \sum_i \int_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)} dz_i$$

E-step (EM for FA)

- We need to compute $Q_i(z_i) = p(z_i \mid x_i; \mu, \Lambda, \Psi)$
- Using a conditional distribution of a Gaussian we find that $z_i \mid x_i \sim N(\mu_{z_i|x_i}, \Sigma_{z_i|x_i})$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$\Sigma = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}$$

$$\mu_{z_i|x_i} = \underbrace{\vec{0}}_{\mu_1} - \underbrace{\Lambda^T}_{\Sigma_{12}} \underbrace{(\Lambda^T\Lambda + \Psi)^{-1}}_{\Sigma_{22}^{-1}} \underbrace{(x_i - \mu)}_{(x_2 - \mu_2)}$$

$$\Sigma_{z_i|x_i} = \underbrace{I}_{\Sigma_{11}} - \underbrace{\Lambda^T}_{\Sigma_{12}} \underbrace{(\Lambda^T\Lambda + \Psi)^{-1}}_{\Sigma_{22}^{-1}} \underbrace{\Lambda}_{\Sigma_{12}}$$

$$Q_i(z_i) = \frac{1}{(2\pi)^{2k} \left|\Sigma_{z_i|x_i}\right|^{1/2}} \exp\left(-\frac{1}{2}(z_i - \mu_{z_i|x_i})^T \Sigma_{z_i|x_i}^{-1} (z_i - \mu_{z_i|x_i}) \right)$$

M-step (EM for FA)

Maximize:

$$\sum_{i=1}^{m} \int_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i; \mu, \Lambda, \Psi)}{Q_i(z_i)} dz_i$$

with respect to the parameters μ, Λ, Ψ

- We will work out the optimization with respect to Λ
- Derivations of the updates for μ, Ψ is an exercise (Do it!)

Update for Λ

$$\sum_{i=1}^{m}\int_{z_i} Q_i(z_i)\log\frac{p(x_i, z_i; \mu, \Lambda, \Psi)}{Q_i(z_i)}\,dz_i$$

$$=\sum_{i=1}^{m}\int_{z_i} Q_i(z_i)[\log p(x_i \mid z_i; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i)]dz_i$$

Expectation with respect to z_i , drawn from Q_i

$$=\sum_{i=1}^{m} E_{z_i \sim Q_i}[\log p(x_i \mid z_i; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i)]$$

Update for Λ (cont.)

$$\sum_{i=1}^{m} E_{z_i \sim Q_i}[\log p(x_i \mid z_i; \mu, \Lambda, \Psi) + \log p(z_i) - \log Q_i(z_i)]$$

Remember that We want to maximize this expression with respect to Λ

$$\sum_{i=1}^{m} E_{z_i \sim Q_i}[\log p(x_i \mid z_i; \mu, \Lambda, \Psi)]$$

$$x \mid z \sim N(\mu + \Lambda z, \Psi)$$

$$= \sum_{i=1}^{m} E\left[\log \frac{1}{(2\pi)^{n/2} |\Psi|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu - \Lambda z_i)^T \Psi^{-1} (x_i - \mu - \Lambda z_i)\right)\right]$$

$$= \sum_{i=1}^{m} E\left[-\frac{1}{2}\log|\Psi| - \frac{n}{2}\log(2\pi) - \frac{1}{2}(x_i - \mu - \Lambda z_i)^T \Psi^{-1} (x_i - \mu - \Lambda z_i)\right]$$

Do not depend on Λ

Update for Λ (cont.)

Take derivative with respect to Λ

$$\nabla_\Lambda \sum_{i=1}^{m} -E\left[\frac{1}{2}(x_i-\mu-\Lambda z_i)^T \Psi^{-1}(x_i-\mu-\Lambda z_i)\right]$$

scalar

$\mathrm{tr}\, a = a,\ \ a \in \Re;$

$$= \nabla_\Lambda \sum_{i=1}^{m} -E\left[\mathrm{tr}\frac{1}{2}(x_i-\mu-\Lambda z_i)^T \Psi^{-1}(x_i-\mu-\Lambda z_i)\right]$$

Simplify:

$$= \sum_{i=1}^{m} \nabla_\Lambda E\left[-\mathrm{tr}\frac{1}{2} z_i^T \Lambda^T \Psi^{-1} \Lambda z_i + \mathrm{tr}\, z_i^T \Lambda^T \Psi^{-1}(x_i-\mu)\right]$$

Update for Λ (cont.)

$$\sum_{i=1}^{m} \nabla_{\Lambda} E\left[-\operatorname{tr} \frac{1}{2} z_i^{T} \Lambda^{T} \Psi^{-1} \Lambda z_i + \operatorname{tr} z_i^{T} \Lambda^{T} \Psi^{-1} (x_i - \mu) \right]$$

$$\operatorname{tr} AB = \operatorname{tr} BA$$

$$= \sum_{i=1}^{m} \nabla_{\Lambda} E\left[-\operatorname{tr} \frac{1}{2} \Lambda^{T} \Psi^{-1} \Lambda z_i z_i^{T} + \operatorname{tr} \Lambda^{T} \Psi^{-1} (x_i - \mu) z_i^{T} \right]$$

$$\nabla_{A^T} \operatorname{tr} ABA^{T} C = B^{T} A^{T} C^{T} + B^{T} A^{T} C$$

$$= \sum_{i=1}^{m} E\left[-\Psi^{-1} \Lambda z_i z_i^{T} + \Psi^{-1} (x_i - \mu) z_i^{T} \right]$$

Update for Λ (cont.)

$$\sum_{i=1}^{m} E\left[-\Psi^{-1}\Lambda z_i z_i^T + \Psi^{-1}(x_i-\mu) z_i^T\right]$$

Setting this to zero and simplifying, we get:

$$\sum_{i=1}^{m} \Lambda E_{z_i \sim Q_i}\left[z_i z_i^T\right] = \sum_{i=1}^{m} (x_i-\mu) E_{z_i \sim Q_i}\left[z_i^T\right]$$

Solving for Λ, we obtain:

$$\Lambda = \left(\sum_{i=1}^{m} (x_i-\mu) E_{z_i \sim Q_i}\left[z_i^T\right]\right)\left(\sum_{i=1}^{m} E_{z_i \sim Q_i}\left[z_i z_i^T\right]\right)^{-1}$$

Since Q is Gaussian with mean $\mu_{z_i|x_i}$ and covariance $\Sigma_{z_i|x_i}$

$$E_{z_i \sim Q_i}[z_i^T] = \mu_{z_i|x_i}^T$$

$$E_{z_i \sim Q_i}[z_i z_i^T] = \mu_{z_i|x_i}\mu_{z_i|x_i}^T + \Sigma_{z_i|x_i}$$

$Cov(Y) = E[YY^T] - E[Y]E[Y^T]$

hence,

$E[YY^T] = E[Y]E[Y^T] + Cov(Y)$

Update for Λ (cont.)

$$E_{z_i \sim Q_i}[z_i z_i^T] = \mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i}$$

$$E_{z_i \sim Q_i}[z_i^T] = \mu_{z_i|x_i}^T$$

substitute

$$\Lambda = \left(\sum_{i=1}^{m} (x_i - \mu) E_{z_i \sim Q_i}\left[z_i^T\right] \right) \left(\sum_{i=1}^{m} E_{z_i \sim Q_i}\left[z_i z_i^T\right] \right)^{-1}$$

$$\Lambda = \left(\sum_{i=1}^{m} (x_i - \mu) \mu_{z_i|x_i}^T \right) \left(\sum_{i=1}^{m} \mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i} \right)^{-1}$$

M-step updates for μ and Ψ

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Doesn't depend on $Q_i(z_i) = p(z_i \mid x_i; \mu, \Lambda, \Psi)$, hence can be computed once for all the iterations .

$$\Phi = \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T - x_i \mu_{z_i|x_i}^T \Lambda^T - \Lambda \mu_{z_i|x_i} x_i^T + \Lambda(\mu_{z_i|x_i} \mu_{z_i|x_i}^T + \Sigma_{z_i|x_i})\Lambda^T$$

The diagonal $\Psi_{ii} = \Phi_{ii}$

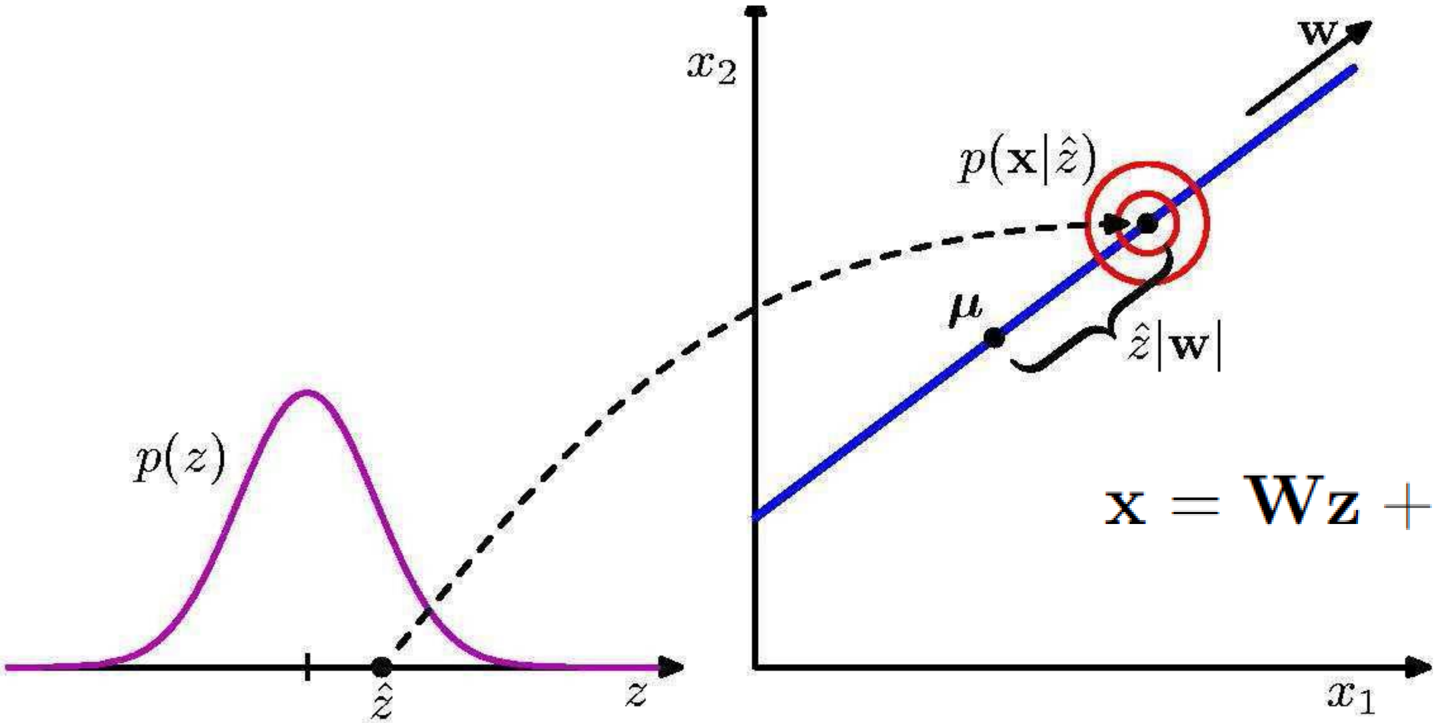
(contains only diagonal entrees)

Probabilistic PCA

- Probabilistic, generative view of data

$$x \in \Re^D, z \in \Re^M$$

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \\ p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I}) \end{aligned}$$

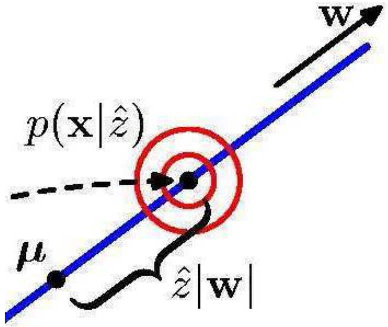


$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \epsilon$$

Compare

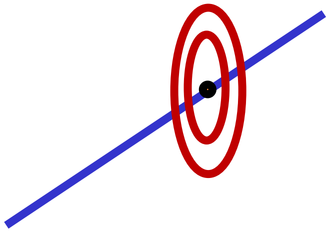
- Probabilistic PCA

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



spherical

- FA



$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \mathbf{\Psi})$$

diagonal, axis-aligned

Probabilistic PCA

- The columns of W are the principle components.
- Can be found using
  - ML in closed form
  - EM (more efficient when only few eigenvectors are required, avoids evaluation of data covariance matrix)
  - Other advantages (see Bishop, Ch.12.2)