Review of clustering

Inputs \( \{ \vec{x}_1, \ldots, \vec{x}_N \} \), \( \vec{x}_n \in \mathbb{R}^D \). How to assign \( \vec{x} \in \mathbb{R}^D \) to \( \{1, 2, \ldots, k\} \)?

1.1 \textit{k}-means

- Prototypes \( \vec{u}_i \in \mathbb{R}^D \)
- Assignment matrix \( Y_{in} \in \{0, 1\} \)

Alternating minimization:

\[
E(\vec{u}, Y) = \sum_{i=1}^{k} \sum_{n=1}^{N} Y_{in}||\vec{x}_n - \vec{u}_i||^2
\]  

(1)

Iterative update:

\[
Y_{in} = \begin{cases} 
1 & \text{if } i = \text{argmin}_j ||\vec{x}_n - \vec{u}_j||^2 \\
0 & \text{otherwise}
\end{cases},
\]  

(2)

\[
\vec{u}_i = \frac{\sum_n Y_{in} \vec{x}_n}{\sum_n Y_{in}}.
\]  

(3)

1.2 Gaussian Mixture Modeling

\[
P(\vec{x}|z = i) = \frac{1}{(2\pi)^{D/2} \sqrt{\det(\Sigma_i)}} \exp \left( -\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1}(\vec{x} - \vec{\mu}_i) \right). \]  

(4)
Log-likelihood

\[ \mathcal{L} = \sum_n \log \left[ \sum_i P(\vec{x}_n | z = i) P(z = i) \right] \]  

(5)

Relation of EM to k-means: if \( \Sigma = \sigma^2 I \), then

\[
\lim_{\sigma^2 \to 0} P(z = i | \vec{x}_n) = \begin{cases}
1 & \text{if } i = \arg\min_j \\
0 & \text{otherwise}
\end{cases}
\]  

(6)

2 Dimensionality Reduction

Inputs \( \{\vec{x}_1, \ldots, \vec{x}_N\}, \vec{x}_n \in \mathbb{R}^D \). How to map \( \vec{x}_n \in \mathbb{R}^D \to \vec{y}_n \in \mathbb{R}^d \) where \( d \ll D \)? For now, assume \( d \) is given.

2.1 Principle component analysis

Suppose that \( D \)-dimensional inputs lie in (or near) a \( d \)-dimensional subspace. Assume inputs are centered around the origin \( \sum_n \vec{x}_n = \vec{0} \). Note that this assumption is difficult for a large, sparse data set since subtracting the mean to center the data will create many nonzero entries which might be difficult to store.

PCA: orthogonal projections of data that lose the least information.

Compute direction \( \hat{u} \) with \( ||\hat{u}||^2 = \hat{u} \cdot \hat{u} = 1 \) which maximizes the projected variance of inputs. Maximize \( V(\hat{u}) \), the variance when projected to \( \hat{u} \)

\[
V(\hat{u}) = \frac{1}{N} \sum_{n=1}^{N} (\vec{x}_n \cdot \hat{u})^2 - \left[ \frac{1}{N} \sum_n (\vec{x}_n \cdot \hat{u}) \right]^2 ,
\]  

(7)

\[
= \frac{1}{N} \sum_{n=1}^{N} \left( \sum_{\alpha=1}^{D} x_{n\alpha} \cdot \hat{u}_{\alpha} \right) \left( \sum_{\beta=1}^{D} x_{n\beta} \cdot \hat{u}_{\beta} \right) ,
\]  

(8)

\[
= \sum_{\alpha,\beta=1}^{D} \hat{u}_{\alpha} \hat{u}_{\beta} \left[ \frac{1}{N} \sum_{n=1}^{N} x_{n\alpha} x_{n\beta} \right] ,
\]  

(9)

\[
= \hat{u}^T C \hat{u} ,
\]  

(10)

because by definition, \( C_{\alpha\beta} \) are the elements of \( D \times D \) covariance matrix. Note that

\[
\frac{1}{N} \sum_n (\vec{x}_n \cdot \hat{u}) = 0 ,
\]  

(11)
from our assumption that inputs are centered. We maximize the variance with constraints on \( \hat{u} \).

**Constrained maximization.**

\[
\tilde{V}(\hat{u}, \lambda) = \hat{u}^T C \hat{u} + \lambda (1 - \hat{u} \cdot \hat{u}),
\]

\(\Rightarrow\) \[
\frac{\partial \tilde{V}}{\partial \hat{u}} = 2C\hat{u} - 2\lambda \hat{u},
\]

\(\Rightarrow\) \[
\frac{\partial \tilde{V}}{\partial \hat{u}} = 0,
\]

\[\Rightarrow C\hat{u} = \lambda \hat{u}.
\]

This is an eigenvalue equation.

**Which eigenvector solution?** Let \( \hat{u}^{(i)} \) denote eigenvector of \( C \) with \( i \)th largest eigenvalue \( \lambda^{(i)} \) (sorted). Then

\[
V (\hat{u}^{(i)}) = \hat{u}^{(i)T} C \hat{u}^{(i)},
\]

\[= \hat{u}^{(i)T} \lambda \hat{u}^{(i)},
\]

\[= \lambda^{(i)}.
\]

The direction with maximum variance is \( \hat{u}^{(i)} \) with largest eigenvalue. To compute \( d \)-dimensional subspace with maximum variance, we can take the top \( d \) eigenvalues as orthogonal basis (because \( C \) is symmetric real matrix) for subspace.

We can map \( \vec{x}_n \in \mathbb{R}^D \rightarrow \vec{y}_n \in \mathbb{R}^d \) by \( y_i = \vec{x} \cdot \hat{u}^{(i)} \).

We can also look at reconstruction error

\[
\mathcal{E} = \sum_n \left\| \vec{x}_n - \sum_{i=1}^{d} (\vec{x}_n \cdot \hat{u}^{(i)}) \hat{u}^{(i)} \right\|^2.
\]

- **Pros of PCA**
  1. No local minima: eigenvalue computation
  2. No tuning parameters other than choosing \( d \ll D \). How do we choose \( d \)? The sum of the eigenvalues tells you how much of the variance you capture with those eigenvectors.

- **Cons of PCA**
  1. No probabilistic interpretation
     - no explicit model of out-of-space noise \( \rightarrow \) implicit parametric assumption?
     - hard to compose models in principled way

3
2. Scaling with dimension
   - $O(ND^2)$ to compute covariance matrix
   - $O(dD^2)$ to compute top eigenvalues

3. Restricted to discovering linear structure
   - Good for PCA: plane
   - Bad for PCA: manifold

3. Factor analysis

Assume inputs $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ sampled i.i.d. from PDF $P(\vec{x})$. Model $P(\vec{x})$ by Gaussian BN with hidden variable. Estimate model to maximize $\mathcal{L} = \sum_n \log P(\vec{x}_n)$.

$\vec{z} \in \mathbb{R}^d$

(hidden)

$\vec{x} \in \mathbb{R}^D$

(observed)

BN:

Gaussian prior distribution for latent variable $\vec{z}$:

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

with zero mean and identity covariance matrix.
Generative model for $\tilde{X}$: Project $\tilde{z}$ up to higher-dimensional space $\mathbb{R}^D$, add some noise.

- sample $\tilde{z}$ from prior distribution $P(\tilde{z})$
- linearly project $\tilde{z} \in \mathbb{R}^d$ into $D$ dimensions
- add independent zero-mean Gaussian noise with variance $\Psi_{\alpha \alpha}$ to $\alpha$th component of $(\Lambda \tilde{z} + \tilde{u})_\alpha$. Let
  \[ \Psi = \text{diag}(\Psi_{11}, \Psi_{22}, \ldots, \Psi_{DD}), \]  
  denote the diagonal covariance matrix of high dimensional noise.

**Gaussian conditional distribution:**

\[
P(\tilde{x}|\tilde{z}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det(\Psi)}} \exp\left(-\frac{1}{2} [\tilde{x} - \Lambda \tilde{z} - \tilde{u}]^T \Psi^{-1} [\tilde{x} - \Lambda \tilde{z} - \tilde{u}] \right) \]  

(22)

Key intuition: for small $\Psi_{ii}$, most variation of data occurs inside subspace spanned by columns of $\Lambda$.

**Noise model:** for non-uniform $\Psi_{ii}$, vectors with same orthogonal projection onto subspace can have different likelihoods.

**Marginal distribution:**

\[
P(\tilde{x}) = \int_{\tilde{z} \in \mathbb{R}^d} P(\tilde{x}|\tilde{z})P(\tilde{z})d^dz \text{(Gaussian)},
\]

(23)

\[
= \frac{1}{(2\pi)^{D/2}\sqrt{\det(\Psi + \Lambda \Lambda^T)}} \exp\left(-\frac{1}{2}(\tilde{x} - \tilde{u})(\Psi + \Lambda \Lambda^T)^{-1}(\tilde{x} - \tilde{u}) \right)
\]

(24)

with mean $E[\tilde{x}] = \tilde{u}$ and covariance $E[(\tilde{x} - \tilde{u})(\tilde{x} - \tilde{u})^T] = \Psi + \Lambda \Lambda^T$.  

5