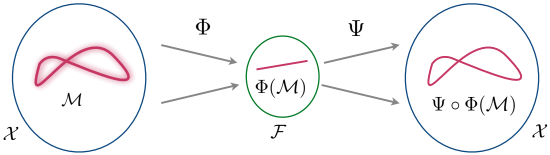
MIT 9.520/6.860, Fall 2017
Statistical Learning Theory and Applications

Class 20: Dictionary Learning

What is data representation?

Let $\mathcal{X}$ be a data-space



A **data representation** is a map

$$\Phi : \mathcal{X} \to \mathcal{F},$$

from the data space to a **representation space** $\mathcal{F}$.

A **data reconstruction** is a map

$$\Psi : \mathcal{F} \to \mathcal{X}.$$

Road map

Last class:

- Prologue: **Learning theory** and data representation
- Part I: Data representations by **design**

This class:

- Part II: Data representations by **unsupervised learning**
  - Dictionary Learning
  - PCA
  - Sparse coding
  - K-means, K-flats

Next class:

- Part III: **Deep** data representations

Notation

$\mathcal{X}$: data space

- $\mathcal{X} = \mathbb{R}^d$ or $\mathcal{X} = \mathbb{C}^d$ (also more general later).
- $x \in \mathcal{X}$

Data representation: $\Phi : \mathcal{X} \to \mathcal{F}$.

$$\forall x \in \mathcal{X}, \exists z \in \mathcal{F} : \Phi(x)$$

$\mathcal{F}$: representation space

- $\mathcal{F} = \mathbb{R}^p$ or $\mathcal{F} = \mathbb{C}^p$
- $z \in \mathcal{F}$

Data reconstruction: $\Psi : \mathcal{F} \to \mathcal{X}$.

$$\forall z \in \mathcal{F}, \exists x \in \mathcal{X} : \Psi(z) = x$$

Why learning?

Ideally: automatic, autonomous learning

- with as **little prior information** as possible,

  but also......

- ...with as **little human supervision** as possible.

$$f(x) = \langle w, \Phi(x) \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}$$

Two-step learning scheme:

- **supervised or unsupervised learning** of $\Phi : \mathcal{X} \to \mathcal{F}$
- *supervised learning* of w in $\mathcal{F}$

Unsupervised representation learning

Samples from a distribution ρ on input space $\mathcal{X}$

$$S = \{x_1, \dots, x_n\} \sim \rho^n$$

Training set S from ρ (supported on $\mathcal{X}_\rho$).

Goal: find $\Phi(x)$ which is "good" not only for S but for other $x \sim \rho$.

Principles *for unsupervised learning of "good" representations?*

Unsupervised representation learning principles

Two main concepts:

1. **Similarity preservation**, it holds

$$\Phi(x) \sim \Phi(x') \Leftrightarrow x \sim x', \quad \forall x \in \mathcal{X}$$

2. **Reconstruction**, there exists a map $\Psi : \mathcal{F} \to \mathcal{X}$ such that

$$\Psi \circ \Phi(x) \sim x, \quad \forall x \in \mathcal{X}$$

Plan

We will first introduce a **reconstruction based** framework for learning data representation, and then discuss in some detail several **examples**.

We will mostly consider $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F} = \mathbb{R}^p$

- **Representation**: $\Phi : \mathcal{X} \to \mathcal{F}$.
- **Reconstruction**: $\Psi : \mathcal{F} \to \mathcal{X}$.

If linear maps:

- **Representation**: $\Phi(x) = Cx$ (coding)
- **Reconstruction**: $\Psi(z) = Dz$ (decoding)

Reconstruction based data representation

Basic idea: the quality of a representation Φ is measured by the **reconstruction error** provided by an associated reconstruction Ψ

$$\|x - \Psi \circ \Phi(x)\|,$$

$\Psi \circ \Phi$: denotes the composition of Φ and Ψ

Empirical data and population

Given $S = \{x_1, \dots, x_n\}$ minimize the **empirical reconstruction error**

$$\widehat{\mathcal{E}}(\Phi, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \Psi \circ \Phi(x_i)\|^2 ,$$

as a proxy to the **expected reconstruction error**

$$\mathcal{E}(\Phi, \Psi) = \int_{\mathcal{X}} d\rho(x) \, \|x - \Psi \circ \Phi(x)\|^2 ,$$

where ρ is the data distribution (fixed but uknown).

Empirical data and population

$$\min_{\Phi,\Psi} \mathcal{E}(\Phi,\Psi), \quad \mathcal{E}(\Phi,\Psi) = \int_{\mathcal{X}} d\rho(x) \left\| x - \Psi \circ \Phi(x) \right\|^2,$$

Caveat

Reconstruction alone is **not enough**...
copying data, i.e. $\Psi \circ \Phi = I$, gives zero reconstruction error!

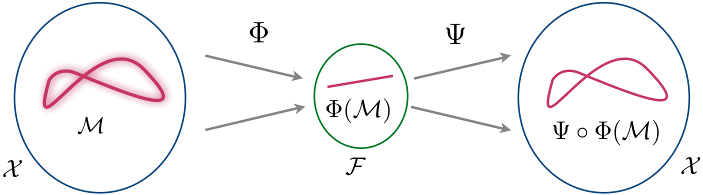
Parsimonious reconstruction

Reconstruction is meaningful only with **constraints**!

- constraints implement some form of **parsimonious** reconstruction,
- identified with a form of **regularization**,
- choice of the constraints corresponds to **different algorithms**.

Fundamental difference with supervised learning: problem is not well defined!

Parsimonious reconstruction



Dictionary learning

$$\|x - \Psi \circ \Phi(x)\|$$

Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{F} = \mathbb{R}^p$.

1. **linear reconstruction**

$$\Psi(z) = Dz, \quad D \in \mathcal{D},$$

   with $\mathcal{D}$ a subset of the space of linear maps from $\mathcal{X}$ to $\mathcal{F}$.

2. **nearest neighbor representation**,

$$\Phi(x) = \Phi_\Psi(x) = \operatorname*{arg\,min}_{z \in \mathcal{F}_\lambda} \|x - Dz\|^2, \quad D \in \mathcal{D}, \quad \mathcal{F}_\lambda \subset \mathcal{F}.$$

Linear reconstruction and dictionaries

Reconstruction $D \in \mathcal{D}$ can be identified by a $d \times p$ **dictionary** matrix with columns

$$a_1, \dots, a_p \in \mathbb{R}^d.$$

Reconstruction of $x \in \mathcal{X}$ corresponds to a suitable **linear expansion** on the dictionary D with coefficients $\beta_k = z^k, z \in \mathcal{F}_\lambda$

$$x = Dz = \sum_{k=1}^{p} a_k z^k = \sum_{k=1}^{p} a_k \beta_k, \qquad \beta_1, \dots, \beta_k \in \mathbb{R}.$$

Nearest neighbor representation

$$\Phi(x) = \Phi_\Psi(x) = \arg\min_{z \in \mathcal{F}_\lambda} \|x - Dz\|^2, \quad D \in \mathcal{D}, \quad \mathcal{F}_\lambda \subset \mathcal{F}.$$

Nearest neighbor (NN) representation since, for $D \in \mathcal{D}$ and letting

$$\mathcal{X}_\lambda = D\mathcal{F}_\lambda,$$

$\Phi(x)$ provides the **closest** point to x in $\mathcal{X}_\lambda$,

$$d(x, \mathcal{X}_\lambda) = \min_{x' \in \mathcal{X}_\lambda} \|x - x'\|^2 = \min_{z' \in \mathcal{F}_\lambda} \|x - Dz'\|^2.$$

Nearest neighbor representation (cont.)

NN representation are defined by a **constrained inverse problem**,

$$\min_{z \in \mathcal{F}_\lambda} \|x - Dz\|^2 .$$

Alternatively, let $\mathcal{F}_\lambda = \mathcal{F}$ and add a **regularization term** $R : \mathcal{F} \to \mathbb{R}$

$$\min_{z \in \mathcal{F}} \left\{ \|x - Dz\|^2 + \lambda R(z) \right\} .$$

Note: Formulations **coincide** for $R(z) = \mathbb{1}_{F_\lambda}$, $z \in \mathcal{F}$.

Dictionary learning

Empirical reconstruction error minimization

$$\min_{\Phi,\Psi} \widehat{\mathcal{E}}(\Phi,\Psi) = \min_{\Phi,\Psi} \frac{1}{n}\sum_{i=1}^{n} \|x_i - \Psi \circ \Phi(x_i)\|^2$$

for **joint** dictionary and representation learning:

$$\underbrace{\min_{D \in \mathcal{D}}}_{\text{Dictionary learning}} \quad \frac{1}{n}\sum_{i=1}^{n} \underbrace{\min_{z_i \in \mathcal{F}_\lambda} \|x_i - Dz_i\|^2}_{\text{Representation learning}}.$$

Dictionary learning

- learning a **regularized representation** on a dictionary,
- **while** simultaneously **learning the dictionary** itself.

Examples

The DL framework encompasses a number of approaches.

- PCA (& kernel PCA)
- K-SVD
- Sparse coding
- K-means
- K-flats
- ...

Principal Component Analysis (PCA)

Let $\mathcal{F}_\lambda = \mathcal{F}_k = \mathbb{R}^k$, $k \leq \min\{n, d\}$, and

$$\mathcal{D} = \{D : \mathcal{F} \to \mathcal{X}, \text{ linear} \mid D^* D = I\}.$$

- D is a $d \times k$ matrix with **orthogonal, unit norm** columns
- Reconstruction:

$$Dz = \sum_{j=1}^{k} a_j z^j, \quad z \in \mathcal{F}$$

- Representation:

$$D^* : \mathcal{X} \to \mathcal{F}, \quad D^* x = (\langle a_1, x \rangle, \ldots, \langle a_k, x \rangle), \quad x \in \mathcal{X}$$

PCA and subset selection

$$DD^* : \mathcal{X} \to \mathcal{X}, \quad DD^* x = \sum_{j=1}^{k} a_j \langle a_j, x \rangle, \quad x \in \mathcal{X}.$$

$P = DD^*$ is a **projection**[1] on subspace of $\mathbb{R}^d$ **spanned** by $a_1, \dots, a_k$.

[1] $P = P^2$ (idempotent)

Rewriting PCA

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \underbrace{\min_{z_i \in \mathcal{F}_k} \left\| x_i - D z_i \right\|^2}_{\text{Representation learning}} .$$

Note that:

$$\Phi(x) = D^* x = \arg\min_{z \in \mathcal{F}_k} \left\| x - D z \right\|^2, \quad \forall x \in \mathcal{X},$$

Rewrite minimization (set $z = D^* x$) as

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - D D^* x_i \right\|^2 .$$

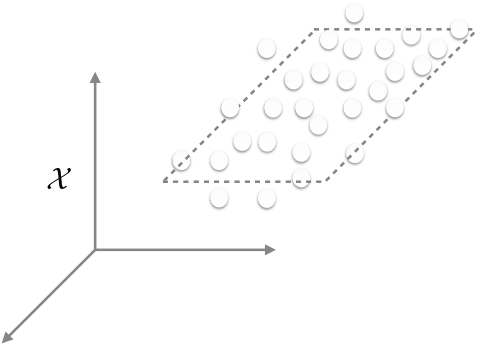
Subspace learning

Finding the $k-$dimensional orthogonal projection D^* with the best (empirical) reconstruction.

Learning a linear representation with PCA

Subspace learning

Finding the $k-$dimensional orthogonal projection with the best reconstruction.



PCA computation

Recall the solution for $k = 1$.

For all $x \in \mathcal{X}$,

$$DD^* x = \langle a, x \rangle a,$$

$$\|x - \langle a, x \rangle a\|^2 = \|x\|^2 - |\langle a, x \rangle|^2$$

with $a \in \mathbb{R}^d$ such that $\|a\| = 1$.

Then, equivalently:

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - DD^* x_i\|^2 \Leftrightarrow \max_{a \in \mathbb{R}^d, \|a\|=1} \frac{1}{n} \sum_{i=1}^{n} |\langle a, x_i \rangle|^2.$$

PCA computation (cont.)

Let $\widehat{X}$ the $n \times d$ data matrix and $V = \frac{1}{n}\widehat{X}^T\widehat{X}$.

$$\frac{1}{n}\sum_{i=1}^{n} |\langle a, x_i \rangle|^2 = \frac{1}{n}\sum_{i=1}^{n} \langle a, x_i \rangle \langle a, x_i \rangle = \left\langle a, \frac{1}{n}\sum_{i=1}^{n} \langle a, x_i \rangle x_i \right\rangle = \langle a, Va \rangle .$$

Then, equivalently:

$$\max_{a \in \mathbb{R}^d, \|a\|=1} \frac{1}{n}\sum_{i=1}^{n} |\langle a, x_i \rangle|^2 \Leftrightarrow \max_{a \in \mathbb{R}^d, \|a\|=1} \langle a, Va \rangle$$

PCA is an eigenproblem

$$\max_{a \in \mathbb{R}^d, \|a\|=1} \langle a, Va \rangle$$

- Solutions are the stationary points of the *Lagrangian*

$$\mathcal{L}(a, \lambda) = \langle a, Va \rangle - \lambda(\|a\|^2 - 1).$$

- Set $\partial \mathcal{L} / \partial a = 0$, then

$$Va = \lambda a, \quad \langle a, Va \rangle = \lambda$$

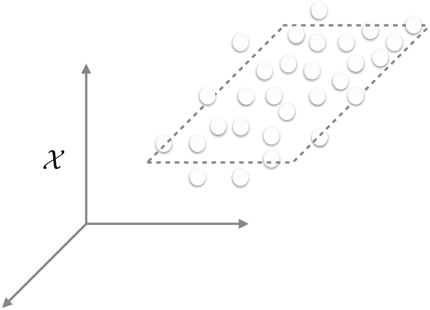
.

Optimization problem is solved by the eigenvector of V associated to the largest eigenvalue.

Note: reasoning extends to $k > 1$ – solution is given by the first k eigenvectors of V.

PCA model

Assumes the support of the data distribution is well approximated by a low dimensional *linear* subspace.



Can we consider an **affine** *representation?*

Can we consider **non-linear** *representations using PCA?*

PCA and affine dictionaries

Consider the problem, with $\mathcal{D}$ as in PCA:

$$\min_{D\in\mathcal{D}, b\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \min_{z_i\in\mathcal{F}_k} \left\| x_i - Dz_i - b \right\|^2 .$$

The above problem is **equivalent** to

$$\min_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \left\| \overline{x}_i - \underbrace{DD^*}_{P} \overline{x}_i \right\|^2$$

with $\overline{x}_i = x_i - m$, $i = 1\ldots, n$.

Note:
- Computations are unchanged but need to consider *centered* data.

PCA and affine dictionaries (cont.)

$$\min_{D\in\mathcal{D},b\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \min_{z_i\in\mathcal{F}_k} \|x_i - Dz_i - b\|^2 \Leftrightarrow \min_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \|\overline{x}_i - DD^*\overline{x}_i\|^2$$

Proof.

- Note that $\Phi(x) = D^*(x-b)$ (by optimality for z), so that

$$\min_{D\in\mathcal{D},b\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \|x_i - b - P(x_i - b)\|^2 = \min_{D\in\mathcal{D},b\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \|Q(x_i - b)\|^2,$$

  with $P = DD^*$ and $Q = I - P$.

- Solving with respect to b,

$$Qb = Qm, \quad m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

  so that

$$\Phi(x) = D^*(x-m).$$

□

Projective coordinates

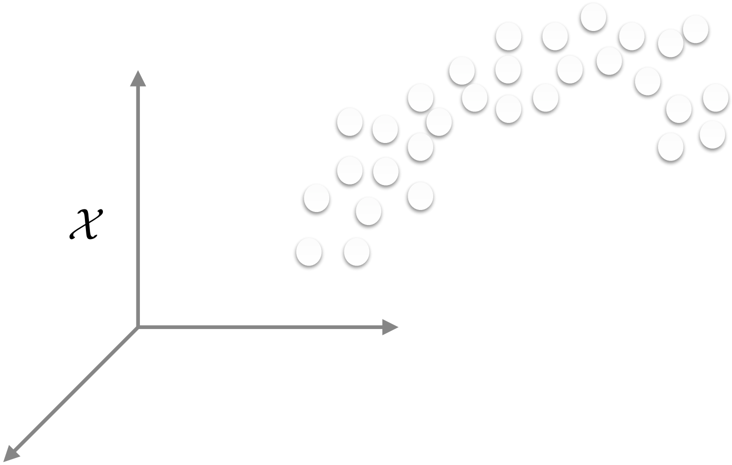
We can rewrite

$$Dz - b = D'z',$$

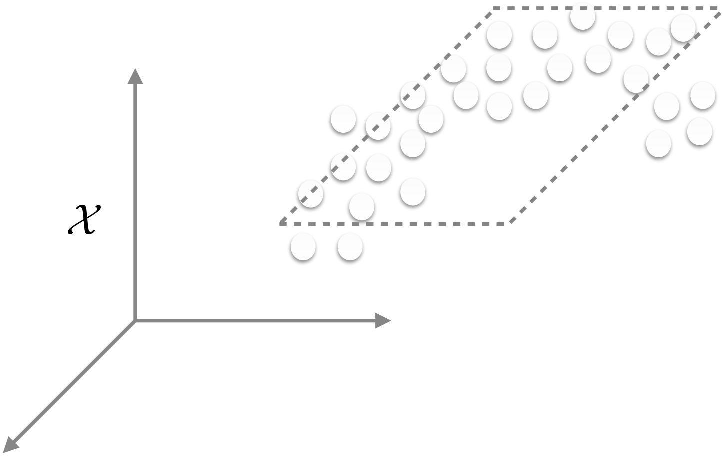
if we let

- D': matrix obtained by adding to D a column equal to b
- z': vector obtained by adding to z a coordinate equal to 1.

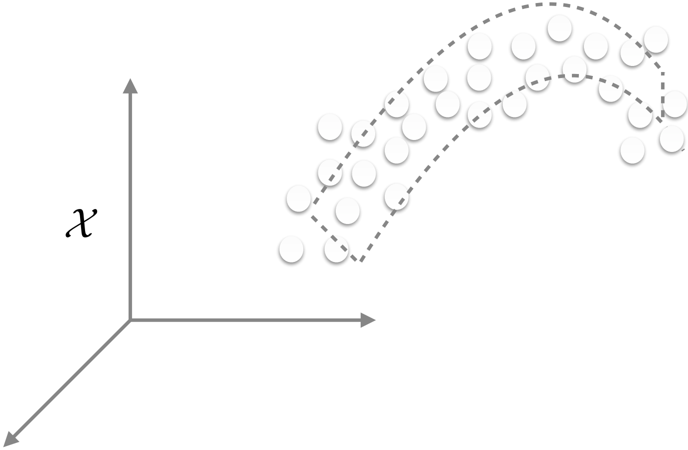
PCA beyond linearity



PCA beyond linearity



PCA beyond linearity



Kernel PCA

Consider a **feature map and associated (reproducing) kernel**.

$$\tilde{\Phi} : \mathcal{X} \to \mathcal{F}, \quad \text{and} \quad K(x, x') = \left\langle \tilde{\Phi}(x), \tilde{\Phi}(x') \right\rangle_{\mathcal{F}}$$

Empirical **reconstruction error in the feature space**,

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \min_{z_i \in \mathcal{F}_k} \left\| \tilde{\Phi}(x_i) - D z_i \right\|_{\mathcal{F}}^2 .$$

Kernel PCA (cont.)

Similar to (linear) PCA (for $k = 1$),

$$\max_{a \in \mathcal{F}, \|a\|_{\mathcal{F}} = 1} \langle a, Va \rangle_{\mathcal{F}}$$

where

$$Va = \frac{1}{n} \sum_{i=1}^{n} \left\langle \tilde{\Phi}(x_i), a \right\rangle_{\mathcal{F}} \tilde{\Phi}(x_i).$$

Representation is given by:

$$\Phi(x) = \left\langle v, \tilde{\Phi}(x) \right\rangle_{\mathcal{F}}, \forall x \in \mathcal{X},$$

with v is the eigenvector of V with largest eigenvalue.

This can be **computed for arbitrary feature map/kernel**.

A representer theorem for kernel PCA

$$\Phi(x) = \left\langle \tilde{\Phi}(x), v \right\rangle_{\mathcal{F}} = \frac{1}{n\sigma} \sum_{i=i}^{n} K(x_i, x) u^i.$$

Proof Linear case: $K(x, x') = \langle x, x' \rangle$, for all $x, x' \in \mathcal{X}$.

- Let $\frac{1}{n}\widehat{K} = \frac{1}{n}\widehat{X}\widehat{X}^T$, $V = \frac{1}{n}\widehat{X}^T\widehat{X}$.
- V and $\widehat{K}$ have same (non-zero) eigenvalues.
- If u is an eigenvector of $\widehat{K}$ with eigenvalue σ, $\widehat{K}u = \sigma u$

$$v = \frac{1}{n\sigma} X^T u = \frac{1}{n\sigma} \sum_{i=i}^{n} x_i u^i$$

  is an eigenvector of V also with eigenvalue σ.

Then, for all $x \in \mathcal{X}$,

$$\Phi(x) = \langle x, v \rangle = \frac{1}{n\sigma} \sum_{i=i}^{n} \langle x_i, x \rangle \, u^i.$$

Extends to any **arbitrary kernel**: $x \mapsto \tilde{\Phi}(x)$, $\left\langle \tilde{\Phi}(x), \tilde{\Phi}(x') \right\rangle_{\mathcal{F}} = K(x, x')$.

Comments on PCA, KPCA

- PCA allows to find good representation for data distribution supported close to a **linear/affine subspace**.
- **Non-linear** extension using kernels.

Note:

- Connection between KPCA and **manifold learning**, e.g. Laplacian/Diffusion maps.
- Off-set/re-centering **not needed** if kernel is *rich enough*.

Sparse coding

One of the first and most famous dictionary learning techniques.

It corresponds to

- $\mathcal{F} = \mathbb{R}^p$,
- $p \geq d$, $\mathcal{F}_\lambda = \{z \in \mathcal{F} \;:\; \|z\|_1 \leq \lambda\}, \quad \lambda > 0$,
- $\mathcal{D} = \{D : \mathcal{F} \to \mathcal{X} \mid \|De_j\|_{\mathcal{F}} \leq 1\}$.

Hence,

$$\underbrace{\min_{D \in \mathcal{D}}}_{\text{dictionary learning}} \quad \frac{1}{n} \sum_{i=1}^{n} \underbrace{\min_{z_i \in \mathcal{F}_\lambda} \|x_i - Dz_i\|^2}_{\text{sparse representation}}$$

Computations for sparse coding

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \min_{z_i \in \mathbb{R}^p, \|z_i\|_1 \leq \lambda} \|x_i - Dz_i\|^2$$

- **not convex** jointly in $(D, \{z_i\})$...
- **separately convex** in the $\{z_i\}$ and D.

- Alternating Minimization is natural
  - Fix D, compute $\{z_i\}$.
  - Fix $\{z_i\}$, compute D.
- (other approaches possible–see e.g. [Schnass '15, Elad et al. '06])

Representation computation

1. Given dictionary D,

$$\min_{z_i \in \mathbb{R}^p, \|z_i\|_1 \leq \lambda} \|x_i - Dz_i\|^2, i = 1, \dots, n$$

Problems are convex and correspond to a **sparse estimation**.

Solved using **convex optimization** techniques.

Splitting/proximal methods

$$z^{(0)}, \quad z^{(t+1)} = S_\lambda(z^{(t)} - \gamma_t D^*(x_i - Dz^{(t)})), \quad t = 0, \dots, t_{\max}$$

with S_λ the soft-thresholding operator,

$$S_\lambda(u) = \max\{|u| - \lambda, 0\}\frac{u}{|u|}, \quad u \in \mathbb{R}$$

.

Dictionary computation

2. Given the representation $\{\Phi(x_i) = z_i\}, i = 1, \dots, n$

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - D\Phi(x_i)\|^2 = \min_{D \in \mathcal{D}} \frac{1}{n} \left\| \widehat{X} - Z^* D \right\|_F^2 ,$$

where Z is the $n \times p$ matrix with rows z_i and $\|\cdot\|_F$, the Frobenius norm.

Problem is convex. Solvable using **convex optimization** techniques.

Splitting/proximal methods

$$D^{(0)}, \quad D^{(t+1)} = P(D^{(t)} - \gamma_t B^*(X - D^{(t)} B)), \quad t = 0, \dots, t_{\max}$$

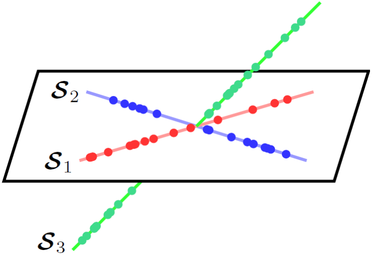
with P the prox operator (projection) from the constraints $(\|De_j\|_{\mathcal{F}} \leq 1)$

$$P(D^j) = D^j / \left\| D^j \right\|, \quad \text{if} \left\| D^j \right\| > 1,$$

$$P(D^j) = D^j, \quad \text{if} \left\| D^j \right\| \leq 1.$$

Sparse coding model

- Assumes support of the data distribution to be a **union of** $\binom{p}{s}$ **subspaces**, i.e. all possible s-dimensional subspaces in $\mathbb{R}^p$, where s is the sparsity level. [2]

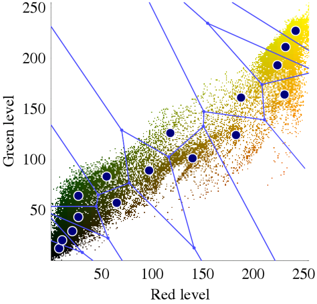


- More general penalties, more general geometric assumptions.

[2] Image credit: Elhamifar, Eldar, 2013

K-means & vector quantization

Typically seen as a **clustering** algorithm in machine learning. . .
but it is also a classical **vector quantization (VQ)** approach. [3]



We revisit this point of view from a **data representation** perspective.

[3] Image:Wikipedia

K-means & vector quantization (cont.)

K-means corresponds to

- $\mathcal{F}_\lambda = \mathcal{F}_k = \{e_1, \dots, e_k\}$, the canonical basis in $\mathbb{R}^k$, $k \leq n$
- $\mathcal{D} = \{D : \mathcal{F} \to \mathcal{X} \mid \text{linear}\}$.

Empirical reconstruction error:

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \min_{z_i \in \{e_1, \dots, e_k\}} \|x_i - Dz_i\|^2$$

Problem is **not convex** (in $(D, \{z_i\})$. Approximate solution through AM.

K-means solution

Alternating minimization (Lloyd's algorithm)

Initialize dictionary D.

1. Let $\{\Phi(x_i) = z_i\}, i = 1, \ldots, n$ be the solutions of problems

$$\min_{z_i \in \{e_1, \ldots, e_k\}} \|x_i - Dz_i\|^2, \quad i = 1, \ldots, n.$$

Assignment:

$$V_j = \{x \in S \mid \Phi(x) = z = e_j\}.$$

(multiple points have same representation since $k \leq n$).

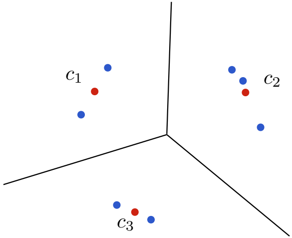
2. Update: Let $a_j = De_j$ (single dictionary atom)

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \|x_i - D\Phi(x_i)\|^2 = \min_{a_1, \ldots, a_k \in R^d} \frac{1}{n} \sum_{j=1}^{k} \sum_{x \in V_j} \|x - a_j\|^2.$$

Step 1: assignment

Solving the discrete problem:

$$\min_{z_i \in \{e_1, \ldots, e_k\}} \|x_i - Dz_i\|^2, \quad i = 1, \ldots, n.$$



Voronoi sets - Data clusters

$$V_j = \{x \in S \mid z = \Phi(x) = e_j\}, \quad j = 1 \ldots k$$

Step 2: dictionary update

$$\min_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \|x_i - D\Phi(x_i)\|^2 = \min_{a_1,\ldots,a_k\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{k}\sum_{x\in V_j} \|x - a_j\|^2 .$$

where $\Phi(x_i) = z_i,\ a_j = De_j$.

Minimization wrt. each column a_j of D is **independent** to all others.

Centroid computation

$$c_j = \operatorname*{arg\,min}_{a_j\in\mathbb{R}^d} \sum_{x\in V_j} \|x - a_j\|^2 = \frac{1}{|V_j|}\sum_{x\in V_j} x =, \quad j = 1,\ldots,k.$$

Minimimum for each column is the centroid of corresponding Voronoi set.

K-means convergence

Algorithm for solving K-means is known as **Lloyd's algorithm**.

- **Alternating minimization** approach:
  $\Longrightarrow$ value of the objective function can be shown to be **non-increasing** with the iterations.

- Only a **finite** number of possible partitions in k clusters:
  $\Longrightarrow$ ensured to **converge to a local minimum** in a finite number of steps.

K-means initialization

Convergence to a **global** minimum can be ensured (with high probability), provided a suitable initialization.

Intuition: spreading out the initial k centroids.

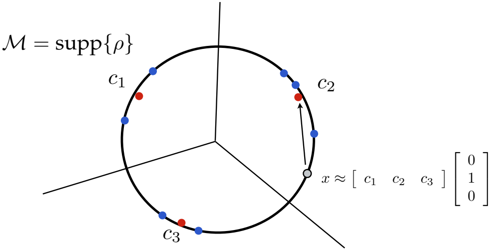
K-means++ [Arthur, Vassilvitskii;07]

1. Choose a centroid uniformly at random from the data.
2. Compute distances of data to the nearest centroid already chosen.

$$D(x, \{c_j\}) = \min_{c_j} \|x - c_j\|^2, \forall x \in S, j < k$$

3. Choose a new centroid from the data using probabilities proportional to such distances.
4. Repeat steps 2 and 3 until k centers have been chosen.

K-means model

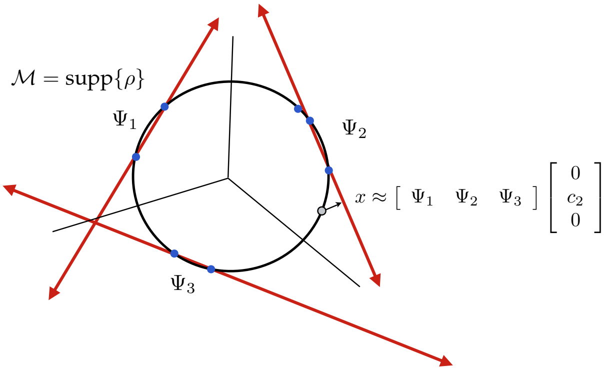


- representation: **extreme sparse representation**, only one non-zero coefficient (**vector quantization**).

- reconstruction: **piecewise constant** approximation of the data, each point is reconstructed by the nearest mean.

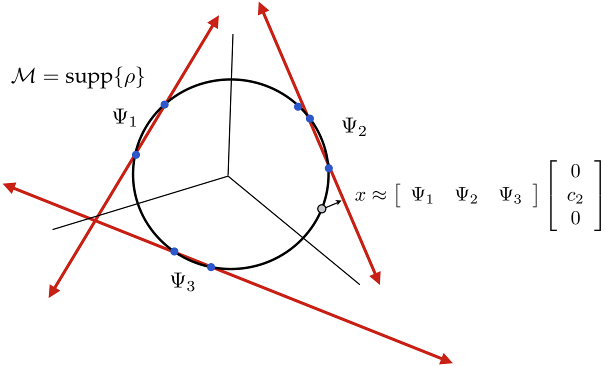
Extensions considering higher order approximation, e.g. **piecewise linear.**

K-flats & piece-wise linear representation



- k-flats representation: **structured sparse representation**, coefficients are projection on *flat*.
- k-flats reconstruction: **piecewise linear** approximation of the data, each point is reconstructed by projection on the nearest flat.

Remarks on K-flats



- Principled way to **enrich** k-means representation (cfr *softmax*).
- Generalized VQ.
- **Geometric structured** dictionary learning.
- **Non-local** approximations.

K-flats computations

Alternating minimization

1. **Initialize** flats $\Psi_1, \dots, \Psi_k$.
2. **Assign** point to nearest flat,

$$V_j = \{x \in S \mid \left\| x - \Psi_j \Psi_j^* x \right\| \leq \left\| x - \Psi_t \Psi_t^* x \right\|, \quad t \neq j\}.$$

3. **Update** flats by computing (local) PCA in each cell V_j, $j = 1, \dots, k$.

Kernel K-means & K-flats

It is easy to extend K-means & K-flats using **kernels**.

$$\tilde{\Phi} : \mathcal{X} \to \mathcal{H}, \quad \text{and} \quad K(x, x) = \left\langle \tilde{\Phi}(x), \tilde{\Phi}(x') \right\rangle_{\mathcal{H}}$$

Consider the empirical reconstruction problem in the feature space,

$$\min_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \min_{z_i \in \{e_1, \ldots, e_k\} \subset \mathcal{H}} \left\| \tilde{\Phi}(x_i) - D z_i \right\|_{\mathcal{H}}^2 .$$

Note: Computation can be performed in closed form

- Kernel K-means: **distance computation**.
- Kernel K-flats: **distance computation + local KPCA**.

Wrap up

Parsimonious reconstruction

Algorithms, computations & models.

Have not talk about:

- **Statistics/stability**

$$\mathbb{P}\left(\left| \min_{\mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \min_{z_i \in \mathcal{F}_k} \|x_i - Dz_i\|^2 - \min_{\mathcal{D}} \int d\rho(x) \min_{z \in \mathcal{F}_k} \|x - Dz\|^2 \right| > \epsilon \right)$$

- **Geometry/quantization**

$$\lim_{k \to \infty} \min_{\mathcal{D}} \int d\rho(x) \min_{z \in \mathcal{F}_k} \|x - Dz\|^2 \to 0$$

- **Computations**: non convex optimization? algorithmic guarantees?

Road map

This class:

- Part II: Data representations by **unsupervised learning**
  - Dictionary Learning
  - PCA
  - Sparse coding
  - K-means, K-flats

Next class:

- Part III: **Deep** data representations (unsupervised, supervised)
  - Neural Networks basics
  - Autoencoders
  - ConvNets