k-SVD for dictionary learning

Aharon, Elad, Bruckstein ‘06

Geelon So
(agso@eng.ucsd.edu)

May 20, 2020

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 54, NO. 11, NOVEMBER 2006 4311

K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation

Michal Aharon, Michael Elad, and Alfred Bruckstein

Introduction

k-means problem, or vector quantization

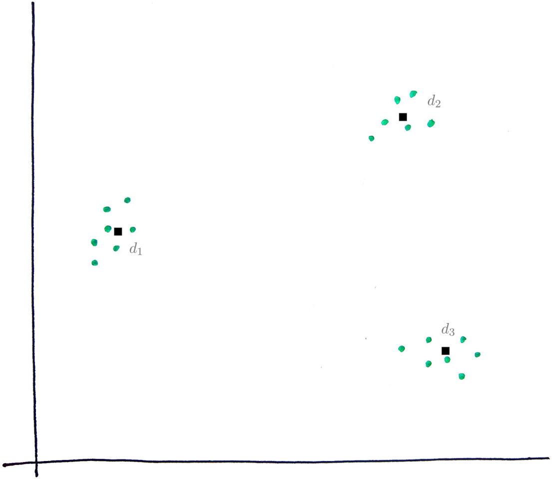


Figure 1: Given data points $y_1, \ldots, y_N \in \mathbb{R}^d$, find *atoms* $d_1, \ldots, d_k \in \mathbb{R}^d$.

k-means problem, or vector quantization

Find **signal-atoms** $d_1, \ldots, d_k$ and **representations** $x_1, \ldots, x_N$ such that $x_i \in \{0,1\}^k$ and $\|x_i\|_0 = 1$ where:

$$\begin{bmatrix} | & & | \\ y_1 & \cdots & y_N \\ | & & | \end{bmatrix} \approx \begin{bmatrix} | & | & & | \\ d_1 & d_2 & \cdots & d_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} | & & | \\ | & & | \\ x_1 & \cdots & x_N \\ | & & | \\ | & & | \end{bmatrix}.$$

The objective is to minimize the **reconstruction error** subject to the constraints on $\mathbf{X}$:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2.$$

Review of k-means clustering

- Initialize cluster means $d_1, \ldots, d_k \in \mathbb{R}^d$
- Repeat until convergence criterion:
  - **Sparse coding:** set $x_i \leftarrow e_{\kappa^*}$ (standard basis element) where

$$\kappa^* = \underset{j \in [k]}{\arg\min} \|y_i - d_j\|_2^2.$$

  - **Dictionary update:** set d_j to be the mean:

$$d_j \leftarrow \frac{1}{|C_j|} \sum_{y \in C_j} y,$$

where $C_j = \{y_i : x_i = e_j\}$.

Relax: gain-shape vector quantization

Find **signal-atoms** $d_1, \ldots, d_k$ and **representations** $x_1, \ldots, x_N$ such that $x_i \in \mathbb{R}^k$ and $\|x_i\|_0 = 1$ where:

$$\begin{bmatrix} | & & | \\ y_1 & \cdots & y_N \\ | & & | \end{bmatrix} \approx \begin{bmatrix} | & | & & | \\ d_1 & d_2 & \cdots & d_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_N \\ | & & | \end{bmatrix}.$$

The objective is to minimize the **reconstruction error** subject to the constraints on $\mathbf{X}$:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2.$$

Relax: gain-shape vector quantization

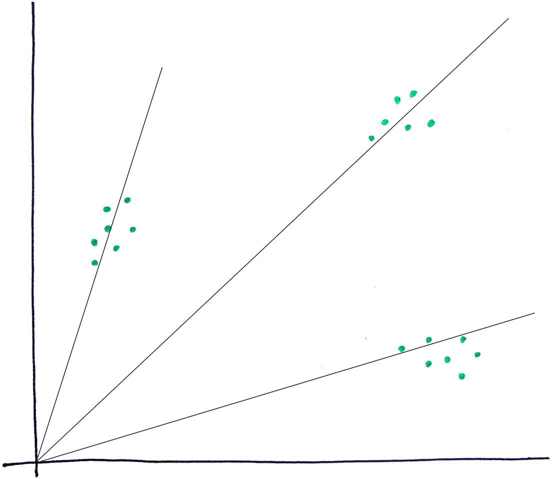


Figure 2: Each y_i is to the closest point on the candidate lines.

Generalize: dictionary learning s-sparse representations

Find **signal-atoms** $d_1, \dots, d_k$ and **representations** $x_1, \dots, x_N$ such that $x_i \in \mathbb{R}^k$ and $\|x_i\|_0 \leq s$ where:

$$\begin{bmatrix} | & & | \\ y_1 & \cdots & y_N \\ | & & | \end{bmatrix} \approx \begin{bmatrix} | & | & & | \\ d_1 & d_2 & \cdots & d_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_N \\ | & & | \end{bmatrix}.$$

The objective is to minimize the **reconstruction error** subject to the constraints on $\mathbf{X}$:

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2.$$

Summary of problems

- **Dictionary:** find k signal-atoms/dictionary elements.
- **Sparse-coding:** approximate data from points within T-dimensional objects generated by the k atoms.
  - $T = 0$ (k-means). Individual points $d_1, \ldots, d_k$.
  - $T = 1$ (gain-shape VQ). Lines through $d_1, \ldots, d_k$.
  - $T \in \mathbb{N}$ (dictionary learning). T-dimensional spaces:

$$\mathrm{span}(d_{i_1}, \ldots, d_{i_T}).$$

Geometric picture

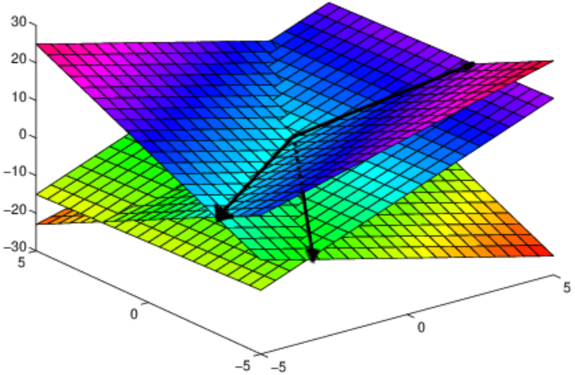


Figure 3: The set of k atoms generate a space $X \subset \mathbb{R}^d$ that is a union of T-manifolds. Each data point y is projected onto X. Which set of atoms minimizes reconstruction error? [image]

Goal of paper

Can we **generalize the k-means algorithm** to the more general problem of dictionary learning s-sparse representations?

Dictionary learning s-sparse representations

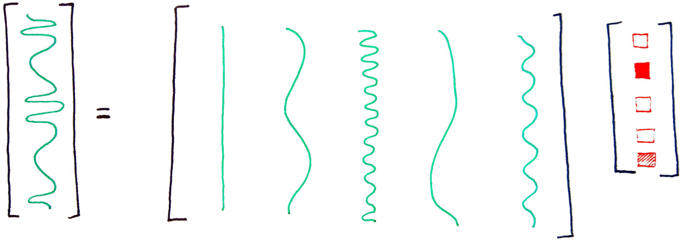


Figure 4: The signal y is a sparse linear combination of atoms $d_1, \ldots, d_k$.

k-SVD

Sparse coding and dictionary update

- **Sparse coding:** given a dictionary $\mathbf{D}$, find sparse representations $\mathbf{X}$ such that

$$\mathbf{Y} \approx \mathbf{DX}.$$

  - e.g. in k-means, $y \mapsto \arg\min \|d_i - y\|_2^2$.
- **Dictionary update:** let $C_j \subset \{y_1, \ldots, y_N\}$ be the set of data points whose representation makes use of d_j,

$$C_j = \{y_i : x_i(j) \neq 0\}.$$

  Can we update d_j and x_i to obtain better fit?
  - e.g. in k-means, $d_j \leftarrow \mathrm{mean}(C_j)$.

Geometric picture of k-SVD

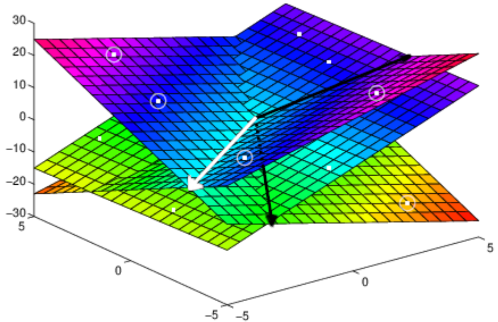


Figure 5: **Sparse coding:** data points are associated to s dictionary atoms. If d_j is the white arrow, then the circled data points are contained in C_j. **Dictionary update:** Jiggling d_j around also jiggles the planes: move to minimize average distance to C_j.

Sparse coding: pursuit algorithms

Many algorithms exist to perform sparse coding:

- Matching pursuit (MP)
- Orthogonal matching pursuit (OMP)
- Basis pursuit (BP)
- Focal underdetermined system solver (FOCUSS)

Techniques can be greedy, convex/non-convex relaxations, etc.

Dictionary update prelude: rank of matrix

Let $\mathbf{M} \in \mathbb{R}^{n \times m}$. The **rank** of $\mathbf{M}$ is the minimal r such that there exists $u_1, \ldots, u_r \in \mathbb{R}^n$ and $v_1, \ldots, v_r \in \mathbb{R}^m$ such that:

$$\mathbf{M} = \sum_{i=1}^{r} u_i v_i^\top.$$

Dictionary update prelude: low-rank approximation

The following says that the best rank-k approximation of a matrix correspond to its top-k singular value decomposition:

Theorem (Eckart-Young)

Let $\mathbf{M} \in \mathbb{R}^{n \times m}$ with singular value decomposition $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

- *Let $\mathbf{\Sigma}_k$ be the submatrix with the top-k singular values.*
- *Let $\mathbf{U}_k$ and $\mathbf{V}_k$ be the corresponding singular vectors.*

Then, for any matrix $\mathbf{N}$ where $\mathrm{rank}(\mathbf{N}) \leq k$,

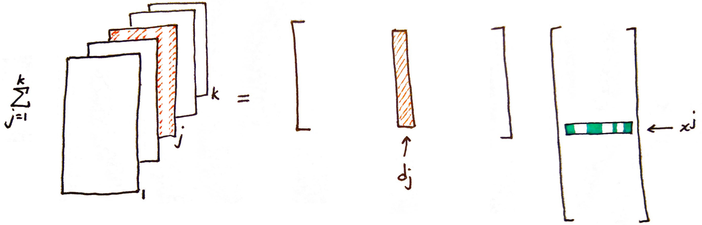
$$\|\mathbf{M} - \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top\|_F \leq \|\mathbf{M} - \mathbf{N}\|_F.$$

Dictionary update

Notice that $\mathbf{DX}$ is a rank-k approximation of $\mathbf{Y}$,

$$\mathbf{DX} = \sum_{j=1}^{k} d_j x^j,$$

where x^j is the jth row of $\mathbf{X}$.



Dictionary update

- Without sparsity constraint, set $\mathbf{D} = \mathbf{U}_k$ and $\mathbf{X} = \mathbf{\Sigma}_k \mathbf{V}_k^\top$.
  - **Idea:** when there is a sparsity constraint, let's fix all of $\mathbf{D}$ except the jth column and all of $\mathbf{X}$ except for the jth row. Now, apply SVD.

Dictionary update

Let $\mathbf{E}_j$ be the reconstruction error without the jth dictionary:

$$\mathbf{E}_j = \mathbf{Y} - \sum_{\kappa \neq j} d_\kappa x^\kappa.$$

- We want to reduce the reconstruction error for those $y_i \in C_j$.
  - Let $\mathbf{\Pi}_j \in \{0,1\}^{N \times |C_j|}$ select the data points in C_j.
  - Obtain SVD of $\mathbf{E}_j \mathbf{\Pi}_j = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.
  - Update $d_j \leftarrow \mathbf{U}_1$ and $x^j \mathbf{\Pi}_j \leftarrow \mathbf{V}_1^\top$

Dictionary update

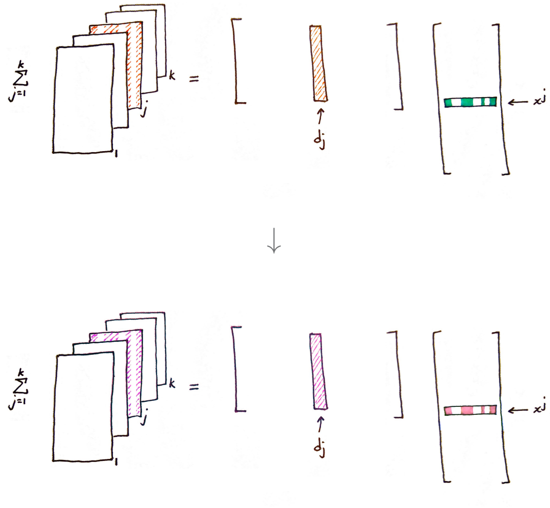


Figure 6: Updates to d_j and the nonzero coordinates of x^j.

k-SVD algorithm

- Initialize dictionary $d_1, \dots, d_k \in \mathbb{R}^d$
- Repeat until convergence criterion:
  - **Sparse coding:** using any pursuit algorithm for the problem:

    $$\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_2^2,$$

    in order to set representations x_i.
  - **Dictionary update:** for each atom d_j:
    - Apply SVD decomposition to $\mathbf{E}_j \mathbf{\Pi}_j = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$
    - Set $d_j \leftarrow \mathbf{U}_1$ and $x^j \mathbf{\Pi}_j \leftarrow \mathbf{\Sigma}_1 \mathbf{V}_1^\top$.

Convergence of k-SVD

Suppose we have pursuit algorithm that solves sparse coding perfectly.[1]

- During dictionary update, the MSE decreases monotonically:

$$\|\mathbf{Y} - \mathbf{DX}\|_2^2 = \left\| \left(\mathbf{Y} - \sum_{\kappa \neq j} d_\kappa x^\kappa \right) - d_j x^j \right\|_2^2 = \|\mathbf{E}_j - d_j x^j\|_2^2.$$

  Thus, convergence to local minimum is guaranteed.

[1] When sparsity $s \ll n$, then pursuit algorithms like OMP, FOCUSS, BP are known to perform well.

A final remark on k-SVD

The dictionary update step does not change which coordinates of x_j are nonzero—only the sparse coding step does this.

- Without sparse coding, we get trapped in a local minimum.

References

[AEB2006] M. Aharon, M. Elad, and A. Bruckstein. “K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation.” *IEE Transactions on Signal Processing*, 2006.