Machine learning

Principal component analysis

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Prof. Giorgio Gambosi

a.a. 2021-2022

Curse of dimensionality

In general, many features: high-dimensional spaces.

- sparseness of data
- increase in the number of coefficients, for example for dimension D and order 3 of the polynomial,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D} \sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

  number of coefficients is $O(D^M)$

High dimensions lead to difficulties in machine learning algorithms (lower reliability or need of large number of coefficients) this is denoted as curse of dimensionality

Dimensionality reduction

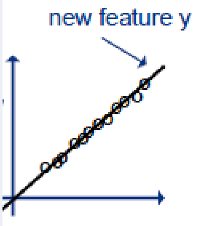
- for any given classifier, the training set size required to obtain a certain accuracy grows exponentially wrt the number of features
- it is important to bound the number of features, identifying the less discriminant ones

Dimensionality reduction

- Feature selection: identify a subset of features which are still discriminant, or, in general, still represent most dataset variance
- Feature extraction: identify a projection of the dataset onto a lower-dimensional space, in such a way to still represent most dataset variance
  - Linear projection: principal component analysis, probabilistic PCA, factor analysis
  - Non linear projection: manifold learning, autoencoders

Searching hyperplanes for the dataset

- verifying whether training set elements lie on a
  hyperplane (a space of lower dimensionality), apart from a limited variability (which could be seen as noise)



- principal component analysis looks for a d'-dimensional subspace ($d' < d$) such that the projection of elements onto such suspace is a "faithful" representation of the original dataset
- as "faithful" representation we mean that distances between elements and their projections are small, even minimal

PCA for $d' = 0$

- Objective: represent all d-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by means of a unique vector $\mathbf{x}_0$, in the most faithful way, that is so that

$$J(\mathbf{x}_0) = \sum_{i=1}^{n} \|\mathbf{x}_0 - \mathbf{x}_i\|^2$$

  is minimum
- it is easy to show that

$$\mathbf{x}_0 = \mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

- In fact,

$$
\begin{aligned}
J(\mathbf{x}_0) &= \sum_{i=1}^{n} \left\|(\mathbf{x}_0 - \mathbf{m}) - (\mathbf{x}_i - \mathbf{m})\right\|^2 \\
&= \sum_{i=1}^{n} \left\|\mathbf{x}_0 - \mathbf{m}\right\|^2 - 2\sum_{i=1}^{n} (\mathbf{x}_0 - \mathbf{m})^T(\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{m}\right\|^2 \\
&= \sum_{i=1}^{n} \left\|\mathbf{x}_0 - \mathbf{m}\right\|^2 - 2(\mathbf{x}_0 - \mathbf{m})^T \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) + \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{m}\right\|^2 \\
&= \sum_{i=1}^{n} \left\|\mathbf{x}_0 - \mathbf{m}\right\|^2 + \sum_{i=1}^{n} \left\|\mathbf{x}_i - \mathbf{m}\right\|^2
\end{aligned}
$$

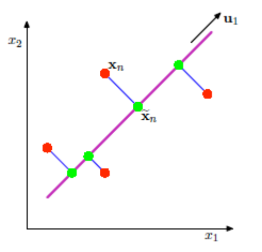
- since

$$
\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) = \sum_{i=1}^{n} \mathbf{x}_i - n \cdot \mathbf{m} = n \cdot \mathbf{m} - n \cdot \mathbf{m} = 0
$$

- the second term is independent from $\mathbf{x}_0$, while the first one is equal to zero for $\mathbf{x}_0 = \mathbf{m}$

PCA for $d' = 1$

- a single vector is too concise a representation of the dataset: anything related to data variability gets lost
- a more interesting case is the one when vectors are projected onto a line passing through $\mathbf{m}$



PCA for $d' = 1$

- let $\mathbf{u}_1$ be unit vector ($\|\mathbf{u}_1\| = 1$) in the line direction: the line equation is then

$$\mathbf{x} = \alpha\mathbf{u}_1 + \mathbf{m}$$

  where α is the distance of $\mathbf{x}$ from $\mathbf{m}$ along the line
- let $\tilde{\mathbf{x}}_i = \alpha_i\mathbf{u}_1 + \mathbf{m}$ be the projection of $\mathbf{x}_i$ ($i = 1, \dots, n$) onto the line: given $\mathbf{x}_1, \dots, \mathbf{x}_n$, we wish to find the set of projections minimizing the quadratic error

The quadratic error is defined as

$$
\begin{aligned}
J(\alpha_1, \dots, \alpha_n, \mathbf{u}_1) &= \sum_{i=1}^{n} \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|^2 \\
&= \sum_{i=1}^{n} \|(\mathbf{m} + \alpha_i \mathbf{u}_1) - \mathbf{x}_i\|^2 \\
&= \sum_{i=1}^{n} \|\alpha_i \mathbf{u}_1 - (\mathbf{x}_i - \mathbf{m})\|^2 \\
&= \sum_{i=1}^{n} +\alpha_i^2 \|\mathbf{u}_1\|^2 + \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m}) \\
&= \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|^2 - 2 \sum_{i=1}^{n} \alpha_i \mathbf{u}_1^T (\mathbf{x}_i - \mathbf{m})
\end{aligned}
$$

Its derivative wrt α_k is

$$\frac{\partial}{\partial \alpha_k} J(\alpha_1, \dots, \alpha_n, \mathbf{u}_1) = 2\alpha_k - 2\mathbf{u}_1^T(\mathbf{x}_k - \mathbf{m})$$

which is zero when $\alpha_k = \mathbf{u}_1^T(\mathbf{x}_k - \mathbf{m})$ (the orthogonal projection of $\mathbf{x}_k$ onto the line).

The second derivative turns out to be positive

$$\frac{\partial}{\partial \alpha_k^2} J(\alpha_1, \dots, \alpha_n, \mathbf{u}_1) = 2$$

showing that what we have found is indeed a minimum.

PCA for $d' = 1$

To derive the best direction $\mathbf{u}_1$ of the line, we consider the covariance matrix of the dataset

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$$

By plugging the values computed for α_i into the definition of $J(\alpha_1, \dots, \alpha_n, \mathbf{u}_1)$, we get

$$\begin{aligned} J(\mathbf{u}_1) &= \sum_{i=1}^{n}\alpha_i^2 + \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{m}\|^2 - 2\sum_{i=1}^{n}\alpha_i^2 \\ &= -\sum_{i=1}^{n}[\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})]^2 + \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{m}\|^2 \\ &= -\sum_{i=1}^{n}\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1 + \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{m}\|^2 \\ &= -n\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{m}\|^2 \end{aligned}$$

- $\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})$ is the projection of $\mathbf{x}_i$ onto the line
- the product
$$\mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1$$
is then the variance of the projection of $\mathbf{x}_i$ wrt the mean $\mathbf{m}$
- the sum
$$\sum_{i=1}^{n} \mathbf{u}_1^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{u}_1 = n\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$$
is the overall variance of the projections of vectors $\mathbf{x}_i$ wrt the mean $\mathbf{m}$

PCA for $d' = 1$

Minimizing $J(\mathbf{u}_1)$ is equivalent to maximizing $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$. That is, $J(\mathbf{u}_1)$ is minimum if $\mathbf{u}_1$ is the direction which keeps the maximum amount of variance in the dataset

Hence, we wish to maximize $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$ (wrt $\mathbf{u}_1$), with the constraint $\|\mathbf{u}_1\| = 1$.

By applying Lagrange multipliers this results equivalent to maximizing

$$u = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 - \lambda_1(\mathbf{u}_1^T\mathbf{u}_1 - 1)$$

This can be done by setting the first derivative wrt $\mathbf{u}_1$:

$$\frac{\partial u}{\partial \mathbf{u}_1} = 2\mathbf{S}\mathbf{u}_1 - 2\lambda_1\mathbf{u}_1$$

to 0, obtaining

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

PCA for $d' = 1$

Note that:

- u is maximized if $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$
- the overall variance of the projections is then equal to the corresponding eigenvalue

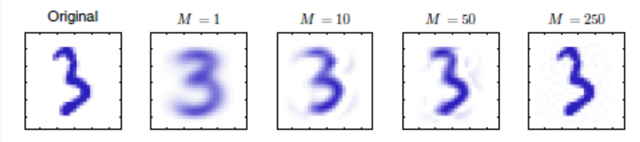
$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \mathbf{u}_1^T\lambda_1\mathbf{u}_1 = \lambda_1\mathbf{u}_1^T\mathbf{u}_1 = \lambda_1$$

- the variance of the projections is then maximized (and the error minimized) if $\mathbf{u}_1$ is the eigenvector of $\mathbf{S}$ corresponding to the maximum eigenvalue λ_1

PCA for $d' > 1$

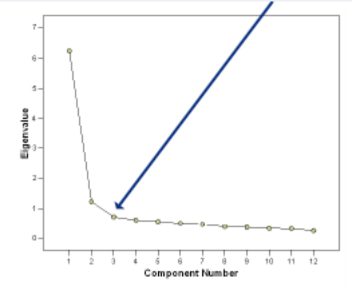
- The quadratic error is minimized by projecting vectors onto a hyperplane defined by the directions associated to the d' eigenvectors corresponding to the d' largest eigenvalues of $\mathbf{S}$
- If we assume data are modeled by a d-dimensional gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, PCA returns a d'-dimensional subspace corresponding to the hyperplane defined by the eigenvectors associated to the d' largest eigenvalues of Σ
- The projections of vectors onto that hyperplane are distributed as a d'-dimensional distribution which keeps the maximum possible amount of data variability

- Digit recognition ($D = 28 \times 28 = 784$)



Choosing d'

Eigenvalue size distribution is usually characterized by a fast initial decrease followed by a small decrease



This makes it possible to identify the number of eigenvalues to keep, and thus the dimensionality of the projections.

Eigenvalues measure the amount of distribution variance kept in the projection.

Let us consider, for each $k < d$, the value

$$r_k = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$$

which provides a measure of the variance fraction associated to the k largest eigenvalues.

When $r_1 < \ldots < r_d$ are known, a certain amount p of variance can be kept by setting

$$d' = \operatorname*{argmin}_{i \in \{1,\ldots,d\}} r_i > p$$

Probabilistic approach to PCA: idea

Introduce a latent variable model to relate a d-dimensional observation vector to a corresponding d'-dimensional gaussian latent variable (with $d' < d$)

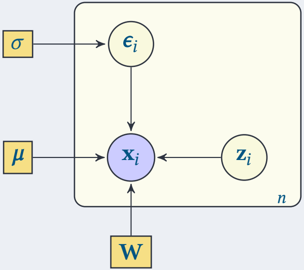
$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

where

- $\mathbf{z}$ is a d'-dimensional gaussian latent variable (the "projection" of $\mathbf{x}$ on a lower-dimensional subspace)
- $\mathbf{W}$ is a $d \times d'$ matrix, relating the original space with the lower-dimensional subspace
- $\boldsymbol{\epsilon}$ is a d-dimensional gaussian noise: noise covariance on different dimensions is assumed to be 0. Noise variance is assumed equal on all dimensions: hence $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
- $\boldsymbol{\mu}$ is the d-dimensional vector of the means

$\boldsymbol{\epsilon}$ and $\boldsymbol{\mu}$ are assumed independent.

Graphical model



1. $\mathbf{z} \in \mathbb{R}^{d'}, \mathbf{x}, \boldsymbol{\epsilon} \in \mathbb{R}^{d}, d' < d$
2. $p(\mathbf{z}) = N(\mathbf{0}, \mathbf{I})$
3. $p(\boldsymbol{\epsilon}) = N(\mathbf{0}, \sigma^2\mathbf{I})$, (isotropic gaussian noise)

Generative process

This can be interpreted in terms of a generative process

1. sample the latent variable $\mathbf{z} \in \mathbb{R}^{d'}$ from

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{d'/2}} e^{-\frac{\|\mathbf{z}\|^2}{2}}$$

2. linearly project onto $\mathbb{R}^d$

$$\mathbf{y} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu}$$

3. sample the noise component $\boldsymbol{\epsilon} \in \mathbb{R}^d$ from

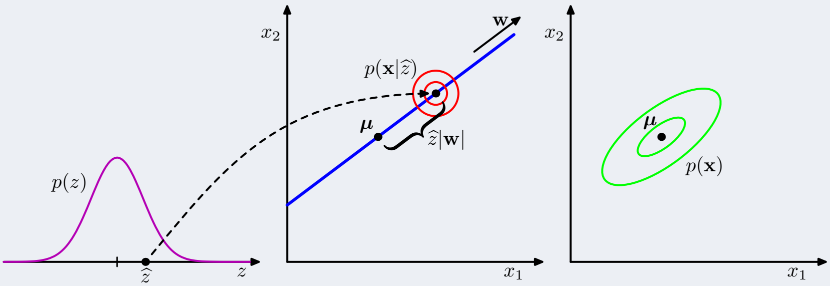
$$p(\boldsymbol{\epsilon}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|\boldsymbol{\epsilon}\|^2}{2\sigma^2}}$$

4. add the noise component $\boldsymbol{\epsilon}$

$$\mathbf{x} = \mathbf{y} + \boldsymbol{\epsilon}$$

This results into $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$

Generative process



Probability recall

Let

$$\mathbf{x}_1 \in \mathbb{R}^r \qquad \mathbf{x}_2 \in \mathbb{R}^s \qquad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$$

Assume $\mathbf{x}$ is normally distributed: $p(\mathbf{x}) = N(\boldsymbol{\mu}, \Sigma)$, and let

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Under the above assumptions:

- The marginal distribution $p(\mathbf{x}_1)$ is a gaussian on $\mathbb{R}^r$, with

$$E[\mathbf{x}_1] = \boldsymbol{\mu}_1$$
$$\mathrm{Cov}(\mathbf{x}_1) = \Sigma_{11}$$

- The conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ is a gaussian on $\mathbb{R}^r$, with

$$E[\mathbf{x}_1|\mathbf{x}_2] = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$
$$\mathrm{Cov}(\mathbf{x}_1|\mathbf{x}_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The joint distribution is

$$p\left(\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \end{bmatrix}\right) = N(\boldsymbol{\mu}_{\mathbf{zx}}, \Sigma)$$

Joint distribution mean

By definition,

$$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{z}} \\ \boldsymbol{\mu}_{\mathbf{x}} \end{bmatrix}$$

- Since $p(\mathbf{z}) = N(\mathbf{0}, \mathbf{I})$, then $\boldsymbol{\mu}_{\mathbf{z}} = 0$.
- Since $p(\mathbf{x}) = \mathbf{Wz} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, then

$$\boldsymbol{\mu}_{\mathbf{x}} = E[\mathbf{x}] = E[\mathbf{Wz} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \mathbf{W}E[\mathbf{z}] + \boldsymbol{\mu} + E[\boldsymbol{\epsilon}] = \boldsymbol{\mu}$$

Hence

$$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix}$$

Joint distribution covariance

For what concerns the distribution covariance

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{zz}} & \Sigma_{\mathbf{zx}} \\ \Sigma_{\mathbf{zx}} & \Sigma_{\mathbf{xx}} \end{bmatrix}$$

where

$$\Sigma_{\mathbf{zz}} = E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{z} - E[\mathbf{z}])^T] = E[\mathbf{z}\mathbf{z}^T] = \mathbf{I}$$
$$\Sigma_{\mathbf{zx}} = E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{x} - E[\mathbf{x}])^T] = \mathbf{W}^T$$
$$\Sigma_{\mathbf{xx}} = E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Latent variable model

Joint distribution

As a consequence, we get

$$\boldsymbol{\mu}_{\mathbf{zx}} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \mathbf{I} & \mathbf{W}^T \\ \mathbf{W} & \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I} \end{bmatrix}$$

Marginal distribution

The marginal distribution of $\mathbf{x}$ is then $p(\mathbf{x}) = N(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$

Conditional distribution

The conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ is $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}}, \Sigma_{\mathbf{z}|\mathbf{x}})$ with

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{x}} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

$$\Sigma_{\mathbf{z}|\mathbf{x}} = \mathbf{I} - \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}\mathbf{W} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1}$$

Setting $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$, the log-likelihood of the dataset in the model is

$$\begin{aligned}\log p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)\\ &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_n - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})^T\end{aligned}$$

Setting the derivative wrt $\boldsymbol{\mu}$ to zero results into

$$\boldsymbol{\mu} = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

and, substituting into the log-likelihood formula,

$$\log p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{nd}{2}\log(2\pi) + \log|\mathbf{C}| + \operatorname{tr}(\mathbf{C}^{-1}\mathbf{S})$$

where $\mathbf{S}$ is the data covariance matrix

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Maximum likelihood for PCA

Maximization wrt $\mathbf{W}$ and σ^2 is more complex: however, a closed form solution exists:

$$\mathbf{W} = \mathbf{U}_{d'}(\mathbf{L}_{d'} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$$

where

- $\mathbf{U}_{d'}$ is the $d \times d'$ matrix whose columns are the eigenvectors corresponding to the d' largest eigenvalues
- $\mathbf{L}_{d'}$ is the $d' \times d'$ diagonal matrix of the largest eigenvalues
- $\mathbf{R}$ is an arbitrary $d' \times d'$ orthogonal matrix, corresponding to a rotation in the latent space

$\mathbf{R}$ can be interpreted as a rotation matrix in latent space.

If $\mathbf{R} = \mathbf{I}$, the columns of $\mathbf{W}$ are the principal components eigenvectors scaled by the variance $\lambda_i - \sigma^2$

Maximum likelihood for PCA

For what concerns maximization wrt σ^2, it results

$$\sigma^2 = \frac{1}{d - d'} \sum_{i=d'+1}^{d} \lambda_i$$

since eigenvalues provide measures of the dataset variance along the corresponding eigenvector direction, this corresponds to the average variance along the discarded directions.

Mapping points to subspace

The conditional distribution

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

can be applied.

In particular, the conditional expectation

$$E[\mathbf{z}|\mathbf{x}] = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

can be assumed as the latent space point corresponding to $\mathbf{x}$.

The projection onto the d'-dimensional subspace can then be performed as

$$\mathbf{x}' = \mathbf{W}E[\mathbf{z}|\mathbf{x}] + \boldsymbol{\mu} = \mathbf{W}\mathbf{W}^T(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \boldsymbol{\mu}$$