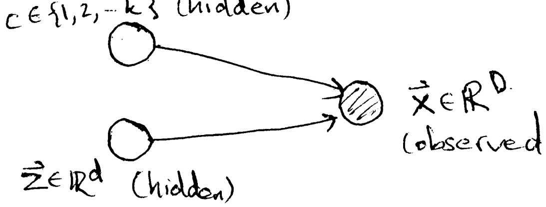
CSE 291B 1/15

Mixture of factor analyzers

- Graphical model



- Probabilistic model

Prior $P(c) = \pi_c$

Prior $P(\vec{z}) \sim \exp\left\{-\frac{1}{2}\|z\|^2\right\}$

Conditional $P(\vec{x}|c,\vec{z}) \sim \exp\left\{-\frac{1}{2}(\vec{x}-\Lambda_c\vec{z}-\vec{\mu}_c)\Psi_c^{-1}(\vec{x}-\Lambda_c\vec{z}-\vec{\mu}_c)\right\}$

Joint $P(\vec{x}, c, \vec{z}) = P(c)\, P(\vec{z})\, P(\vec{x}|c,\vec{z})$

Marginal $P(\vec{x}) = \sum_c \int d^d z\, P(\vec{x}, c, \vec{z})$

- EM algorithm

Same structure as FA updates,
but with examples weighted by $P(c|\vec{x}_n)$

Useful shorthand:

$$N_c = \sum_{n=1}^{N} P(c|\vec{x}_n)$$

$$\left.\begin{aligned}\Delta\vec{x}_n^c &= \vec{x}_n - \frac{1}{N_c}\sum_{\ell=1}^{N} P(c|\vec{x}_\ell)\,\vec{x}_\ell \\ \Delta\vec{z}_n^c &= E[\vec{z}\,|\,c,\vec{x}_n] - \frac{1}{N_c}\sum_{\ell=1}^{N} P(c|\vec{x}_\ell)\,E[\vec{z}\,|\,c,\vec{x}_\ell]\end{aligned}\right\} \text{"deviation from patch center"}$$

$$E[\delta\vec{z}\,\delta\vec{z}^T|c,\vec{x}_n] = E[\vec{z}\vec{z}^T|c,\vec{x}_n] - E[\vec{z}|c,\vec{x}_n]\,E[\vec{z}|c,\vec{x}_n]^T$$

E-step:

Compute $P(c|\vec{x}_n)$, $E[\vec{z}\,|\,\vec{x}_n, c]$, $E[\delta\vec{z}\,\delta\vec{z}^T|\vec{x}_n, c]$ (posterior probabilities) using Bayes rule.

M-step:

$$\pi_c \leftarrow N_c/N$$

$$\Lambda_c \leftarrow \left[\sum_n \underbrace{P(c|\vec{x}_n)}_{\uparrow\ \text{new}} (\Delta\vec{x}_n^{\,c})(\Delta\vec{z}_n^{\,c})^T\right]\left[\sum_n \underbrace{P(c|\vec{x}_n)}_{\uparrow\ \text{new}} \left\{E[\delta\vec{z}\,\delta\vec{z}^T|\vec{x}_n, c] + \Delta\vec{z}_n^{\,c}\,\Delta\vec{z}_n^{\,c\,T}\right\}\right]$$

$$\vec{\mu}_c \leftarrow \frac{1}{N_c}\sum_n P(c|\vec{x}_n)\left[\vec{x}_n - \Lambda_c E[\vec{z}\,|\,\vec{x}_n, c]\right]$$

$$[\Psi_c]_{ii} \leftarrow \frac{1}{N_c}\sum_n P(c|\vec{x}_n)\left[(\Delta\vec{x}_n^{\,c} - \Lambda_c \Delta\vec{z}_n^{\,c})_i^2 + (\Lambda_c E[\delta\vec{z}\,\delta\vec{z}^T|\vec{x}_n, c]\Lambda_c^T)_{ii}\right]$$

Converges to local maximum $\mathcal{L} = \sum_{n=1}^{N} \log P(\vec{x}_n)$ w.r.t. $\{(\pi_c, \vec{\mu}_c, \Lambda_c, \Psi_c)\}_{c=1}^{k}$

Matrix factorization

- How to approximate a large matrix by the (matrix) product of two smaller ones?

$$\underbrace{D\left[\begin{array}{cccc} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_N \end{array}\right]}_{\longleftarrow N \longrightarrow} \approx \underbrace{D\left[\begin{array}{ccc} \vec{v}_1 & \cdots & \vec{v}_d \end{array}\right]}_{\leftarrow d \rightarrow} \underbrace{\left[\begin{array}{cccc} \vec{y}_1 & \vec{y}_2 & \cdots & \vec{y}_N \end{array}\right] d}_{\longleftarrow N \longrightarrow}$$

$d \ll D$

$d \ll N$

$$X \approx V Y$$

X: $D \times N$ data matrix

$\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ with $\vec{x}_i \in \mathbb{R}^D$

$N = $ # examples

$D = $ # dimensions

V: $D \times d$ basis vectors

$\{\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_d\}$

with $\vec{v}_i \in \mathbb{R}^D$

Y: $d \times N$ encoding coefficients

$\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_N\}$

$\vec{y}_n \in \mathbb{R}^d$

- Vector quantization (redux)

To compute prototype $\vec{v}_1, \ldots, \vec{v}_k$:

Minimize quantization error

$$\mathcal{E}(V, Y) = \sum_{n=1}^{N} \sum_{i=1}^{k} \| \vec{x}_n - \vec{v}_i \|^2 Y_{in} \quad \text{subject to} \begin{cases} Y_{in} \in \{0, 1\} \\ \sum_{i=1}^{k} Y_{in} = 1 \end{cases}$$

To compute $X \approx VY$

minimize $\|X - VY\|^2 = \sum_{n=1}^{N} \sum_{i=1}^{k} [X_{in} - (VY)_{in}]^2$ subject to constraints

Completely equivalent.

Ex: images of faces

prototypes are typical whole faces in different parts of face space

• PCA (redux)

To compute basis vectors $\vec{v}_1, \ldots, \vec{v}_d$,

minimize reconstruction error:

$$\sum_{n=1}^{N} \| \vec{x}_n - \sum_{i=1}^{d} \vec{v}_i (\vec{x}_n \cdot \vec{v}_i) \|^2 \text{ subject to } \vec{v}_i \cdot \vec{v}_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise} \end{cases}$$

To compute $X \approx VY$,

minimize $\|X - VY\|^2 = \sum_{in} (X_{in} - (VY)_{in})^2$ subject to $V^T V = I_d$

Ex: images of faces

~~basis~~ vectors ("eigenfaces") are "facial gradients" of greatest variance around mean face.

(eigenvectors give students "sad" faces)

Non-negative matrix factorization

- How to approximate $X \approx VY$ where X, V, and Y only contain non-negative elements?

- Least squares approx error:

$$\mathcal{E}_{LS} = \|X - VY\|^2 = \sum_{in} \left[X_{in} - (VY)_{in}\right]^2$$

Trivially, $\mathcal{E}_{LS} \geq 0$. And $\mathcal{E}_{LS} = 0 \iff X = VY$

- Non-negative divergence

$$\mathcal{E}_{KL} = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in}\right]$$

Reduces to Kullback-Leibler (KL) divergence when $\sum_{in} X_{in} = \sum_{in} (VY)_{in} = 1$.

- Properties of $f(a,b) = a \log \frac{a}{b} - a + b$

(i) $f(a,b) \geq 0$

(ii) $f(a,b) = 0 \iff a = b$

(iii) $f(a,b) \neq f(b,a)$ (that is, not necessary that $f(a,b) = f(b,a)$)

(iv) $\lim_{a \to 0} f(a,b) = b$

(v) $\lim_{b \to 0} f(a,b) = \begin{cases} \infty & \text{if } a \neq 0 \\ 0 & \text{if } a = 0 \end{cases}$

- Properties of $\mathcal{E}_{KL}$

like $\mathcal{E}_{LS}$ → From (i)-(ii): $\mathcal{E}_{KL} = 0 \iff X = VY$

unlike $\mathcal{E}_{KL}$:
- From (iii): approx penalties not symmetric
- From (iv): approx penalty is linear for zero matrix elements of X
- From (v): approx penalty diverges when (VY) does not "explain" all non-zero matrix elements of X

- NMF yields parts-based representations

  Ex: images of faces

  NMF discovers basis vectors that resemble localized facial features (e.g. eyes, nose, mouth, etc...)

  $\Longrightarrow$ "Mr. Potato Head" model

  (non-negatve constraint means reconstruction is additive-only)

- Minimization of $\mathcal{E}_{LS}$ and $\mathcal{E}_{KL}$
  - Neither is possible in closed form due to non-negativity constraints.
  - Look for iterative solutions.

- Decomposition of $\mathcal{E}_{LS}$:

$$\mathcal{E}_{LS} = \sum_{in} \left[X_{in} - (VY)_{in} \right]^2$$

Define

$$\mathcal{E}^{+}_{LS} = \sum_{in} \left[X_{in}^2 + (VY)_{in}^2 \right] \geq 0$$

$$\mathcal{E}^{-}_{LS} = 2 \sum_{in} X_{in} (VY)_{in} \geq 0$$

Clearly: $\mathcal{E}_{LS} = \mathcal{E}^{+}_{LS} - \mathcal{E}^{-}_{LS}$

- Non-negative gradients:

$$\frac{\partial \mathcal{E}^{+}_{LS}}{\partial V_{i\alpha}} = 2 \sum_{n} (VY)_{in} Y_{\alpha n} = 2 (VYY^T)_{i\alpha} \geq 0 \text{ for all } i, \alpha$$

$$\frac{\partial \mathcal{E}^{-}_{LS}}{\partial V_{i\alpha}} = 2 \sum_{n} X_{in} Y_{\alpha n} = 2 (XY^T)_{i\alpha} \geq 0 \quad \text{for all } i, \alpha$$

similar calculations for derivatives w.r.t. $Y_{\alpha n}$

- Multiplicative update

Consider:

$$V_{i\alpha} \longleftarrow V_{i\alpha} \left[\frac{\left(\frac{\partial \mathcal{E}^{-}_{LS}}{\partial V_{i\alpha}} \right)}{\left(\frac{\partial \mathcal{E}^{+}_{LS}}{\partial V_{i\alpha}} \right)} \right]$$

(Why?)