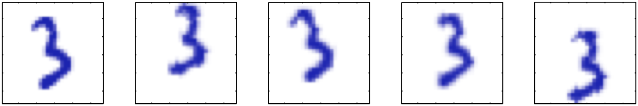
Bayesian PCA

Evgeny Burnaev

Skoltech, Moscow, Russia

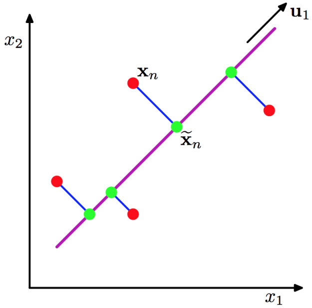
Outline

Introduction



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

Maximum variance formulation



- $\{\mathbf{x}_n\}_{n=1}^N$, $\mathbf{x}_n \in \mathbb{R}^D$ is a sample
- Goal: project the data onto a space (principal subspace) having dimensionality $M < D$, while maximizes the variance of the projected points
- Let $M = 1$ and denote by $\mathbf{u}_1 \in \mathbb{R}^D$ a D-dimensional vector, s.t. $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$

Maximum variance formulation

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

$$\frac{1}{N}\sum_{n=1}^{N}\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1,$$

  where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{x})(\mathbf{x}_n - \overline{x})^{\mathrm{T}}$

- Setting the derivative of $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1)$ to zero, we get that

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

- By induction: the optimal linear projections for which the variance of the projected data is maximized are defined by the M eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$ of the data covariance matrix $\mathbf{S}$, corresponding to the M largest eigenvalues $\lambda_1, \ldots, \lambda_M$

Minimum-error formulation

- We introduce a complete orthonormal set of D-dimensional basis vectors $\{\mathbf{u}_i\}_{i=1}^D$, s.t.

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = \delta_{ij}$$

- Thus it holds for any $\mathbf{x}_n$: $\mathbf{x}_n = \sum_{i=1}^D \alpha_{ni}\mathbf{u}_i$
- Due to orthonormality we get that $\alpha_{nj} = \mathbf{x}_n^{\mathrm{T}}\mathbf{u}_j$, i.e.

$$\mathbf{x}_n = \sum_{i=1}^D (\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i$$

- The M-dimensional linear subspace is represented by the first M of the basis vectors, so the approximation of $\mathbf{x}_n$ is

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^M z_{ni}\mathbf{u}_i + \sum_{i=M+1}^D b_i\mathbf{u}_i,$$

where $\{b_i\}$ are constants, that are the same for all data points

Minimum-error formulation

- The distortion measure

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \tilde{\mathbf{x}}_n\|^2$$

- Setting derivatives to zero we get that

$$\{z_{nj} = \mathbf{x}_n^{\mathrm{T}}\mathbf{u}_j\}_{j=1}^{M},\ \{b_j = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_j\}_{j=M+1}^{D}$$

- Since $\mathbf{x}_n - \tilde{\mathbf{x}}_n = \sum_{i=M+1}^{D}\{(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}\mathbf{u}_i\}\mathbf{u}_i$, then

$$J = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=M+1}^{D}(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i - \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i)^2 = \sum_{i=M+1}^{D}\mathbf{u}_i^{\mathrm{T}}\mathbf{S}\mathbf{u}_i$$

Minimum-error formulation

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

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

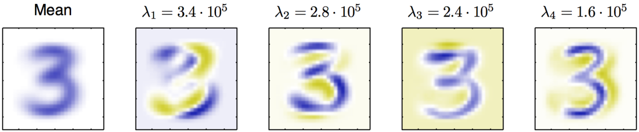
  we get that

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

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

$$J = \sum_{i=M+1}^{D} \lambda_i$$

Applications of PCA



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

$$\begin{aligned}\tilde{\mathbf{x}}_n &= \sum_{i=1}^{M}(\mathbf{x}_n^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i + \sum_{i=M+1}^{D}(\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i \\ &= \overline{\mathbf{x}} + \sum_{i=1}^{M}(\mathbf{x}_n^{\mathrm{T}} - \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{u}_i)\mathbf{u}_i,\end{aligned}$$

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

Applications of PCA

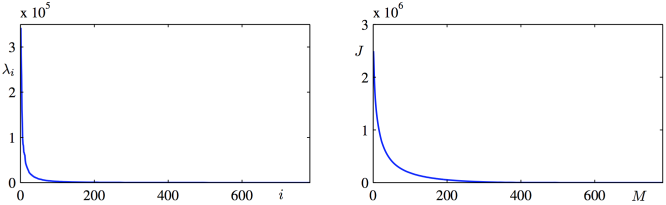


FIGURE : Eigenvalue spectrum (left). Sum of the discarded eigenvalues (right)

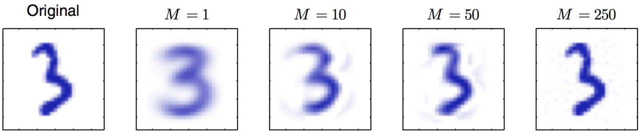
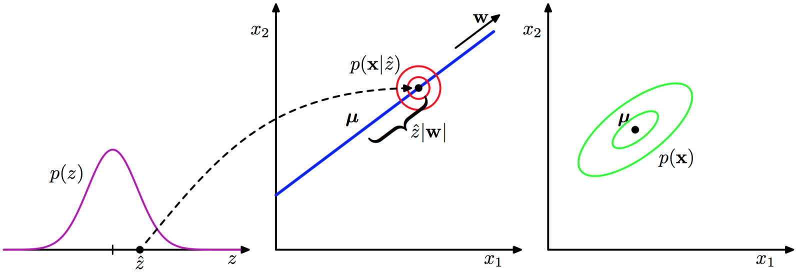


FIGURE : PCA reconstructions of the off-line digits data set. $M = D = 28 \times 28 = 784$ is already perfect reconstruction

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
- Necessary for the Bayesian treatment of PCA
- 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

Probabilistic PCA



- We assume that $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, $\mathbf{z} \in \mathbb{R}^M$, $(M < D)$
- 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I})$

Probabilistic PCA

- 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

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

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

$$\tilde{\mathbf{W}}\tilde{\mathbf{W}}^{\mathrm{T}} = \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$$

Probabilistic PCA

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

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

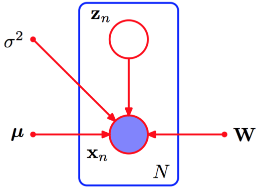
where $M \times M$ matrix M has the form

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

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

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

Maximum likelihood PCA



- Given a data set $\mathbf{X} = \{\mathbf{x}_n\}$ the log-likelihood

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

$$= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Maximum likelihood PCA

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

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

  where $\mathbf{S}$ is the data covariance matrix
- ML for $\mathbf{W}$ and σ^2

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

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

Maximum likelihood PCA

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

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

- Thus $\mathbf{C}$ is independent of $\mathbf{R}$ for $\mathbf{W}_{ML} = \mathbf{U}_M(\mathbf{L}_M - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$
- If $\mathbf{v}$ is orthogonal to the principal subspace, then $\mathbf{v}^{\mathrm{T}}\mathbf{U} = \mathbf{0}$, i.e. $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v} = \sigma^2$
- If $\mathbf{v} = \mathbf{u}_i$, then $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- For $\mathbf{R} = \mathbf{I}$ we get a usual PCA, otherwise columns of $\mathbf{W}$ need not be orthogonal

Conventional PCA vs. Bayesian PCA

- Conventional PCA: projection of points from the D-dimensional data space onto an M-dimensional linear subspace ($D > M$)
- 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

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

- The posterior covariance is $\sigma^2\mathsf{M}^{-1}$

Conventional PCA vs. Bayesian PCA

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

$$DM + 1 - M(M-1)/2,$$

since

— $DM + 1$ for $\mathbf{W}$ and σ^2

— minus $M(M-1)/2$ parameters for $\mathbf{R}$ (redundancy in parameterization associated with rotations)

EM algorithm for PCA

- 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

EM algorithm for PCA

- The complete-data log likelihood function takes the form

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

- 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

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

EM algorithm for PCA

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

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

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

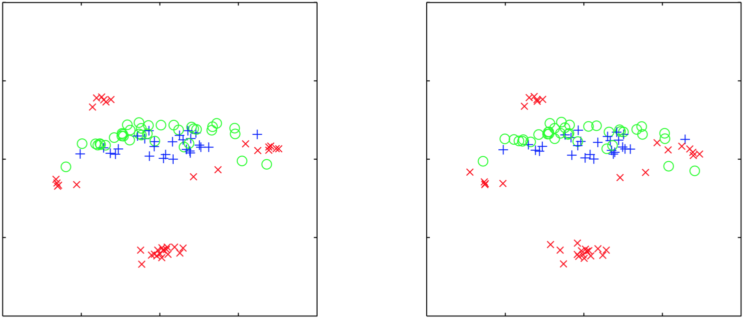
$$\mathbf{W}_{\mathrm{new}} = \left[\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]\right]^{-1}$$

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

EM vs. MLE

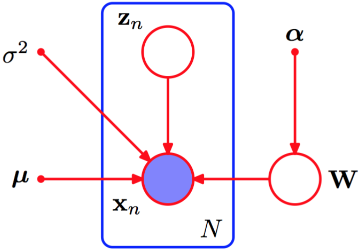
- benefit of the iterative EM algorithm for PCA: computational efficiency for large-scale applications
- PCA: $O(D^3)$ for an eigendecomposition or $O(MD^2)$ if we need the first M eigenvectors
- However, we need $O(ND^2)$ to calculate the covariance matrix. In case of EM algorithm we need only $O(NDM)$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

Effective number of parameters



- 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 M?
- 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

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

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

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

PCA Model Selection

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 $\mathbf{W}$

$$\mathbf{W}_{\mathrm{new}} = \left[\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^{\mathrm{T}}\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}] + \sigma^2\mathbf{A}\right]^{-1},$$

where $\mathbf{A} = \mathrm{diag}(\alpha_i)$. The value of $\boldsymbol{\mu}$ is given by the sample mean, as before

— Re-estimate α_i maximizing $p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2)$:

$$\alpha_i^{\mathrm{new}} = \frac{D}{\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i}$$

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