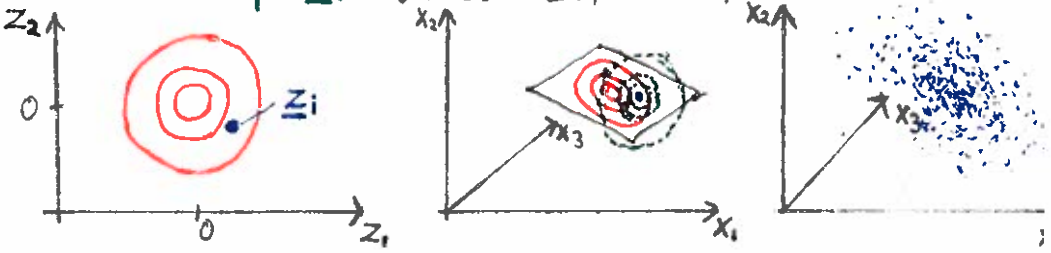
Probabilistic PCA

- Latent $\underline{z}$ with $p(\underline{z}) = \mathcal{N}(\underline{z} \mid \underline{0}, I_k)$
- Observable $\underline{x}$ with $\underline{x} = W\underline{z} + \underline{\mu} + \underline{\varepsilon}$

where $p(\underline{\varepsilon}) = \mathcal{N}(\underline{\varepsilon} \mid \underline{0}, \sigma^2 I_d)$



Joint distribution:

$$p(\underline{z}, \underline{x}) = \frac{1}{\text{const}} \exp\left(-\frac{1}{2}\left[(\underline{x} - W\underline{z} - \underline{\mu})^T \left(\frac{1}{\sigma^2} I_d\right) \cdot (\underline{x} - W\underline{z} - \underline{\mu}) + \underline{z}^T \underline{z}\right]\right)$$

Completing the square:

For a given $-\frac{1}{2}\underline{x}^T A \underline{x} + \underline{x}^T \underline{\eta} + \text{const}$:

1) Isolate 2nd order term $-\frac{1}{2}\underline{x}^T A \underline{x}$ and obtain $\boxed{\Sigma = A^{-1}}$

2) Isolate linear term $\underline{x}^T \underline{\eta}$ and use Σ to obtain $\boxed{\underline{\mu} = \Sigma \underline{\eta}}$

Applied to the PPCA model:

- Take $\begin{pmatrix} z \\ x \end{pmatrix}$ as a single random vector and isolate 2^{nd} order terms of $p(z, x)$:

$$-\frac{1}{2}\left[z^T\left(I_k + W^T\frac{1}{\sigma^2}W\right)z + x^T\frac{1}{\sigma^2}x - x^T\frac{1}{\sigma^2}Wz - z^TW^T\frac{1}{\sigma^2}x\right]$$

$$= -\frac{1}{2}\begin{pmatrix} z \\ x \end{pmatrix}^T\begin{pmatrix} I_k + W^T\frac{1}{\sigma^2}W & -W^T\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2}W & \frac{1}{\sigma^2}I_d \end{pmatrix}\begin{pmatrix} z \\ x \end{pmatrix}$$

$$\Rightarrow \mathrm{Cov}\left[\begin{pmatrix} z \\ x \end{pmatrix}\right] = \left(\quad " \quad\right)^{-1} \overset{\text{lin. alg.}}{=} \begin{pmatrix} I_k & W^T \\ W & WW^T + \sigma^2 I_d \end{pmatrix}$$

- Linear terms:

$$-z^TW^T\frac{1}{\sigma^2}\mu + x^T\frac{1}{\sigma^2}\mu = \begin{pmatrix} z \\ x \end{pmatrix}^T\begin{pmatrix} -W^T\frac{1}{\sigma^2}\mu \\ \frac{1}{\sigma^2}\mu \end{pmatrix}$$

$$\rightsquigarrow \mathbb{E}\left[\begin{pmatrix} z \\ x \end{pmatrix}\right] = \begin{pmatrix} I_k & W^T \\ W & WW^T + \sigma^2 I_d \end{pmatrix}\begin{pmatrix} -W^T\frac{1}{\sigma^2}\mu \\ \frac{1}{\sigma^2}\mu \end{pmatrix} = \begin{pmatrix} 0 \\ \mu \end{pmatrix}$$

$$\Rightarrow p(z, x) = \mathcal{N}\left(\begin{pmatrix} z \\ x \end{pmatrix} \middle| \begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} I_k & W^T \\ W & WW^T + \sigma^2 I_d \end{pmatrix}\right)$$

Using the lower (right) partitions, we also obtain the observation distr.:

$$p(x) = \mathcal{N}(x \mid \mu, WW^T + \sigma^2 I_d)$$

2020 L 8 ②

Maximum likelihood

Given centred $d \times n$ observation matrix

$X = (\underline{x}_1 \dots \underline{x}_n)$

Objective: $\log p(X|W,\sigma^2) \overset{ind}{=} \sum_{i=1}^{n} \log p(\underline{x}_i|W,\sigma^2)$

$$= \sum_{i=1}^{n} \log \left[(2\pi)^d \det \Sigma \right]^{-\frac{1}{2}} - \frac{1}{2} \underline{x}_i^T \Sigma^{-1} \underline{x}_i$$

$$= -\frac{1}{2} \left[n d \log 2\pi + n \log(\det \Sigma) + n \frac{1}{n} \sum_i \underline{x}_i^T \Sigma^{-1} \underline{x}_i \right]$$

$$= -\frac{n}{2} \left[d \log(2\pi) + \log(\det \Sigma) + \operatorname{trace}(\Sigma^{-1} \hat{\Sigma}) \right]$$

where $\Sigma = WW^T + \sigma^2 I_d$ and

$\hat{\Sigma} = \operatorname{cov}(X) = \frac{1}{n} X X^T$

$W_{ML} = U_k (\Lambda_k - \sigma^2 I_k)^{1/2} R$ (Tipping and Bishop 1999)

where $\hat{\Sigma} = U \Lambda U$ the EV decomposition

R : $k \times k$ orth matrix "Solution not unique"

e.g. $R = I_k$ yielding $W_{ML} = U_k (\Lambda_k - \sigma^2 I_k)^{1/2}$

$\sigma^2_{ML} = \frac{1}{d-k} \sum_{i=k+1}^{d} \lambda_i$

$\rightsquigarrow \sigma^2$ represents the average lost variance per residual dimension

Relation to PCA:

Given an observation $\underline{x}_i$, what do we know about the latent variable?

A) $\underline{z}_i = U_k^T \underline{x}_i$ ✗

B) We get a single $\underline{z}_i$ from $\underline{x}_i$ via some other relation ✗

C) We get a distribution over $\underline{z}$'s ✓

D) You don't know ✗

2020 4 8

Relation to PCA:

- PCA: map data $\underline{x}_i$ to PC scores $\underline{z}_i$:

  $\underline{z}_i = U_k^T \underline{x}_i$

- Probabilistic PCA: map $\underline{z}_i$ to $\underline{x}_i$:

  $\underline{x}_i = W\underline{z}_i + \underline{\varepsilon}$ stochastic

  → whole posterior distribution $p(\underline{z} | \underline{x}_i)$

  closest thing to PCA mapping

Posterior $p(\underline{z} | \underline{x})$:

Fix $\underline{x}$ in $p(\underline{z}, \underline{x})$ and complete the square

Completing the square for posterior:

- 2nd order terms:

$$\text{Cov}[\underline{z}|\underline{x}] = \left(I_k + W^T \frac{1}{\sigma^2} W\right)^{-1}$$

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

where $M = W^T W + \sigma^2 I_k$

- Linear terms:

$$\underline{\ell} = \frac{1}{2} W^T \frac{1}{\sigma^2} \underline{x} + \frac{1}{2}\left(\underline{x}^T \frac{1}{\sigma^2} W\right)^T - W^T \frac{1}{\sigma^2} \underline{\mu}$$

$$= W^T \frac{1}{\sigma^2} (\underline{x} - \underline{\mu})$$

$$\Rightarrow \mathbb{E}[\underline{z}|\underline{x}] = \sigma^2 M^{-1} W^T \frac{1}{\sigma^2} (\underline{x} - \underline{\mu})$$

$$= M^{-1} W^T (\underline{x} - \underline{\mu})$$

Posterior: $p(\underline{z}|\underline{x}) = \mathcal{N}(\underline{z}|M^{-1}W^T(\underline{x}-\underline{\mu}), \sigma^2 M^{-1})$

Projections:

For PCA: $\hat{\underline{x}} = U_k U_k^T \underline{x}$

For Probabilistic PCA:

$\hat{\underline{x}} = W_{ML} \mathbb{E}[\underline{z} | \underline{x}] = W_{ML} M_{ML}^{-1} W_{ML}^T \underline{x}$

where $M_{ML} = W_{ML}^T W_{ML} + \sigma^2 I$

"Expectation of posterior projected back into data space."

For $\sigma^2 \to 0$:

$W_{ML} M_{ML}^{-1} W_{ML}^T \underline{x}$

$= \ldots = U_k U_k^T \underline{x}$

$\rightsquigarrow$ PCA is a special case of Probabilistic PCA where covariance of noise is infinitesimally small