Models with Latent Variables. EM algorithm

Evgeny Burnaev

Skoltech, Moscow, Russia

Skolkovo Institute of Science and Technology

Outline

- Measure of divergence between two distributions defined on the same domains

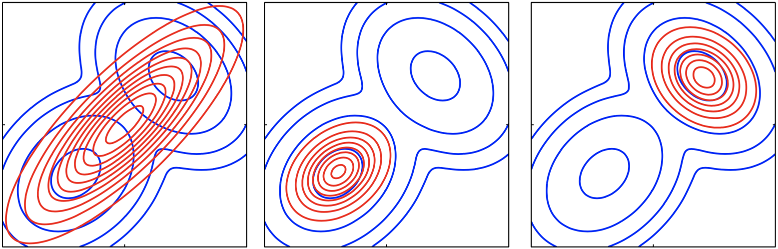
$$
\begin{aligned}
KL(q\|p) &= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \\
&= -\int q(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim q(\cdot)} - \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \\
&\geq -\log \mathbb{E}_{\mathbf{x}\sim q(\cdot)} \frac{p(\mathbf{x})}{q(\mathbf{x})} = -\log \int q(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} = -\log 1 = 0
\end{aligned}
$$

- Information theoretic interpretation

$$
\mathrm{KL} = \mathrm{Cross\,Entropy} - \mathrm{Entropy}
$$

- If we minimize KL w.r.t. $q(\cdot)$ the approximation should be good where $q(\cdot)$ has large values

Kullback-Leibler divergence



(a) (b) (c)

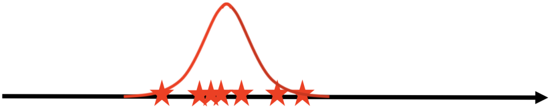
(a) Blue contours: bimodal mixture of two Gaussians distribution $p(\mathbf{z})$. Red contours: single Gaussian distribution $q(\mathbf{z})$ that best approximates $p(\mathbf{z})$ by minimizing $KL(p\|q)$

(b) As in (a) but now $q(\mathbf{z})$ is found by numerical minimization of $KL(q\|p)$

(c) As in (b) but showing a different local minimum of $KL(q\|p)$

- We have a set of points generated from a Gaussian

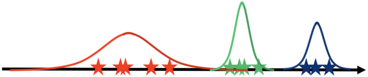
$$x_i \sim \mathcal{N}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$



- We estimate its parameters μ and σ: sample mean and variance

- Several sets of points from different gaussians

- We have to estimate the parameters of those gaussians and their weights



- The problem is as easy if we know what objects were generated from each gaussian
- Using a single gaussian model leads to inaccurate results

- For each x_i we introduce additional z_i denoting the index of Gaussian from which i-th object was generated
- The model is

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) = \\ = \prod_{i=1}^{m} p(x_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}) \\ = \prod_{i=1}^{m} \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \sigma^2_{z_i})$$

- Here $\pi_j = p(z_i = j)$ are prior probability of j-th Gaussian and $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ are the parameters to be estimated
- If we know both $\mathbf{X}$ and $\mathbf{Z}$, we use MLE

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

- We do not know $\mathbf{Z} \Rightarrow$ we maximize w.r.t. $\boldsymbol{\theta}$ the log of incomplete likelihood

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- For any distribution $q(\mathbf{Z})$ we get that

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, we get

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} =$$
$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z}) p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}) p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$
$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$

- We get

$$\log p(\mathbf{X}|\boldsymbol{\theta}) =$$
$$= \underbrace{\int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}}_{\substack{\text{Evidence Lower Bound} \\ \text{ELBO } \mathcal{L}(q, \boldsymbol{\theta})}} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{Non-negative}}$$

- Thus
$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q \| p) \geq \mathcal{L}(q, \boldsymbol{\theta})$$
- Instead of optimizing $\log p(\mathbf{X}|\boldsymbol{\theta})$ we optimize ELBO $\mathcal{L}(q, \boldsymbol{\theta})$ w.r.t. both $\boldsymbol{\theta}$ and $q(\mathbf{Z})$
- The block-coordinate algorithm is known as EM-algorithm

- Function $g(\xi, \mathbf{x})$ is called a variational lower bound for $f(\mathbf{x})$ iff
  - For all ξ and for all $\mathbf{x}$ it follows $f(\mathbf{x}) \geq g(\xi, \mathbf{x})$
  - For any $\mathbf{x}_0$ there exists $\xi(\mathbf{x}_0)$ such that $f(\mathbf{x}_0) = g(\xi(\mathbf{x}_0), \mathbf{x}_0)$
- If we managed to find such variational lower bound, then instead of solving

$$f(\mathbf{x}) \to \max_{\mathbf{x}}$$

we can iteratively perform block-coordinate updates of $g(\xi, \mathbf{x})$

$$\mathbf{x}_i = \arg\max_{\mathbf{x}} g(\xi_{i-1}, \mathbf{x}),\ \xi_i = \xi(\mathbf{x}_i) = \arg\max_{\xi} g(\xi, \mathbf{x}_i)$$

- To solve

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

we start from initial $\boldsymbol{\theta}_0$ and iteratively repeat optimize w.r.t. q and $\boldsymbol{\theta}$

- Let us find $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}_0) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}_0) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \log p(\mathbf{X}|\boldsymbol{\theta}_0) \\
&= -KL(q\|p) + \log p(\mathbf{X}|\boldsymbol{\theta}_0)
\end{aligned}$$

- Thus we get that

$$\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Thus to solve

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

we start from initial $\boldsymbol{\theta}_0$ and iteratively repeat

- **E-step**: find

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

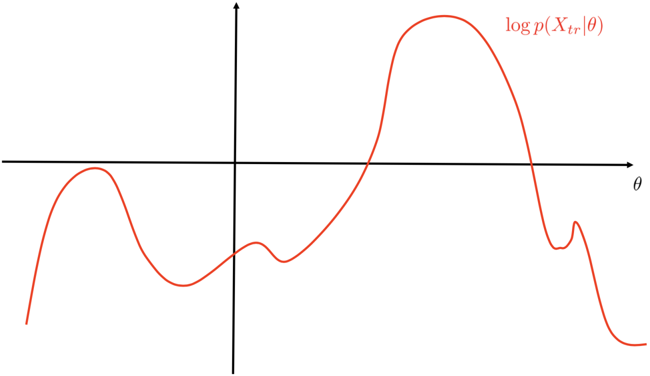
- **M-step**: solve

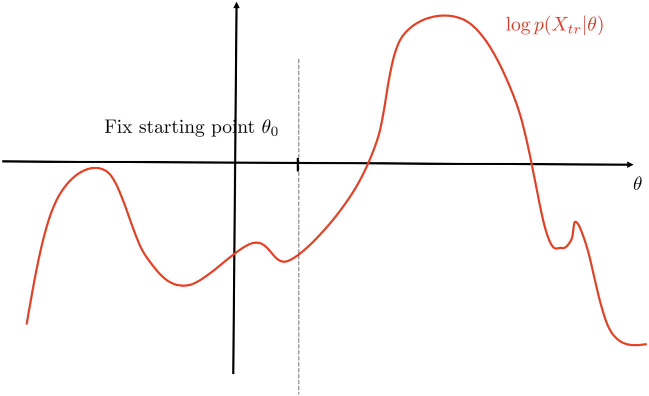
$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}),$$

set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$ and go to **E-step** until convergence

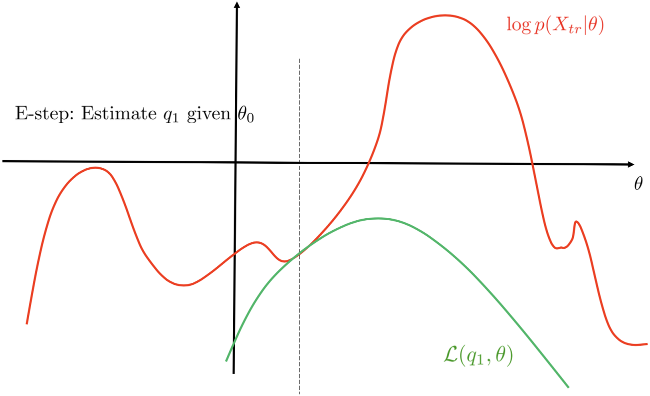
- The EM algorithm monotonically increases the lower bound and converges to stationary point of $\log p(\mathbf{X}|\boldsymbol{\theta})$

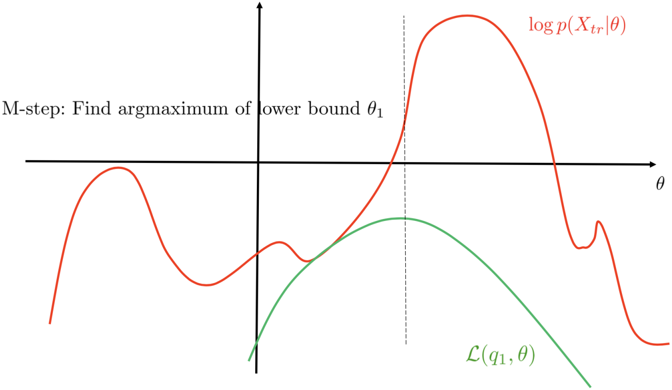
EM-algorithm



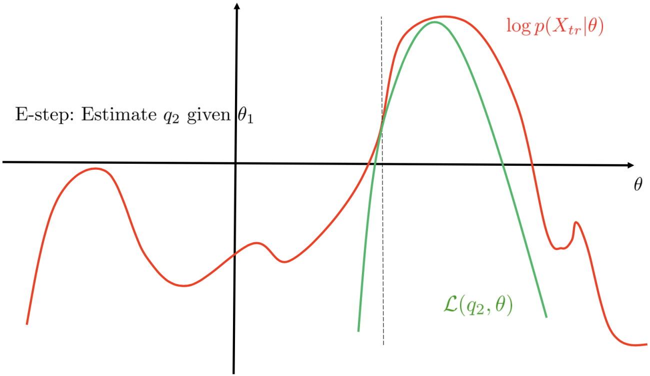

$\log p(X_{tr}|\theta)$
Fix starting point θ_0
θ

EM-algorithm

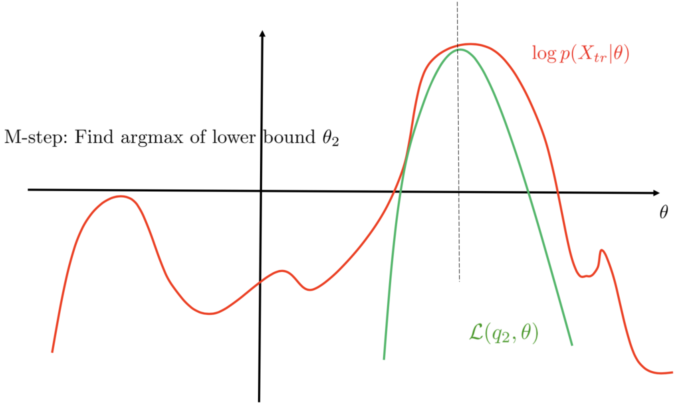




$\log p(X_{tr}|\theta)$
M-step: Find argmaximum of lower bound θ_1
θ
$\mathcal{L}(q_1, \theta)$



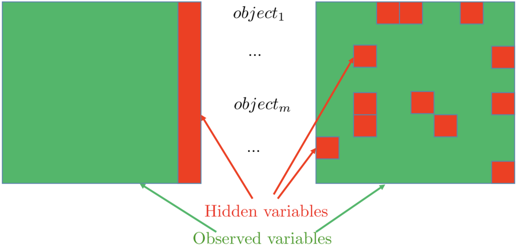
E-step: Estimate q_2 given θ_1
$\log p(X_{tr}|\theta)$
θ
$\mathcal{L}(q_2, \theta)$


M-step: Find argmax of lower bound θ_2
$\log p(X_{tr}|\theta)$
θ
$\mathcal{L}(q_2, \theta)$

Is EM worth to use?

- In many cases (e.g. for the mixture of Gaussians) E-step and M-step can be performed in closed forms
- Allows to build more complicated models of data using mixtures of simple distributions
- If true posterior $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ is intractable we may search for the closest $q(\mathbf{Z})$ among tractable distributions by solving optimization problem
- Allows to process missing data by treating it as latent variables

Is EM worth to use?



- EM algorithm allows to fill in arbitrary gaps in data
- May deal with both discrete and continuous variables
- Always converges
- Allows multiple extensions

- Assume all $z_i \in \{1, \dots, K\}$ then the marginal

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})$$

  is a finite mixture of distributions

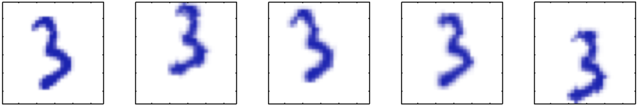
- E-step can be performed in closed form

$$q(z_i = k) = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})}{\sum_{l=1}^{K} p(\mathbf{x}_i|l, \boldsymbol{\theta}) p(z_i = l|\boldsymbol{\theta})}$$

- M-step is simply a sum of finite terms

$$\mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbb{E}_{z_i} \log p(x_i, z_i|\boldsymbol{\theta}) =$$

$$\sum_{i=1}^{m} \sum_{k=1}^{K} q(z_i = k) \log p(x_i, k|\boldsymbol{\theta})$$



- Real datasets: data points lie close to a manifold of much lower dimensionality
- 100×100 grey-scale image, i.e. 10^4 dimensional data space
- three degrees of variability: the vertical/horizontal translations and the rotations, described by some latent variables
- three dimensional nonlinear manifold
- real digit image data: a further degrees of freedom from scaling, variability in an individuals writing, writing styles
- In practice, the data points will not be confined precisely to a smooth low-dimensional manifold: can be interpreted as noise

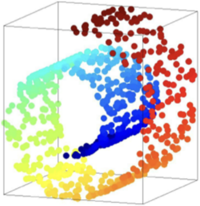
- Continuous variables can be considered as a mixture of a continuum of distributions

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \int p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) d\mathbf{z}_i = \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i|\boldsymbol{\theta}) d\mathbf{z}_i$$

- E-step can be done in closed form only in case of conjugate distributions

$$q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i|\boldsymbol{\theta})}{\int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i|\boldsymbol{\theta}) d\mathbf{z}_i}$$

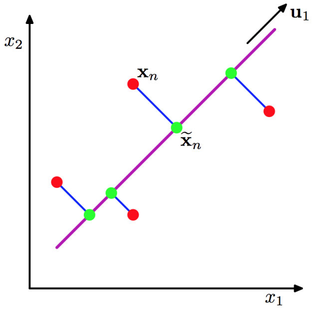
- Typically continuous latent variables are used for dimension reduction also known as representation learning



- Develop probabilistic parametric data model
- Include additional (latent) variables until model becomes simple enough, e.g. belongs to exponential class
- Treat all missing values in data as latent variables
- When fitting the model to data (e.g. using MLE) run EM
- Estimate a distribution on latent variables
- Maximize the expectation w.r.t. latent variables of joint log-likelihood w.r.t. parameters

- Each object has multi-dimensional discrete latent variable $\Rightarrow$ exponentially large sums
- Object has both discrete and continuous latent variables (e.g. mixture of low-dimensional manifolds) $\Rightarrow$ mixed discrete-continuous distributions over latent variables
- Continuous latent variables come from non-conjugate priors $\Rightarrow$ intractable multi-dimensional integrals
- Further approach: Large-Scale Variational Bayes

Maximum variance formulation



- $\{\mathbf{x}_i\}_{i=1}^m$, $\mathbf{x}_i \in \mathbb{R}^d$ is a sample
- Goal: project the data onto a space with dimensionality $q < d$, while maximizes the variance of the projected points
- Let $q = 1$ and denote by $\mathbf{u}_1 \in \mathbb{R}^d$ a d-dimensional vector, s.t. $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

- If we denote by $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$, then the variance of the projected data is

$$\frac{1}{m}\sum_{i=1}^{m}\{\mathbf{u}_1^\top \mathbf{x}_i - \mathbf{u}_1^\top \overline{\mathbf{x}}\}^2 = \mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1,$$

  where $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^\top$
- Setting the derivative of $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top \mathbf{u}_1)$ to zero, we get that

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

- By induction: the optimal linear projections with maximal variance are defined by the q eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_q$ of the data covariance matrix $\mathbf{S}$, corresponding to the q largest eigenvalues $\lambda_1, \ldots, \lambda_q$

- We introduce a complete orthonormal set of d-dimensional basis vectors $\{\mathbf{u}_i\}_{i=1}^d$, s.t.
$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$
- Thus it holds for any $\mathbf{x}_i$: $\mathbf{x}_i = \sum_{j=1}^d \alpha_{ij} \mathbf{u}_j$
- Due to orthonormality we get that $\alpha_{ij} = \mathbf{x}_i^\top \mathbf{u}_j$, i.e.
$$\mathbf{x}_i = \sum_{j=1}^d (\mathbf{x}_i^\top \mathbf{u}_j) \mathbf{u}_j$$
- The q-dimensional linear subspace is represented by the first q of the basis vectors, so the approximation of $\mathbf{x}_i$ is
$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^q z_{ij} \mathbf{u}_j + \sum_{j=q+1}^d b_j \mathbf{u}_j,$$
where $\{b_j\}$ are constants, that are the same for all data points

Minimum-error formulation

- The distortion measure

$$J = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$$

- Setting derivatives to zero we get that

$$\{z_{ij} = \mathbf{x}_i^\top \mathbf{u}_j\}_{j=1}^{q}, \ \{b_j = \overline{\mathbf{x}}^\top \mathbf{u}_j\}_{j=q+1}^{d}$$

- Since $\mathbf{x}_i - \widetilde{\mathbf{x}}_i = \sum_{j=q+1}^{d} \{(\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{u}_j\} \mathbf{u}_j$, then

$$J = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=q+1}^{d} (\mathbf{x}_i^\top \mathbf{u}_j - \overline{\mathbf{x}}^\top \mathbf{u}_j)^2 = \sum_{j=q+1}^{d} \mathbf{u}_j^\top \mathbf{S} \mathbf{u}_j$$

- E.g. in case $d = 2$: by minimizing

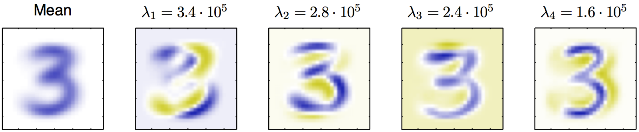
$$J = \mathbf{u}_2^\top \mathbf{S} \mathbf{u}_2 + \lambda_2 (1 - \mathbf{u}_2^\top \mathbf{u}_2)$$

  we get that

$$\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2, \; J = \lambda_2,$$

  i.e. we should choose the principal subspace to be aligned with the eigenvector having the larger eigenvalue
- In general case $\{\mathbf{u}_i\}_{i=1}^q$ are eigenvectors $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ and

$$J = \sum_{i=q+1}^{d} \lambda_i$$



- PCA approximation to a data vector $\mathbf{x}_n$

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^{q} (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j + \sum_{j=q+1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j$$

$$= \overline{\mathbf{x}} + \sum_{j=1}^{q} (\mathbf{x}_j^\top - \overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j,$$

where we used the relation $\overline{\mathbf{x}} = \sum_{i=1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_i)\mathbf{u}_i$

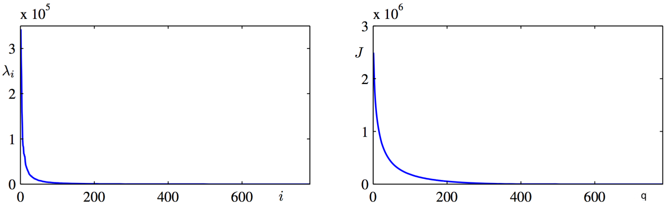


Figure – Eigenvalue spectrum (left). Sum of the discarded eigenvalues (right)

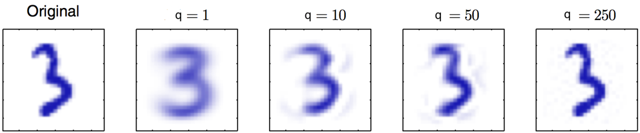
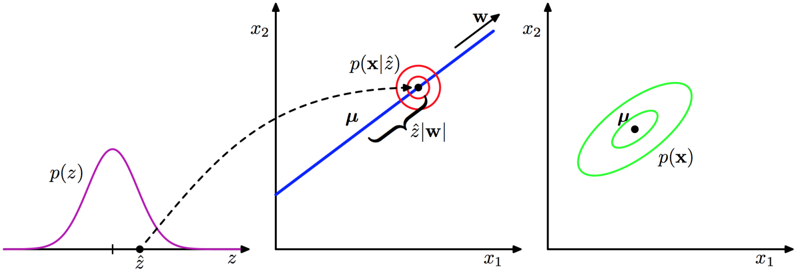


Figure – PCA reconstructions of the off-line digits data set. $q = d = 28 \times 28 = 784$ is already perfect reconstruction

Benefits of the Probabilistic PCA

- Probabilistic PCA represents a constrained form of the Gaussian distribution
- Provides EM algorithm for PCA: computationally efficient since we can calculate only needed components
- Probabilistic model $+$ EM $=$ to deal with missing values
- Mixtures of probabilistic PCA models can be formulated in a principled way and trained using the EM algorithm
- The existence of a likelihood function $\Rightarrow$ direct comparison with other probabilistic density models
- Probabilistic PCA can be used to model class-conditional densities
- The probabilistic PCA model can be run generatively to provide samples from the distribution



- We assume that $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, $\mathbf{z} \in \mathbb{R}^q$ $(q < d)$
- Similarly

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}), \text{ i.e. } \mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \; \mathbf{x} \in \mathbb{R}^d,$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I})$

- We would like to determine $\mathbf{W}$ and σ^2. Thus we need a marginal $p(\mathbf{x})$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- We get that $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$, where

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

- There is redundancy in this parametrization corresponding to rotations of the latent space coordinates: for $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$, where $\mathbf{R}$ is an orhtogonal matrix, we get that

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top = \mathbf{W}\mathbf{R}\mathbf{R}^\top\mathbf{W}^\top = \mathbf{W}\mathbf{W}^\top$$

- Inversion of $d \times d$ matrix $\mathbf{C}$:

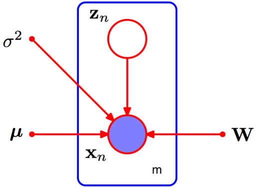
$$\mathbf{C}^{-1} = \sigma^{-1}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathsf{M}^{-1}\mathbf{W}^{\top},$$

  where $q \times q$ matrix M has the form

$$\mathsf{M} = \mathbf{W}^{\top}\mathbf{W} + \sigma^{2}\mathbf{I}$$

- Thus the cost of inverting $\mathbf{C}$ is reduced from $O(d^3)$ to $O(q^3)$
- The posterior $p(\mathbf{z}|\mathbf{x})$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathsf{M}^{-1}\mathbf{W}^{\top}(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathsf{M})$$



- Given a data set $\mathbf{X}_m = \{\mathbf{x}_i\}_{i=1}^m$ the log-likelihood

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \sum_{i=1}^{m} \log p(\mathbf{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$$

$$= -\frac{md}{2}\log(2\pi) - \frac{m}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{x}_i - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- Optimizing w.r.t. $\boldsymbol{\mu}$ we get $\boldsymbol{\mu} = \overline{\mathbf{x}}$ and

$$\log p(\mathbf{X}_m | \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d \log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

  where $\mathbf{S}$ is the data covariance matrix
- ML for $\mathbf{W}$ and σ^2: all the stationary points of the log-likelihood has the form

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

  where
  - $\mathbf{U}_q \in \mathbb{R}^{d \times q}$ is a matrix whose columns are given by any subset (of size q) of the eigenvectors of the data covariance matrix $\mathbf{S}$,
  - $\mathbf{L}_q$ is a $q \times q$ diagonal matrix with elements λ_i,
  - $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal matrix

- For an unconditional $p(\mathbf{x})$ we get that

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Thus $\mathbf{C}$ is independent of $\mathbf{R}$ for

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma_{ML}^2\mathbf{I})^{1/2}\mathbf{R}$$

- If $\mathbf{v}$ is orthogonal to the principal subspace, then $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, i.e. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- If $\mathbf{v} = \mathbf{u}_i$, then $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- For $\mathbf{R} = \mathbf{I}$ we get a usual PCA, otherwise columns of $\mathbf{W}$ need not be orthogonal

- Conventional PCA: projection of points from the d- dimensional data space onto an q-dimensional linear subspace ($d > q$)
- Probabilistic PCA: mapping from the latent space into the data space. We can reverse this mapping using Bayes theorem (visualization and data compression)
- The mean is given by

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathsf{M}^{-1}\mathbf{W}_{ML}^{\top}(\mathbf{x} - \overline{\mathbf{x}})$$

- The posterior covariance is $\mathrm{cov}[\mathbf{z}] = \sigma^2\mathsf{M}^{-1}$

- Usual Gaussian distribution: $d(d+1)/2$ parameters.
- Probabilistic PCA: define d-dimensional Gaussian retaining the q most significant correlations. The number of degress of freedom in the covariance matrix $\mathbf{C}$ is given by

$$dq + 1 - q(q-1)/2,$$

since

  - $dq + 1$ for $\mathbf{W}$ and σ^2
  - minus $q(q-1)/2$ parameters for $\mathbf{R}$ (redundancy in parametrization associated with rotations)

- We have already obtained an exact closed-form solution for the MLE. Why do we need EM?
- In spaces of high dimensionality, there may be computational advantages in using an iterative EM procedure rather than working directly with the sample covariance matrix
- General framework for EM
  - we write down the complete-data log likelihood
  - take its expectation w.r.t. the posterior distribution of the latent distribution with "old" parameters
  - maximization of this expected complete data log-likelihood then yields the "new" parameter values

- The complete-data log likelihood function takes the form

$$\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{m} \{\log p(\mathbf{x}_n | \mathbf{z}_n) + \log p(\mathbf{z}_n)\}$$

- MLE for $\boldsymbol{\mu}$ is equal to $\overline{\mathbf{x}}$, thus substituting the sample mean, and taking the expectation with respect to the posterior distribution over the latent variables

$$\begin{aligned}
\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2)] = -\sum_{n=1}^{m} \Big\{ & \frac{d}{2} \log(2\pi\sigma^2) + \frac{1}{2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top]) \\
& + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{z}_n]^\top \mathbf{W}^\top (\mathbf{x}_n - \boldsymbol{\mu}) \\
& + \frac{1}{2\sigma^2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top] \mathbf{W}^\top \mathbf{W}) \Big\}
\end{aligned}$$

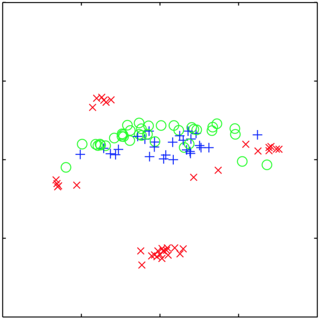
In the E step we use the old parameter values to evaluate

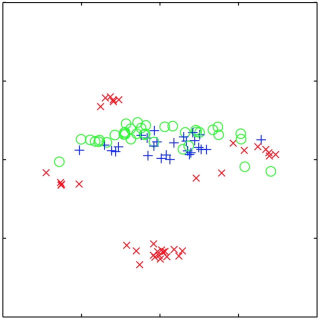
$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathrm{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^\top$$

In the M step we maximize w.r.t. $\mathbf{W}$ and σ^2:

$$\mathbf{W}_{\mathrm{new}} = \left[\sum_{n=1}^{m}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^\top\right]\left[\sum_{n=1}^{m}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\right]^{-1}$$
$$\sigma^2_{\mathrm{new}} = \frac{1}{md}\sum_{n=1}^{m}\{\|\mathbf{x}_n - \overline{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_n]^\top\mathbf{W}^\top_{\mathrm{new}}(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$+ \mathrm{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\mathbf{W}^\top_{\mathrm{new}}\mathbf{W}_{\mathrm{new}}\right)$$

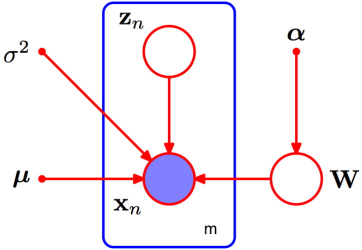
- Benefit of the iterative EM algorithm for PCA: computational efficiency for large-scale applications
- PCA: $O(d^3)$ for an eigendecomposition or $O(qd^2)$ if we need the first q eigenvectors
- However, we need $O(md^2)$ to calculate the covariance matrix.
- In case of EM algorithm we need only $O(mdq)$ steps which is better than $O(md^2)$ for $d \gg q$
- We can do EM incrementally
- Probabilistic PCA can deal with missing values by marginalizing over the distribution over unobserved variables





- Probabilistic PCA: visualization of 100 data points.
- Left: the posterior mean projections of the data points on the principal subspace.
- Right: is obtained by first randomly omitting 30% of the variable values and then using EM to handle the missing values

PCA Model Selection



- How to select q?
- We need to marginalize out the model parameters $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2
- Here we consider a simpler approach: evidence approximation
- $\boldsymbol{\alpha}$ governs which latent dimensions should be pruned

PCA Model Selection

- We use ARD prior (Automatic Relevance Determination) that allows surplus dimensions in the principal subspace to be pruned out of the model

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left\{-\frac{1}{2}\alpha_i \mathbf{w}_i^\top \mathbf{w}_i\right\}$$

- The values of α_i are re-estimated during training by maximizing the log marginal likelihood given by

$$p(\mathbf{X}_m|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{W}|\boldsymbol{\alpha}) d\mathbf{W}$$

Since the integral is not tractable, we use the Laplace approximation and an iterative estimation algorithm:

- Initialize α_i
- Apply EM-algorithm to estimate $\mathbf{W}$ and σ^2. The only change is to the M-step equation for $\mathbf{W}$

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{m} (\mathbf{x}_n - \overline{\mathbf{x}}) \mathbb{E}[\mathbf{z}_n]^\top \right] \left[\sum_{n=1}^{m} \mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top] + \sigma^2 \boldsymbol{\alpha} \right]^{-1},$$

  where $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$. The value of $\boldsymbol{\mu}$ is given by the sample mean, as before
- Re-estimate α_i maximizing $p(\mathbf{X}_m | \boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2)$:

$$\alpha_i^{\text{new}} = \frac{d}{\mathbf{w}_i^\top \mathbf{w}_i}$$

- Usually we start from some $q \leq d - 1$. If some α_i go to infinity we can delete the corresponding dimensions

EM can

- fill in missing data
- reveal data structure (manifolds, clusters)
- find hidden information in training data
- handle unknown factors caused by our choice of $\boldsymbol{\theta}$, e.g. in reinforcement learning
- be used to construct more flexible models of data with better predictive abilities
- used for large datasets, as training time is approximately the same as for analogous models without latent variables