Mixture of factor analyzers

- Graphical model
  \[ c \in \{1, 2, -k \} \text{ (hidden)} \]
  \[ \bar{x}_e \in \mathbb{R}^d \text{ (observed)} \]
  \[ \bar{z} \in \mathbb{R}^d \text{ (hidden)} \]

- Probabilistic model
  \[ \text{Prior } P(c) = \pi_c \]
  \[ \text{Prior } P(\bar{z}) \sim \exp \left\{ -\frac{1}{2} ||z||^2 \right\} \]
  \[ \text{Conditional } P(\bar{x} \mid c, \bar{z}) \sim \exp \left\{ -\frac{1}{2} (\bar{x} - \Lambda_c \bar{z} - \mu_c)^T \Psi_c^{-1} (\bar{x} - \Lambda_c \bar{z} - \mu_c) \right\} \]
  \[ \text{Joint } P(\bar{x}, c, \bar{z}) = P(c) P(\bar{z}) P(\bar{x} \mid c, \bar{z}) \]
  \[ \text{Marginal } P(\bar{x}) = \sum_c \int_d^d P(\bar{x}, c, \bar{z}) \]

- EM algorithm
  
  Same structure as FA updates, but with examples weighted by \( P(c \mid \bar{x}_n) \)

  Useful shorthand:
  \[ N_c = \sum_{n=1}^N P(c \mid \bar{x}_n) \]
  \[ \Delta \bar{x}_n^c = \bar{x}_n - \frac{1}{N_c} \sum_{\ell=1}^N P(c \mid \bar{x}_\ell) \bar{x}_\ell \]
  \[ \Delta \bar{z}_n^c = E[\bar{z} \mid c, \bar{x}_n] - \frac{1}{N_c} \sum_{\ell=1}^N P(c \mid \bar{x}_\ell) E[\bar{z} \mid c, \bar{x}_\ell] \]
  \[ E[\delta \bar{z} \delta \bar{z}^T \mid c, \bar{x}_n] = E[\bar{z} \bar{z}^T \mid c, \bar{x}_n] - E[\bar{z} \mid c, \bar{x}_n] E[\bar{z} \mid c, \bar{x}_n]^T \]
E-step:
Compute \( P(c|x_n), E[z|x_n, c], E[\delta z \delta z^T|x_n, c] \) (posterior probabilities) using Bayes rule.

M-step:
\[
\Pi_c \leftarrow \frac{N_c}{N}
\]
\[
\Lambda_c \leftarrow \left[ \sum_n \sum_{i \text{ new}}^T \frac{P(c|x_n)}{P(c|x_n)} \left( \Delta x_n^c \right) \left( \Delta z_n^c \right)^T \right] \left[ \sum_n \sum_{i \text{ new}}^T \frac{E[\delta z \delta z^T|x_n, c]}{E[\delta z \delta z^T|x_n, c]+\Delta z_n^c \Delta z_n^c^T} \right]
\]
\[
\mu_c \leftarrow \frac{1}{N_c} \sum_n P(c|x_n) \left[ x_n^c - \Lambda_c \ E[z|x_n, c] \right]
\]
\[
[\Psi_c]_{i;i} \leftarrow \frac{1}{N_c} \sum_n P(c|x_n) \left[ (\Delta x_n^c - \Lambda_c \Delta z_n^c)^2_{i;i} + (\Lambda_c \ E[\delta z \delta z^T|x_n, c] \Lambda_c^T)_{i;i} \right]
\]
Converges to local maximum \( L = \sum_{n=1}^N \log P(x_n) \) w.r.t. \( \{\Pi_c, \Lambda_c, \mu_c, \Psi_c\}_{c=1}^k \).
Matrix factorization

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

\[
\begin{bmatrix}
\tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_N
\end{bmatrix} \approx \begin{bmatrix}
\tilde{y}_1 & \tilde{y}_2 & \cdots & \tilde{y}_N
\end{bmatrix}^T \begin{bmatrix}
\tilde{v}_1^T & \tilde{v}_2^T & \cdots & \tilde{v}_d^T
\end{bmatrix}
\]

\( d \ll N \)

\( d \ll D \)

\[ X \approx VY \]

- Vector quantization (redux)

To compute prototype \( \tilde{v}_1, \ldots, \tilde{v}_k \):

Minimize quantization error

\[ E(V,Y) = \sum_{n=1}^{N} \sum_{i=1}^{k} ||\tilde{x}_n - \tilde{v}_i||^2 Y_{in} \] subject to \( \{ Y_{in} \in \{0,1\} \} \)

\[ \sum_{i=1}^{k} \sum_{n=1}^{N} Y_{in} = 1 \]

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
To compute basis vectors $\hat{v}_1, \ldots, \hat{v}_d$,
minimize reconstruction error:
$$\sum_{n=1}^{N} \| x_n - \sum_{i=1}^{d} \hat{v}_i (x_n \cdot \hat{v}_i) \|^2$$
subject to $\hat{v}_i \cdot \hat{v}_j = \{ 1 \text{ if } i=j \}$
0 otherwise.

To compute $X \approx VY$,
minimize $\| X - VY \|^2 = \sum_{in} (X_{in} - (VY)_{in})^2$
subject to $V^TV = I_d$

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:
  $\varepsilon_{LS} = \| X - VY \|_2^2 = \sum_{in} [X_{in} - (VY)_{in}]^2$

  Trivially, $\varepsilon_{LS} \geq 0$. And $\varepsilon_{LS} = 0 \iff X = VY$

- Non-negative divergence
  $\varepsilon_{KL} = \sum_{in} \left[ X_{in} \log \frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in} \right]$

  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 $\varepsilon_{KL}$
  Like $\varepsilon_{LS}$: From (i)-(ii): $\varepsilon_{KL} = 0 \iff X = VY$
  Unlike $\varepsilon_{KL}$: approx penalty is not symmetric
    - From (iii): approx penalty diverges when $(VY)$ does not "explain" all non-zero matrix elements of $X$
• NMF yields parts-based representations
  
  \( \bar{E}_k \): images of faces

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

  \( \Rightarrow \) "Mr. Potato Head" model

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

• Minimization of \( E_{ls} \) and \( E_{kl} \)
  
  - Neither is possible in closed form due to non-negativity constraints.
  - Look for iterative solutions.

• Decomposition of \( E_{ls} \):

  \[
  E_{ls} = \sum_{i,n} \left[ x_{in} - (v y)_{in} \right]^2
  \]

  Define \( E_{ls}^+ = \sum_{i,n} \left[ x_{in}^2 + (v y)_{in}^2 \right] \geq 0 \)

  \[ E_{ls}^- = 2 \sum_{i,n} x_{in} (v y)_{in} \geq 0 \]

  Clearly: \( E_{ls} = E_{ls}^+ - E_{ls}^- \)

• Non-negative gradients:

  \[
  \frac{\partial E_{ls}^+}{\partial v_{i\alpha}} = 2 \sum_n (v y)_{in} y_{\alpha n} = 2 (v y y^T)_{i\alpha} \geq 0 \text{ for all } i, \alpha
  \]

  \[
  \frac{\partial E_{ls}^-}{\partial v_{i\alpha}} = 2 \sum_n x_{in} y_{\alpha n} = 2 (x y^T)_{i\alpha} \geq 0 \text{ for all } i, \alpha
  \]

  Similar calculations for derivatives w.r.t. \( y_{\alpha n} \)

• Multiplicative update

  Consider: \( v_{i\alpha} \leftarrow v_{i\alpha} \left[ \begin{array}{c} \frac{\partial E_{ls}^+}{\partial v_{i\alpha}} \\ -\frac{\partial E_{ls}^-}{\partial v_{i\alpha}} \end{array} \right] \)

  (Why?)