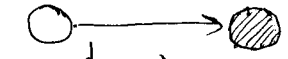
CSE 291 B 1/13

Factor analysis

- Belief network



$\vec{z} \in \mathbb{R}^d$ (hidden) low dimensional ($d \ll D$)

$\vec{x} \in \mathbb{R}^D$ (observed) high dimensional

- Gaussian distributions

Prior $P(\vec{z}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}\|\vec{z}\|^2\right\}$

Conditional $P(\vec{x}|\vec{z}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det\Psi}} \exp\left\{-\frac{1}{2}(\vec{x}-\Lambda\vec{z}-\vec{\mu})^T\Psi^{-1}(\vec{x}-\Lambda\vec{z}-\vec{\mu})\right\}$

Marginal $P(\vec{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\Psi+\Lambda\Lambda^T)}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T(\Psi+\Lambda\Lambda^T)^{-1}(\vec{x}-\vec{\mu})\right\}$

Posterior $P(\vec{z}|\vec{x}) = ??$

- Bayes rule

$$P(\vec{z}|\vec{x}) = \frac{P(\vec{x}|\vec{z})P(\vec{z})}{P(\vec{x})}$$

$$\sim \exp\left\{-\frac{1}{2}(\vec{x}-\Lambda\vec{z}-\vec{\mu})^T\Psi^{-1}(\vec{x}-\Lambda\vec{z}-\vec{\mu}) - \frac{1}{2}\|\vec{z}\|^2\right\}$$

chopping terms that don't have $\vec{z}$

- "Complete the square" inside the exponent

$$P(\vec{z}|\vec{x}) \sim \exp\left\{-\frac{1}{2}(\vec{z}-\vec{b})^T A(\vec{z}-\vec{b}) + \underbrace{\ldots}_{\text{independent of } \vec{z}}\right\}$$

for appropriately defined $\vec{b} \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$

After some algebra:

$$A = I_d + \Lambda^T\Psi^{-1}\Lambda$$

$$b = A^{-1}\Lambda^T\Psi^{-1}(\vec{x}-\vec{\mu})$$

Important statistics:

Let $E[\vec{z}|\vec{x}] = \int \vec{z}\, P(\vec{z}|\vec{x})$ posterior mean

Let $\delta\vec{z}_x \triangleq \vec{z} - E[\vec{z}|\vec{x}]$

Posterior $P(\vec{z}|\vec{x})$ is gaussian b/c joint is gaussian.
Identifying terms in exponent:

$$E[\vec{z}|\vec{x}] = (I+\Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}(\vec{x}-\vec{\mu})$$

$$E[\delta\vec{z}_x\,\delta\vec{z}_x^T|\vec{x}] = (I+\Lambda^T\Psi^{-1}\Lambda)^{-1}$$

- ML estimation for complete data:
  given $\{(\vec{z}_n,\vec{x}_n)\}_{n=1}^N$, how to maximize $\mathcal{L}=\sum_n \log P(\vec{z}_n,\vec{x}_n)$ in terms of $(\vec{\mu},\Lambda,\Psi)$?

Log-likelihood:

$$\mathcal{L} = \sum_n \log P(\vec{z}_n,\vec{x}_n)$$

$$= -\frac{1}{2}\sum_n \{\log\det\Psi + (\vec{x}_n-\Lambda\vec{z}_n-\vec{\mu})^T\Psi^{-1}(\vec{x}_n-\Lambda\vec{z}_n-\vec{\mu}) + \dots\}$$

(... : independant of (μ,Λ,Ψ))

Note: RHS is quadratic in μ,Λ.
Simple to differentiate and optimize.

Define:

$$\Delta\vec{x}_\ell = \vec{x}_\ell - \frac{1}{N}\sum_n \vec{x}_n$$

$$\Delta\vec{z}_\ell = \vec{z}_\ell - \frac{1}{N}\sum_n \vec{z}_n$$

} deviations from sample means

By setting $\frac{\partial\mathcal{L}}{\partial\vec{\mu}}=\vec{0}$ (aka $\boxed{\frac{\partial\mathcal{L}}{\partial\mu_i}=0}$), $\frac{\partial\mathcal{L}}{\partial\Lambda_{i\alpha}}=0$, $\frac{\partial\mathcal{L}}{\partial\Psi_{ii}}=0$:

$$\Lambda = \underbrace{\left[\sum_n (\Delta\vec{x}_n)(\Delta\vec{z}_n)^T\right]}_{D\times d\text{ matrix}}\underbrace{\left[\sum_n (\Delta\vec{z}_n)(\Delta\vec{z}_n)^T\right]^{-1}}_{d\times d\text{ matrix}}$$

$$\vec{\mu} = \frac{1}{N}\sum_n (\vec{x}_n - \Lambda\vec{z}_n)$$

$$\Psi_{ii} = \frac{1}{N}\sum_n (\vec{x}_n-\Lambda\vec{z}_n-\vec{\mu})_i^2$$

- ML estimation for incomplete data:
  - EM algorithm estimates $\Lambda, \vec{\mu}, \Psi$ to maximize marginal log-likelihood $\mathcal{L} = \sum_n \log P(\vec{x}_n)$
  - E-step: compute mean $E[\vec{z}|\vec{x}_n]$ and covariance $E[\delta z \delta z^T | \vec{x}_n]$ of posterior distribution $P(\vec{z}|\vec{x}_n)$
  - M-step (stated w/o proof):

$$\Lambda \leftarrow \underbrace{\left[\sum_n (\Delta\vec{x}_n)(\Delta\vec{z}_n)^T\right]}_{D \times d \text{ matrix}} \overbrace{\left[\sum_n (\Delta\vec{z}_n)(\Delta\vec{z}_n)^T + \sum_n E[\delta z \delta z^T | \vec{x}_n]\right]^{-1}}^{d \times d \text{ matrix}}$$

$$\Delta\vec{z}_n = E[\vec{z}|\vec{x}_n] - \frac{1}{N}\sum_n E[\vec{z}|\vec{x}_n]$$

$$\vec{\mu} \leftarrow \frac{1}{N}\sum_n (\vec{x}_n - \Lambda E[\vec{z}|\vec{x}_n])$$

$$\Psi_{ii} \leftarrow \frac{1}{N}\sum_n \left[(\vec{x}_n - \Lambda E[\vec{z}|\vec{x}_n] - \vec{\mu})_i^2 + (\Lambda E[\delta z \delta z^T | \vec{x}_n] \Lambda^T)_{ii}\right]$$

New terms arise from uncertainty in $\vec{z}$.

Since $P(\vec{z}|\vec{x})$ is gaussian, all the uncertainty is captured by cov matrix $E[\delta z \delta z^T | \vec{x}_n]$.

(If cov entries are very small, ~~extra terms~~ update begins to look like ML estimation w/ complete data)

- Computational complexity:
  $O(NDd)$ per iteration

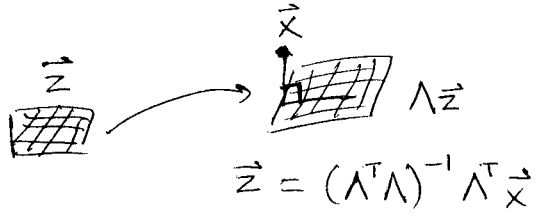
- Relation to PCA?

Suppose (i) sample has zero mean $\sum \vec{x}_n = 0$. Take $\vec{\mu} = 0$.

(ii) high dimensional noise is isotropic: $\Psi = \sigma^2 I$ ← global scalar variance

Then in low noise limit:

$$\lim_{\sigma^2 \to 0} P(\vec{z}|\vec{x}) = \delta(\vec{z} - (\Lambda^T\Lambda)^{-1}\Lambda^T\vec{x})$$



$\vec{z}$ is completely determined by linear projection onto subspace.

- EM algorithm in $\lim_{\sigma^2 \to 0}$:

E-step: $\vec{z}_n = (\Lambda^T\Lambda)^{-1}\Lambda^T\vec{x}_n$

M-step: $\Lambda \leftarrow \left(\sum_n \vec{x}_n \vec{z}_n^T\right)\left(\sum_n \vec{z}_n \vec{z}_n^T\right)^{-1}$ iteratively compute best map from $\vec{z}$ to $\vec{x}$

Complexity: $O(NDd)$ per iteration (* How many iterations?)

EM in this limit converges to computing <u>max variance subspace</u>

Compare to eigenvector computations for PCA:

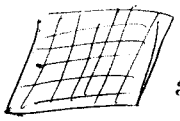
$O(ND^2)$ to compute cov matrix

$O(dD^2)$ to compute subspace

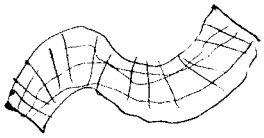
EM-PCA often more efficient in practice. (*)

Unifying clustering and dimensionality reduction

- High dimensional data may lie on (or near) a low dimensional sub-manifold



subspace



submanifold

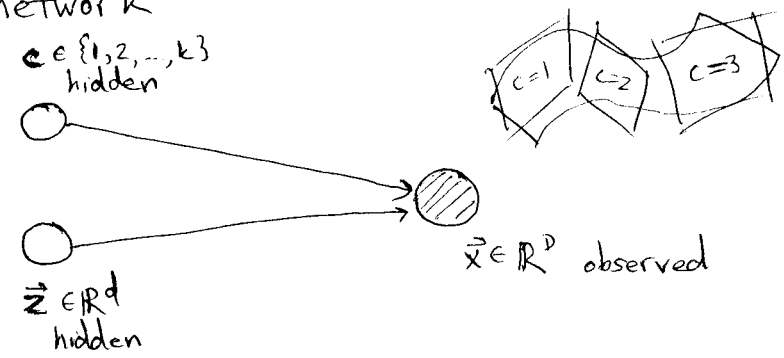
Ex: D = 100 x 100 = 10,000 dimensional (# pixels/image)
images of faces

$d \ll D$ degrees of variability (# muscles)

- Submanifold can be approximated by a collection of locally linear patches
- Model each patch by factor analysis
  Model collection by mixture model

Mixture of factor analyzers

- Belief network



- Model parameters $\{(\pi_c, \vec{\mu}_c, \Lambda_c, \Psi_c)\}_{c=1}^{k}$

$P(c) = \pi_c$ mixture weight; prior prob. for being on cth patch

$P(\vec{z}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}\|z\|^2\right\}$

$P(\vec{x}|\vec{z}, c) \sim \exp\left\{-\frac{1}{2}(\vec{x}-\Lambda_c\vec{z}-\vec{\mu}_c)^T \Psi^{-1}(\vec{x}-\Lambda_c\vec{z}-\vec{\mu}_c)\right\}$ patch-specific factor analysis

(Lack of $\vec{z}$ depending on c may mean that high-dim points $\vec{x}$ at the boundary of diff. patches that map to diff $\vec{z}_i, \vec{z}_j$ depend on patch-i, patch-j cannot be modeled.)