CSE 575
Statistical Machine Learning

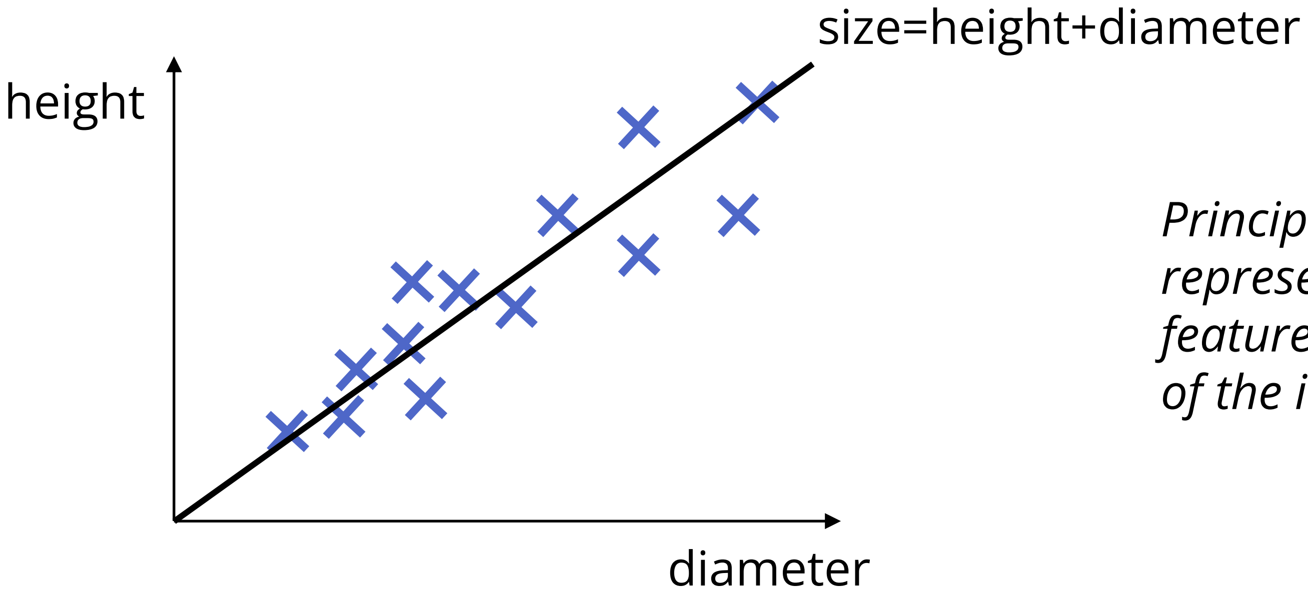
Lecture 18

YooJung Choi

Fall 2022

Dimensionality reduction: example

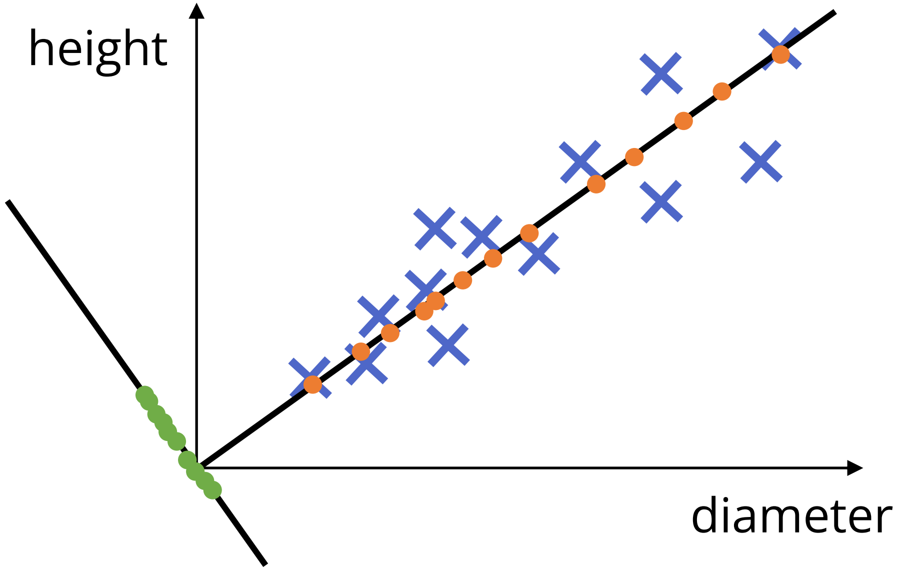
- Two features: height and diameter of trees
- The features are correlated
- We can characterize the size of a tree using a single feature



Principal component analysis: represent the data using fewer features, each a linear combination of the input features

Principal component analysis

- Problem: Given a D-dimensional data, map each $\mathbf{x}_n$ to an M-dimensional $\mathbf{z}_n = \mathbf{U}^T\mathbf{x}_n$
- First, consider projection onto a one-dimensional space
- Let $\mathbf{u}$ be the vector defining the direction of projection
- Then each data point $\mathbf{x}_n$ is projected onto 1D (scalar) $\mathbf{u}^T\mathbf{x}_n$
- What is the best direction of projection?
- Want to maximize the variance!



PCA: one-dimensional

- Mean of the projected data:

$$\frac{1}{\mathrm{N}}\sum_{n=1}^{N}\mathbf{u}^T\mathbf{x}_n = \mathbf{u}^T\left(\frac{1}{\mathrm{N}}\sum_{n=1}^{N}\mathbf{x}_n\right) = \mathbf{u}^T\boldsymbol{\mu}$$

- Variance of the projected data:

$$\frac{1}{\mathrm{N}}\sum_{n=1}^{N}(\mathbf{u}^T\mathbf{x}_n - \mathbf{u}^T\boldsymbol{\mu})^2 = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}(\mathbf{u}^T\mathbf{x}_n - \mathbf{u}^T\boldsymbol{\mu})(\mathbf{u}^T\mathbf{x}_n - \mathbf{u}^T\boldsymbol{\mu})^T = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\mathbf{u}^T(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T\mathbf{u}$$

$$= \mathbf{u}^T\left(\frac{1}{\mathrm{N}}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T\right)\mathbf{u} = \mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}$$

- Note: you can trivially increase variance by $||\mathbf{u}|| \to \infty$. Thus, we constrain $||\mathbf{u}|| = 1$
- Maximize $\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}$ s.t. $||\mathbf{u}||^2 \le 1$
- Using Lagrange multiplier, maximize $\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} + \lambda(1 - \mathbf{u}^T\mathbf{u})$

PCA: one-dimensional

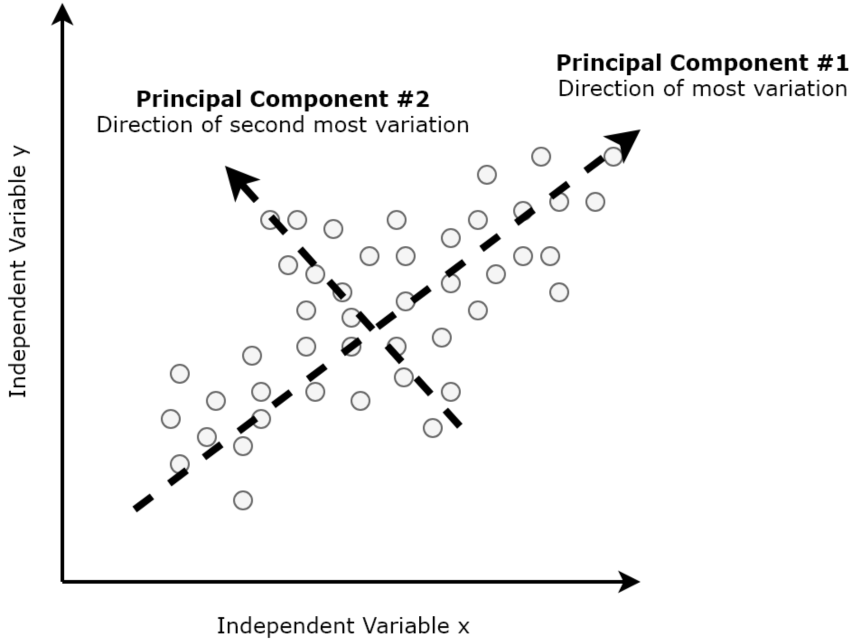
- Maximize $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u})$
- Unconstrained optimization w.r.t. $\mathbf{u}$, so we can set the partial to zero and solve for $\mathbf{u}$:

$$\frac{\partial}{\partial \mathbf{u}}\left(\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} + \lambda(1 - \mathbf{u}^T \mathbf{u})\right) = 2\boldsymbol{\Sigma}\mathbf{u} - 2\lambda\mathbf{u} = 0 \qquad \Rightarrow \boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}$$

- i.e. $\mathbf{u}$ is an eigenvector of $\boldsymbol{\Sigma}$, with the eigenvalue λ
- Variance: $\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} = \mathbf{u}^T(\lambda\mathbf{u}) = \lambda\mathbf{u}^T \mathbf{u} = \lambda$
  - maximized when $\mathbf{u}$ is the eigenvector having the largest eigenvalue λ
  - *i.e. $\mathbf{u}$ is the first principal component*

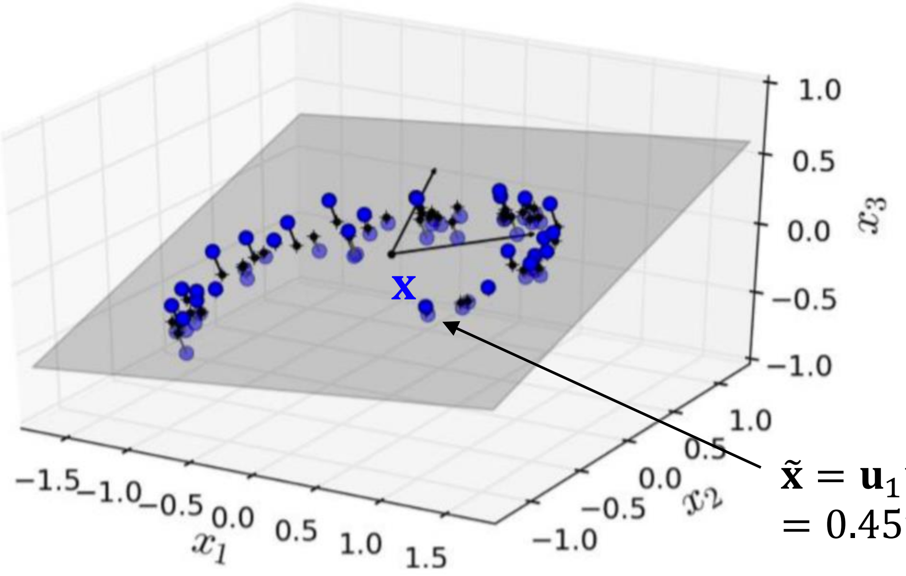
PCA: M-dimensional

- Project onto M eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{M}$ of $\mathbf{\Sigma}$ having the M *largest eigenvalues* $\lambda_1, \ldots, \lambda_M$

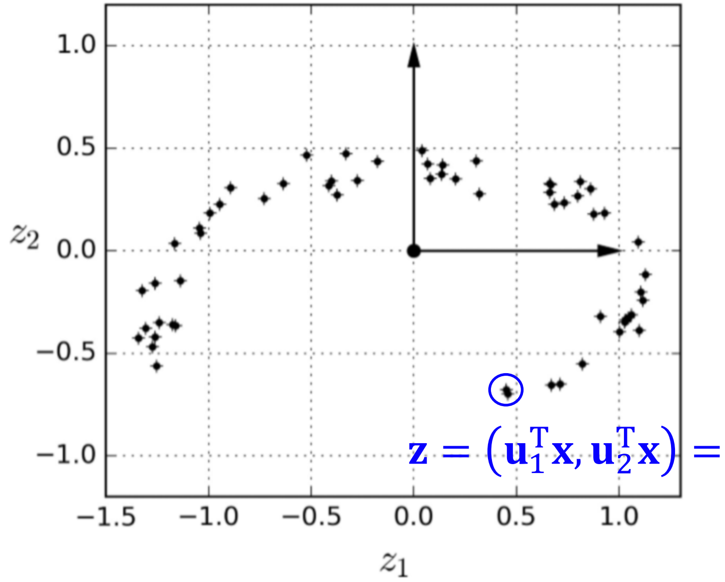


PCA: M-dimensional

- Project onto M eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_\mathrm{M}$ of $\mathbf{\Sigma}$ having the M *largest eigenvalues* $\lambda_1, \ldots, \lambda_M$
- Each $\mathbf{x}$ is transformed into $\mathbf{z} = \mathbf{U}^T\mathbf{x}$ where $\mathbf{U}^T = \begin{bmatrix} -\mathbf{u}_1^T - \\ \vdots \\ -\mathbf{u}_\mathrm{M}^T - \end{bmatrix}$ *i.e.* $\mathbf{z} = \begin{bmatrix} \mathbf{u}_1^T\mathbf{x} \\ \vdots \\ \mathbf{u}_\mathrm{M}^T\mathbf{x} \end{bmatrix}$



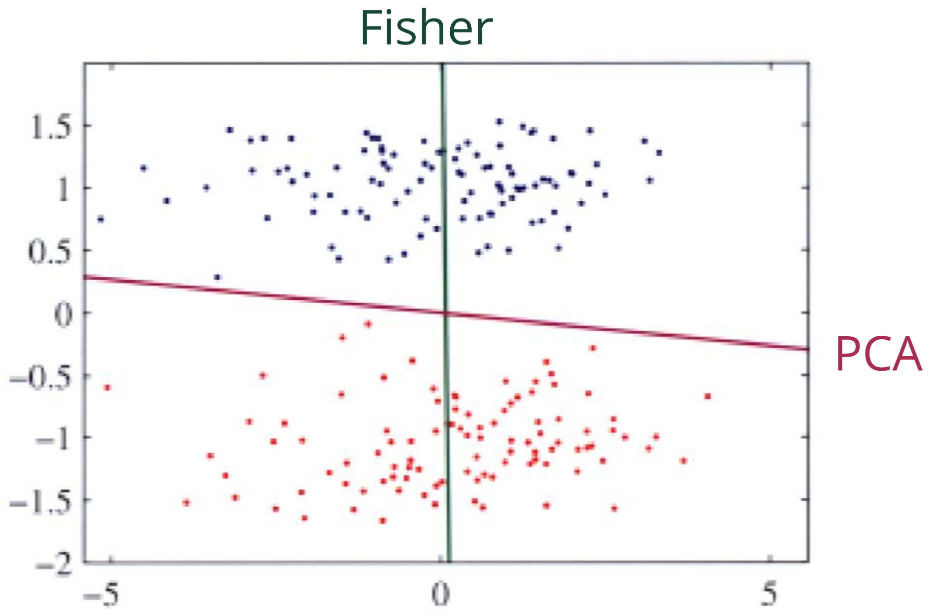
$\tilde{\mathbf{x}} = \mathbf{u}_1\mathbf{u}_1^\mathrm{T}\mathbf{x} + \mathbf{u}_2\mathbf{u}_2^\mathrm{T}\mathbf{x}$
$= 0.45\mathbf{u}_1 - 0.7\mathbf{u}_2$



$\mathbf{z} = \left(\mathbf{u}_1^\mathrm{T}\mathbf{x}, \mathbf{u}_2^\mathrm{T}\mathbf{x}\right) = (0.45, -0.7)$

PCA and Fisher linear discriminant

- Recall: Fisher's linear discriminant projects data points onto a single dimension, on the direction that gives the best class separation
- On the other hand, PCA projects data points onto the direction of maximum variance



Standardization

- Principal component analysis tries to capture the most variance using lower dimensional vectors
- We do not want one feature to have significantly higher variance than other features
- Solution: standardize data to have zero mean and unit variance

$$\frac{x_i - \mu_i}{\sigma_i} \quad \text{for each feature } i = 1, \ldots, D \text{ where } \mu_i = \frac{1}{N}\sum_{n=1}^{N} x_{ni}, \qquad \sigma_i^2 = \frac{1}{N}\sum_{n=1}^{N} (x_{ni} - \mu_i)^2$$

- After standardization, $\boldsymbol{\Sigma} = \frac{1}{\mathrm{N}}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T = \frac{1}{\mathrm{N}}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n{}^T = \frac{1}{\mathrm{N}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$
- For PCA, get the eigendecomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$
- Equivalently, singular value decomposition (SVD) of $\mathbf{X}$

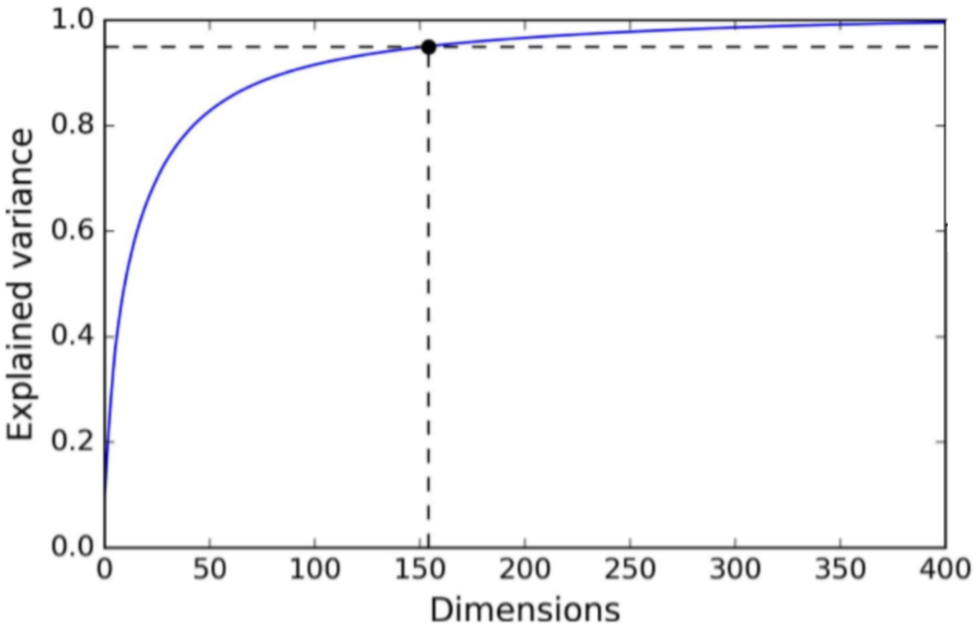
Design matrix

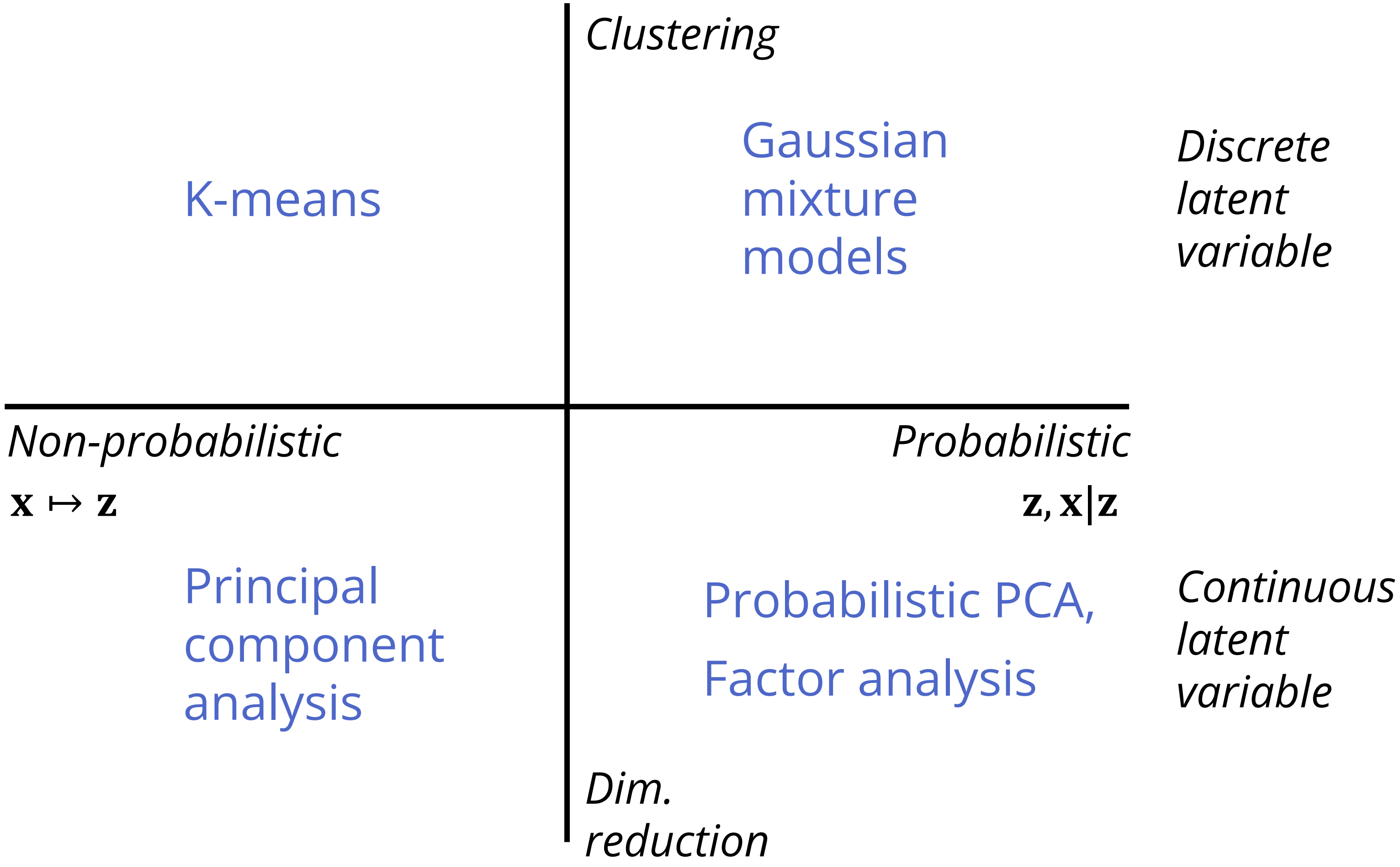
$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T - \\ \vdots \\ -\mathbf{x}_N^T - \end{bmatrix}$$

Explained variance

- We can first find all eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ then choose the M principal components
- I.e. we can choose the value of M after seeing the eigenvectors & eigenvalues
- Choose the M to get sufficiently high *explained variance*

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



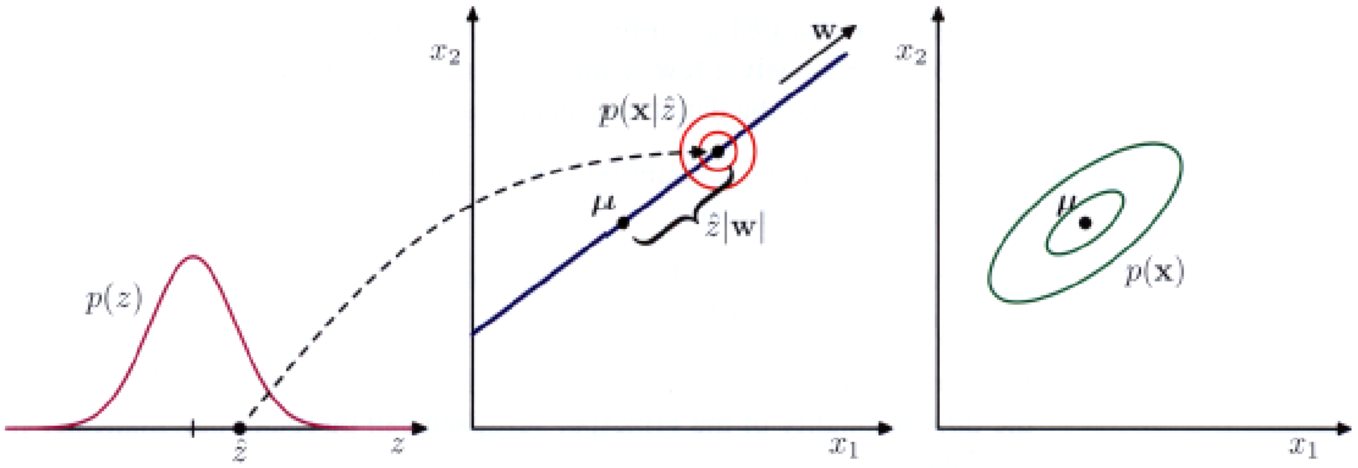

Clustering
K-means
Gaussian mixture models
Discrete latent variable
Non-probabilistic
$\mathbf{x} \mapsto \mathbf{z}$
Probabilistic
$\mathbf{z}, \mathbf{x}|\mathbf{z}$
Principal component analysis
Probabilistic PCA,
Factor analysis
Continuous latent variable
Dim. reduction

Probabilistic PCA

- Recall: Gaussian mixture models $p(z=k)=\pi_k,\;\; p(\mathbf{x}|z=k)=\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$
- Probabilistic PCA: assume a *continuous* M-dimensional latent variable $\mathbf{z}$

$$p(\mathbf{z})=\mathcal{N}(\mathbf{z}\,|\,\mathbf{0},\mathbf{I})$$

- Conditional distribution of observed variables given by: $p(\mathbf{x}\,|\,\mathbf{z})=\mathcal{N}(\mathbf{x}\,|\,\mathbf{W}\mathbf{z}+\boldsymbol{\mu},\sigma^2\mathbf{I})$
- Equivalently, $\mathbf{x}=\mathbf{W}\mathbf{z}+\boldsymbol{\mu}+\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{\epsilon}\,|\,\mathbf{0},\sigma^2\mathbf{I})$



Generate $\mathbf{x}$ *by:*

1. *sampling* $\mathbf{z}$ *from a zero-mean, unit-covariance Gaussian*
2. *sampling* $\mathbf{x}$ *from a Gaussian centered at* $\mathbf{W}\mathbf{z}+\boldsymbol{\mu}$ *with a spherical covariance*

Probabilistic PCA

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}), \qquad p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mathbf{Wz} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

- Marginal distribution

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \mathbf{z}) \cdot p(\mathbf{z})\, d\mathbf{z} = \int \mathcal{N}(\mathbf{x} \mid \mathbf{Wz} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \cdot \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})\, d\mathbf{z} = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$$

Product of Gaussian densities is also a Gaussian & Marginal of Gaussian is Gaussian

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{Wz} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \mathbb{E}[(\mathbf{Wz} + \boldsymbol{\epsilon})(\mathbf{Wz} + \boldsymbol{\epsilon})^T]$$

$$= \mathbb{E}[\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^T] + \mathbb{E}[\mathbf{Wz}]\mathbb{E}[\boldsymbol{\epsilon}^T] + \mathbb{E}[\boldsymbol{\epsilon}]\,\mathbb{E}[\mathbf{z}^T\mathbf{W}^T] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$$

Maximum-likelihood PCA

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \underbrace{\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}}_{\mathbf{C}})$$

- Parameters: $\boldsymbol{\mu}, \mathbf{W}, \sigma^2$
- Log-likelihood: $\sum_{n=1}^{N} \log p(\mathbf{x}_n) = -\frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$
- Exact closed-form solution for MLE:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \qquad \mathbf{W} = \mathbf{U}_{\mathrm{M}}(\mathbf{L}_M - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \qquad \sigma^2 = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

where $\mathbf{U}_{\mathrm{M}}$ a DxM matrix of M principal eigenvectors of the data covariance matrix $\boldsymbol{\Sigma}$,

$\mathbf{L}_M$ an MxM diagonal matrix of the corresponding eigenvalues,

$\mathbf{R}$ an arbitrary MxM orthogonal matrix *(treat as a rotation matrix in the latent space)*

Maximum-likelihood PCA

- Property of multivariate Gaussians: posterior distribution $p(\mathbf{z} \mid \mathbf{x})$ is also a Gaussian

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

- Suppose we map each $\mathbf{x}$ to $\mathbb{E}[\mathbf{z} \mid \mathbf{x}]$

$$\mathbb{E}[\mathbf{z} \mid \mathbf{x}] = (\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu}) \quad \rightarrow (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T(\mathbf{x}-\boldsymbol{\mu}) \text{ as } \sigma^2 \rightarrow 0$$

Orthogonal projection onto the latent space => standard PCA!

- Exact closed-form solution for MLE:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\,, \qquad \mathbf{W} = \mathbf{U}_{\mathrm{M}}(\mathbf{L}_M - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \qquad \sigma^2 = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

where $\mathbf{U}_{\mathrm{M}}$ a DxM matrix of M principal eigenvectors of the data covariance matrix $\mathbf{\Sigma}$,

$\mathbf{L}_M$ an MxM diagonal matrix of the corresponding eigenvalues,

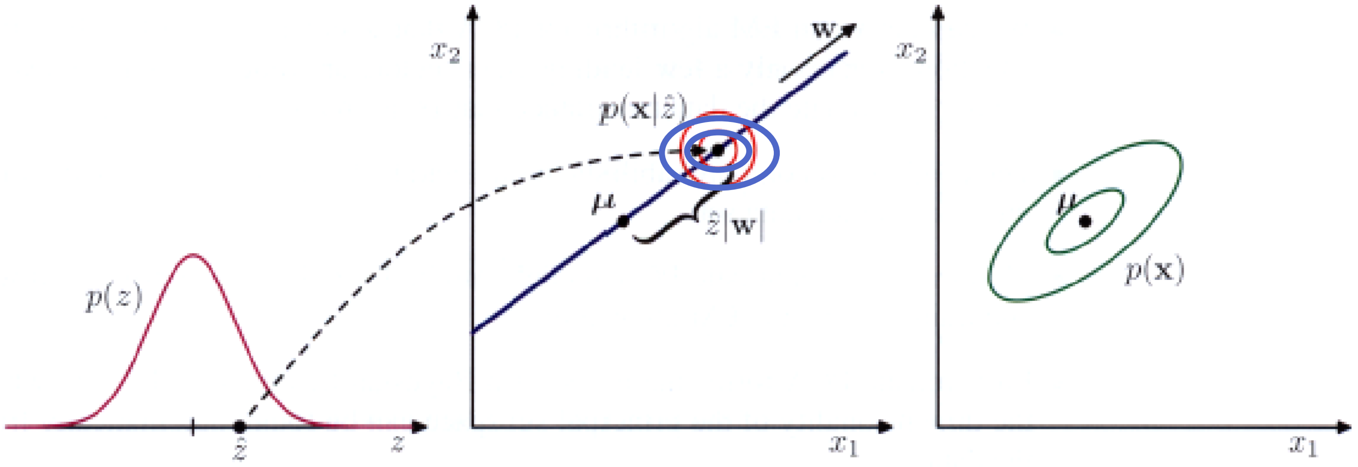
$\mathbf{R}$ an arbitrary MxM orthogonal matrix *(treat as a rotation matrix in the latent space)*

Factor analysis

- A *continuous* M-dimensional latent variable $\mathbf{z}$ and conditional distribution with a *diagonal* covariance:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}), \qquad p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

- Similar to probabilistic PCA, marginal distribution: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi})$
- No longer a closed form MLE solution. Learn by expectation maximization



Non-probabilistic	*Probabilistic*
$\mathbf{x} \mapsto \mathbf{z}$	$\mathbf{z}, \mathbf{x}\vert\mathbf{z}$
Principal component analysis	Probabilistic PCA, Factor analysis

Dim. reduction

Covariance in the latent space

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

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi})$$

"explain" the variance by

$\mathbf{W}\mathbf{W}^T$	*(PCA)*
$\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$	*(PPCA)*
$\mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$	*(FA)*