Learning Scale and Shift-Invariant Dictionary for Sparse Representation

Toshimitsu Aritake, Noboru Murata

Waseda University, Japan

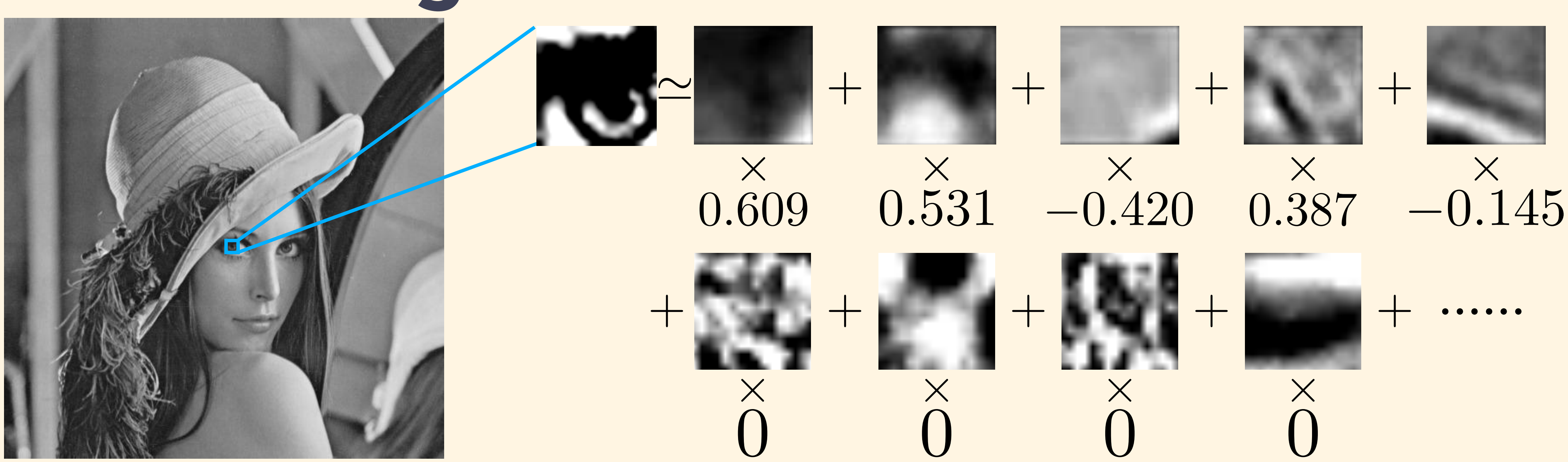
11/09/2019

LOD2019

Sparse Coding

Method to represent a given signal with a small number of features selected from a given large number of candidates

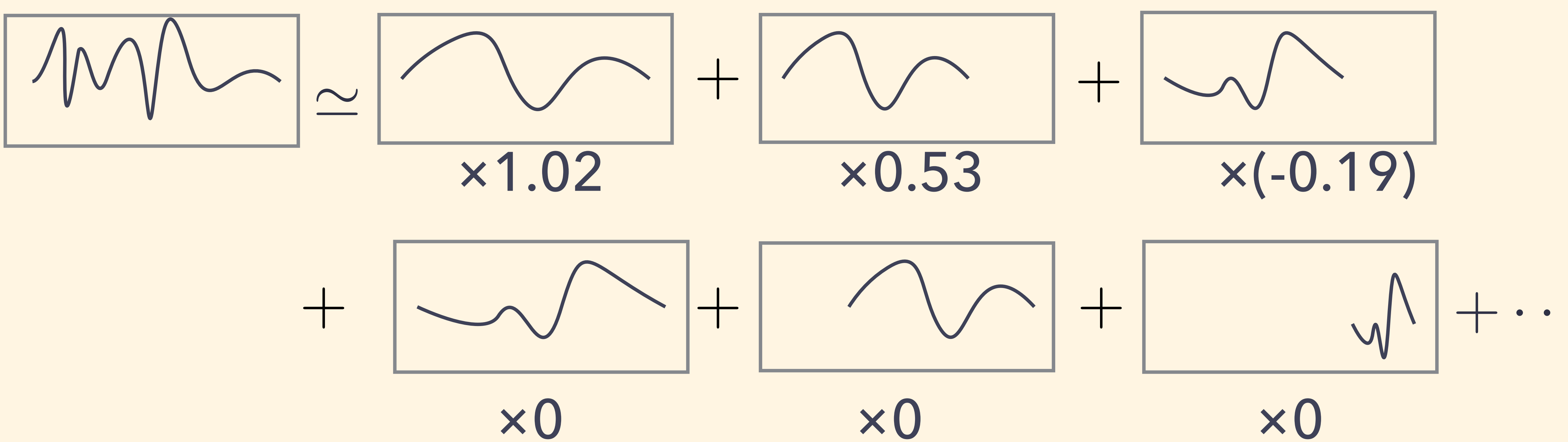
Natural Image



Sparse Coding

Method to represent a given signal with a small number of features selected from a given large number of candidates

Time series



Sparse Coding

- signal (observation) : $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$
- atoms (features) : $\boldsymbol{d}_k = (d_{k1}, d_{k2}, \ldots, d_{kn})^\top \in \mathbb{R}^n$ $(k = 1, 2, \ldots, m)$
- dictionary : $\boldsymbol{D} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m) \in \mathbb{R}^{n \times m}$ $(n < m)$
- coefficient vector : $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^\top \in \mathbb{R}^m$

$$\begin{bmatrix} \\ \boldsymbol{y} \\ \\ \end{bmatrix} \simeq \underbrace{\begin{bmatrix} \\ \boldsymbol{d}_1 & \boldsymbol{d}_2 & \boldsymbol{d}_3 & \boldsymbol{d}_4 & \cdots & \boldsymbol{d}_{m-1} & \boldsymbol{d}_m \\ \\ \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\begin{bmatrix} 0.609 \\ 0 \\ 0 \\ 0.531 \\ \vdots \\ -0.145 \\ 0 \end{bmatrix}}_{\boldsymbol{x}}$$

Sparse Coding

Given a signal $\boldsymbol{y} \in \mathbb{R}^n$ and a dictionary $\boldsymbol{D} \in \mathbb{R}^{n \times m}$
find a sparse coefficient vector $\boldsymbol{x}$

Lasso

$$\underset{\boldsymbol{x}}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$

minimize the approximation error and
the sparsity regularizer

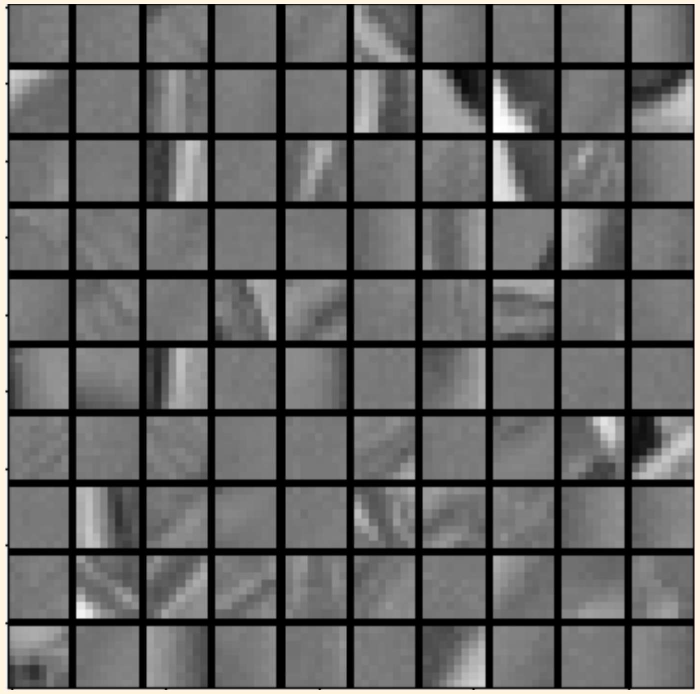
Choice of a Dictionary

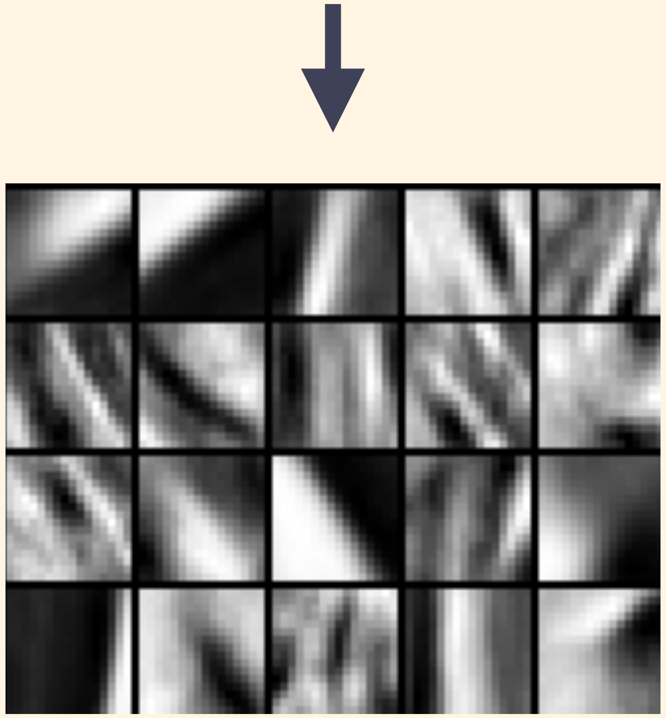
$$\underset{\boldsymbol{x}}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$

- The choice of a dictionary $\boldsymbol{D}$ significantly affects the quality of overall signal processing

- How to choose a dictionary $\boldsymbol{D}$ to represent data by sparse coding ?

Dictionary Learning



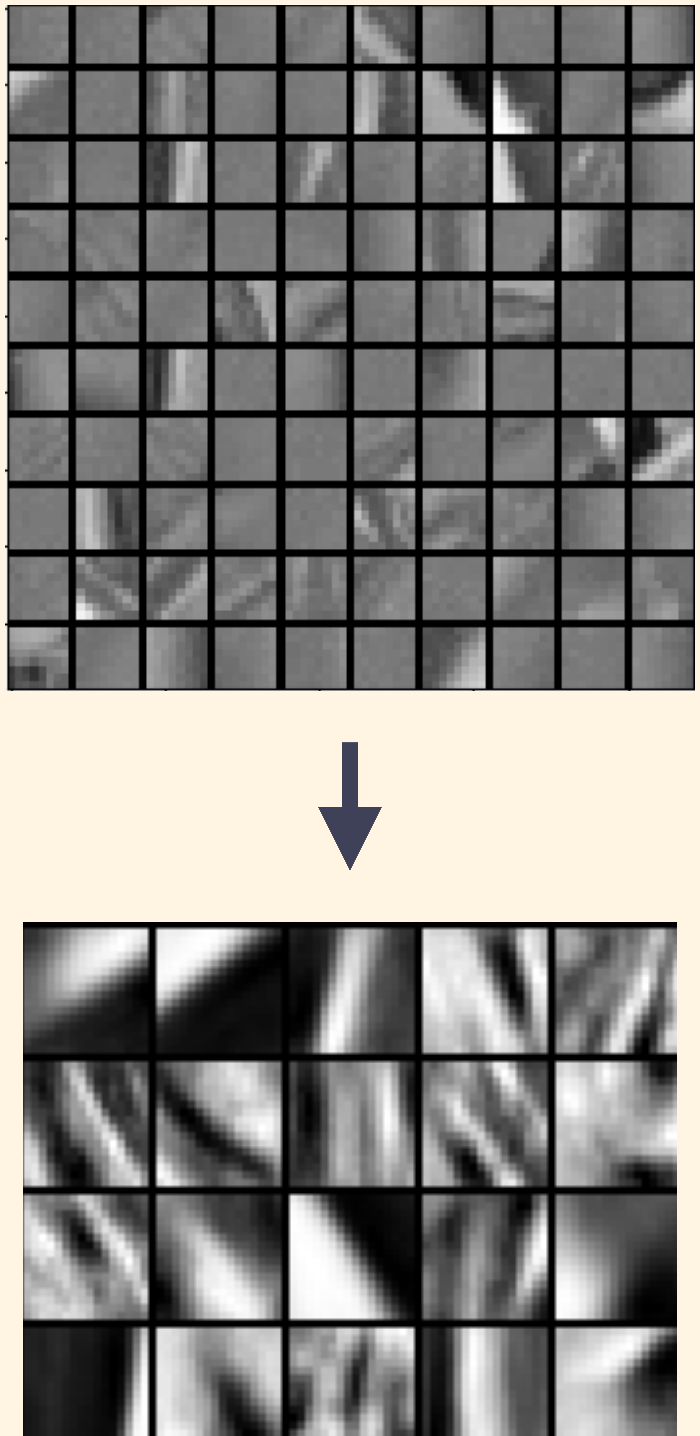


- Learn adaptive features from a set of signals $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$

$$\underset{\{\boldsymbol{x}_j\}_{j=1}^N, \boldsymbol{D}}{\text{minimize}} \sum_{j=1}^{N} \left(\|\boldsymbol{y}_j - \boldsymbol{D}\boldsymbol{x}_j\|_2^2 + \lambda \|\boldsymbol{x}_j\|_1 \right)$$

- This problem is not jointly convex with respect to both $\{\boldsymbol{x}_j\}_{j=1}^N$ and $\boldsymbol{D}$
- Alternating minimization is used to solve the above problem

Dictionary Learning



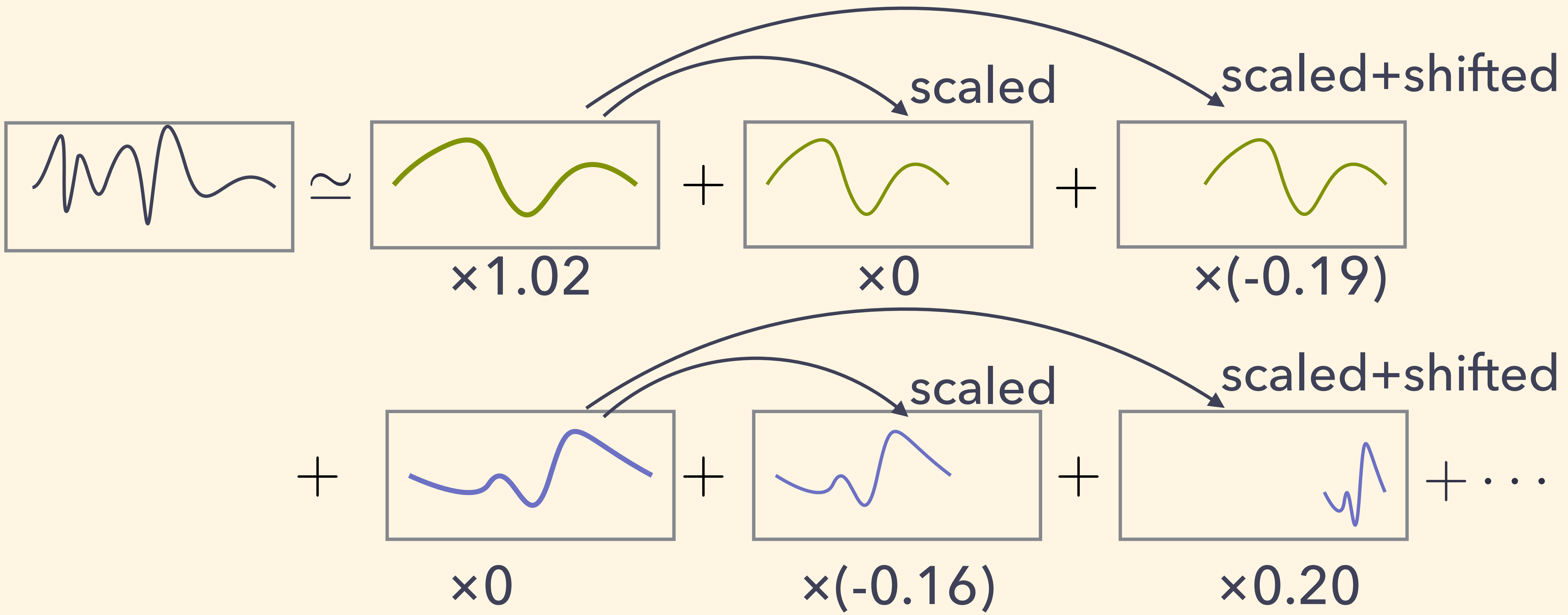
- Learn adaptive features from a set of signals $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$

$$\underset{\{\boldsymbol{x}_j\}_{j=1}^{N}, \boldsymbol{D}}{\text{minimize}} \sum_{j=1}^{N} \left(\|\boldsymbol{y}_j - \boldsymbol{D}\boldsymbol{x}_j\|_2^2 + \lambda \|\boldsymbol{x}_j\|_1 \right)$$

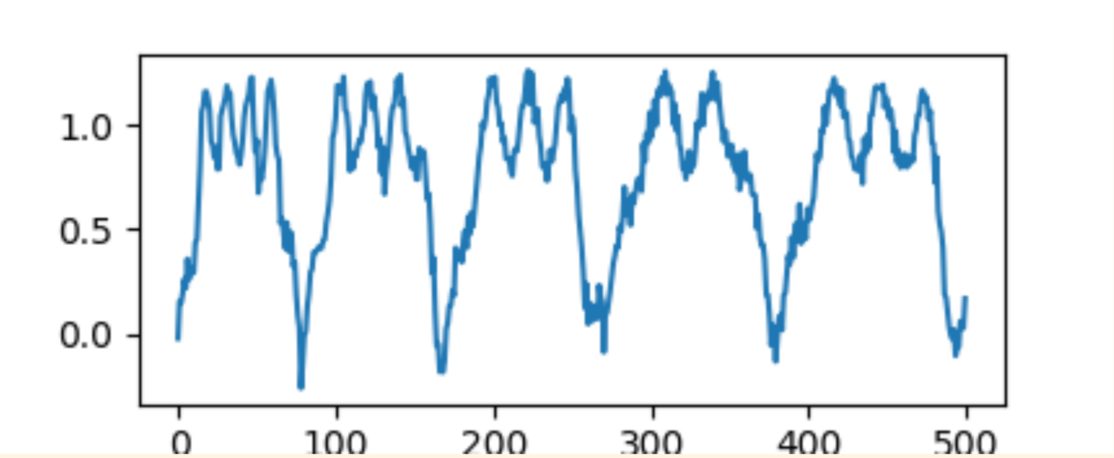
- coefficient vectors are independently optimized for each signal
- A dictionary is optimized as common features for a set of signals

Assumption

Similar features appear at various scales and locations of the observed signals



Assumption

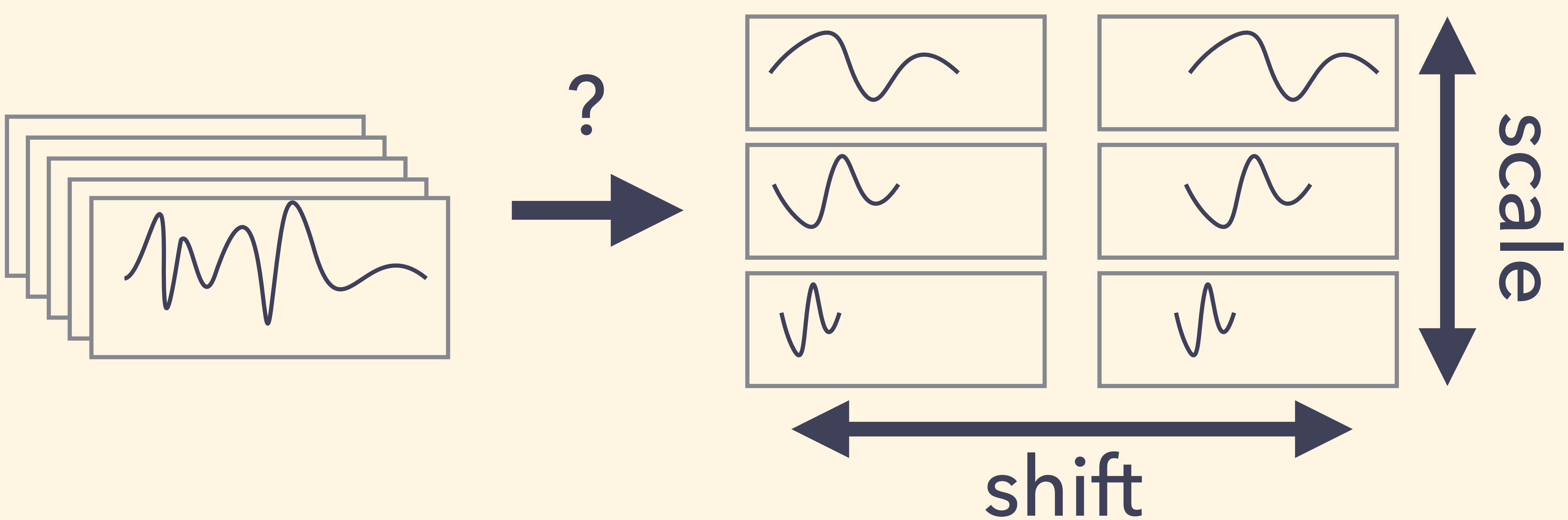


- Natural Image
  - Common to assume there are multi-scale features in images
  - It is reasonable that an object of different size have the same features of different scale
- Time Series Data
  - Assume the signals have similar temporal patterns at various scales and locations

Problem

Can we learn atoms and their scaled or shifted atoms from a set of signals by dictionary learning?

- Learned atoms are essential features to characterize a set of signals



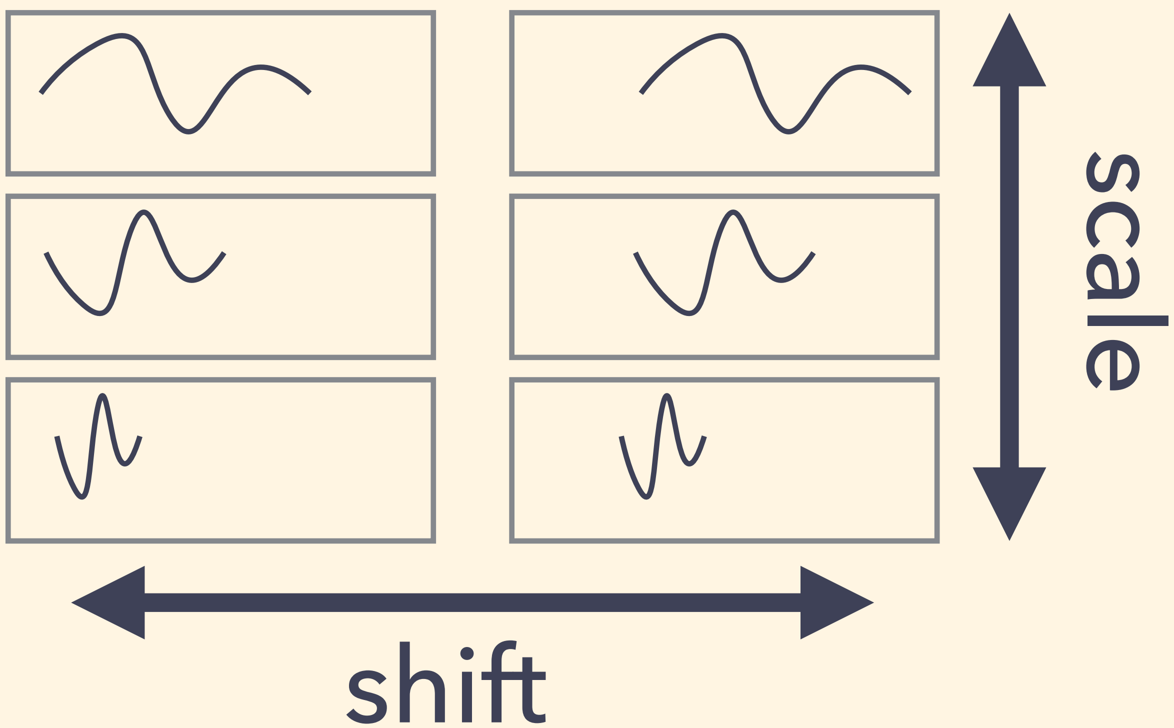
Problem

Can we learn atoms and their scaled or shifted atoms from a set of signals by dictionary learning?
→ **NO**

$$\underset{\{\boldsymbol{x}_j\}_{j=1}^{N}, \boldsymbol{D}}{\text{minimize}} \sum_{j=1}^{N} (\|\boldsymbol{y}_j - \boldsymbol{D}\boldsymbol{x}_j\|_2^2 + \lambda \|\boldsymbol{x}_j\|_1)$$

- In general, a dictionary model does not consider the relationship between atoms

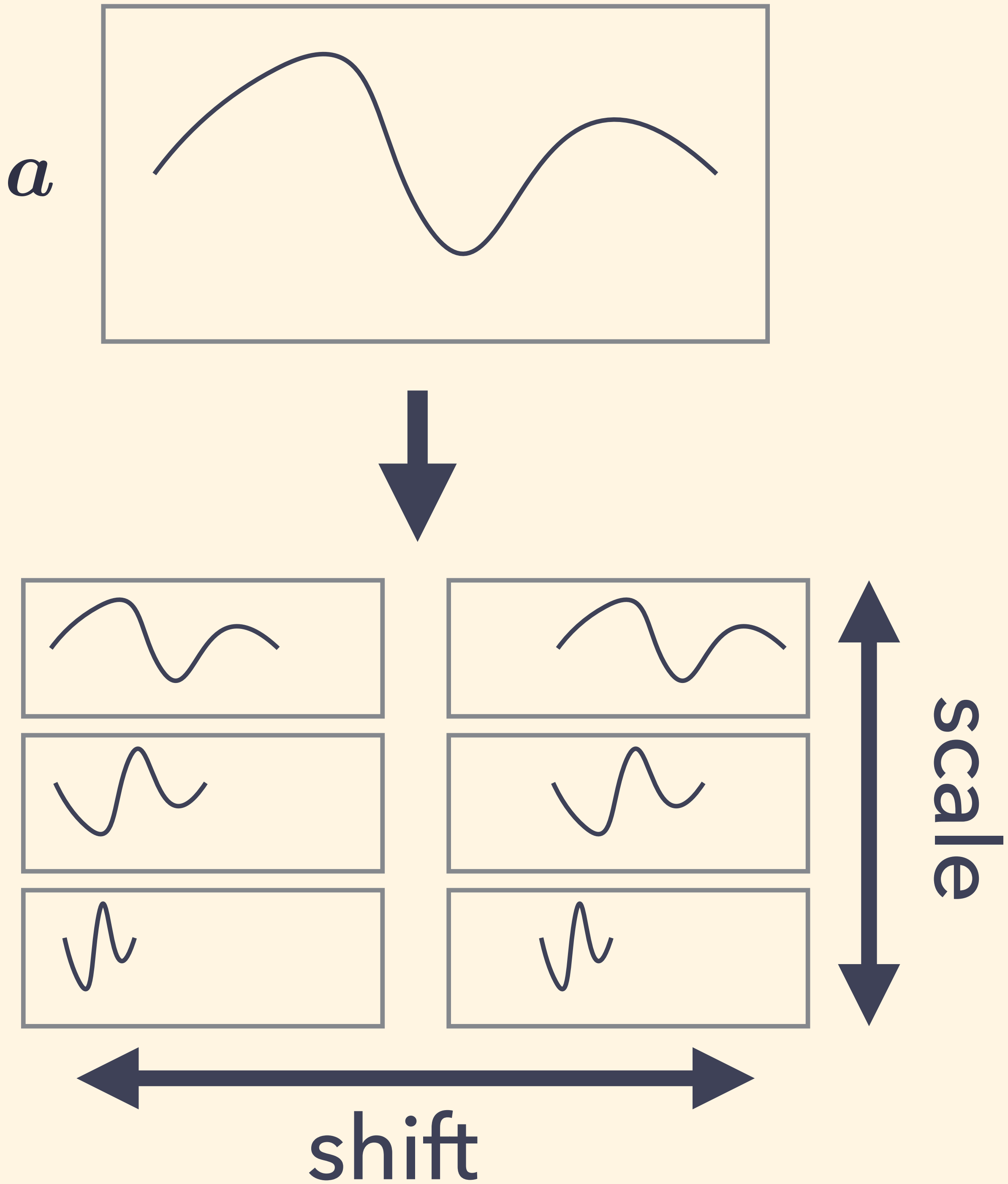
Our Contribution



We propose:

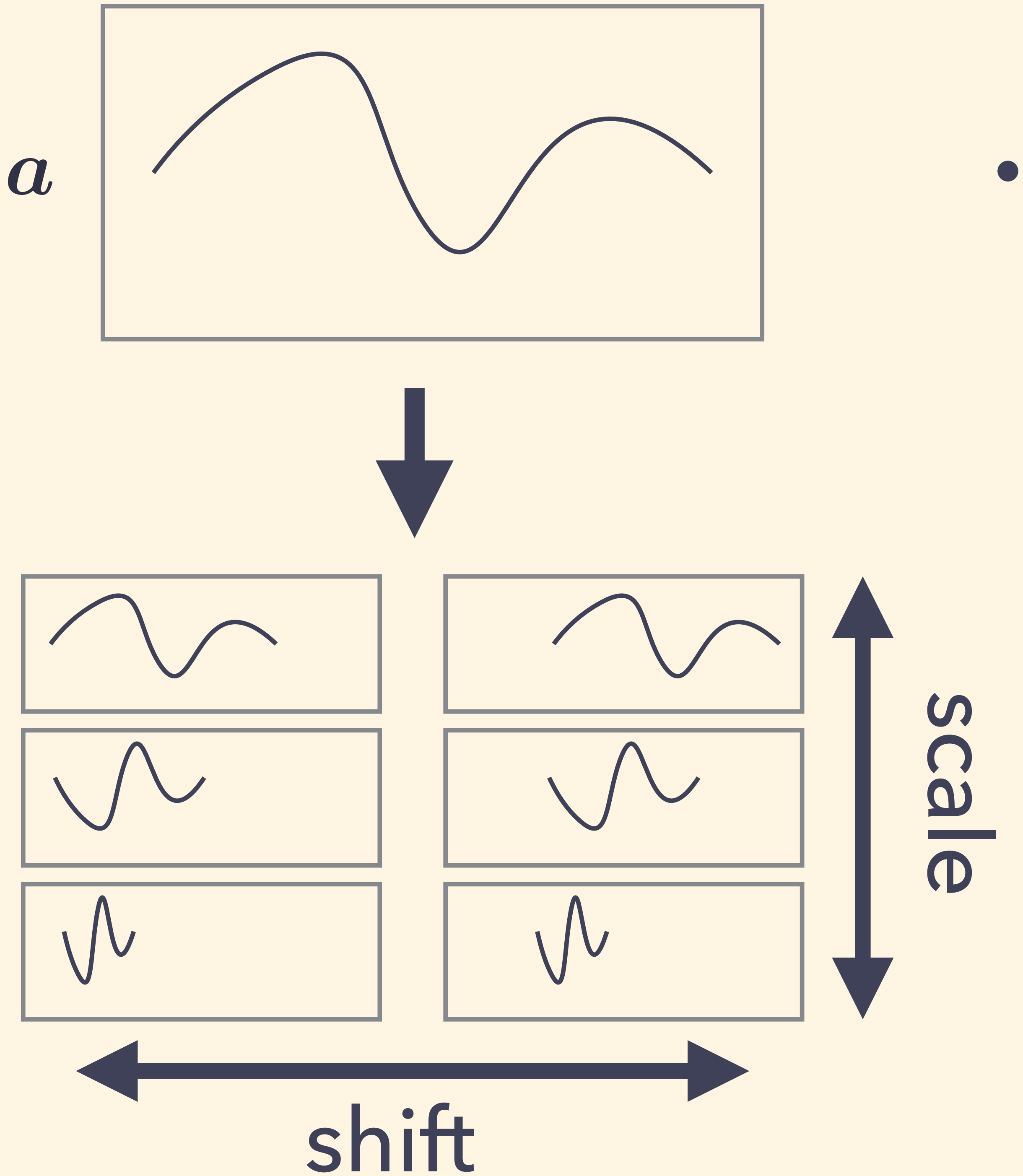
- a dictionary model which considers the scale and shift structure
- an algorithm to learn a structured dictionary from a set of signals

Introducing Shift and Scaling Structure Into a Dictionary



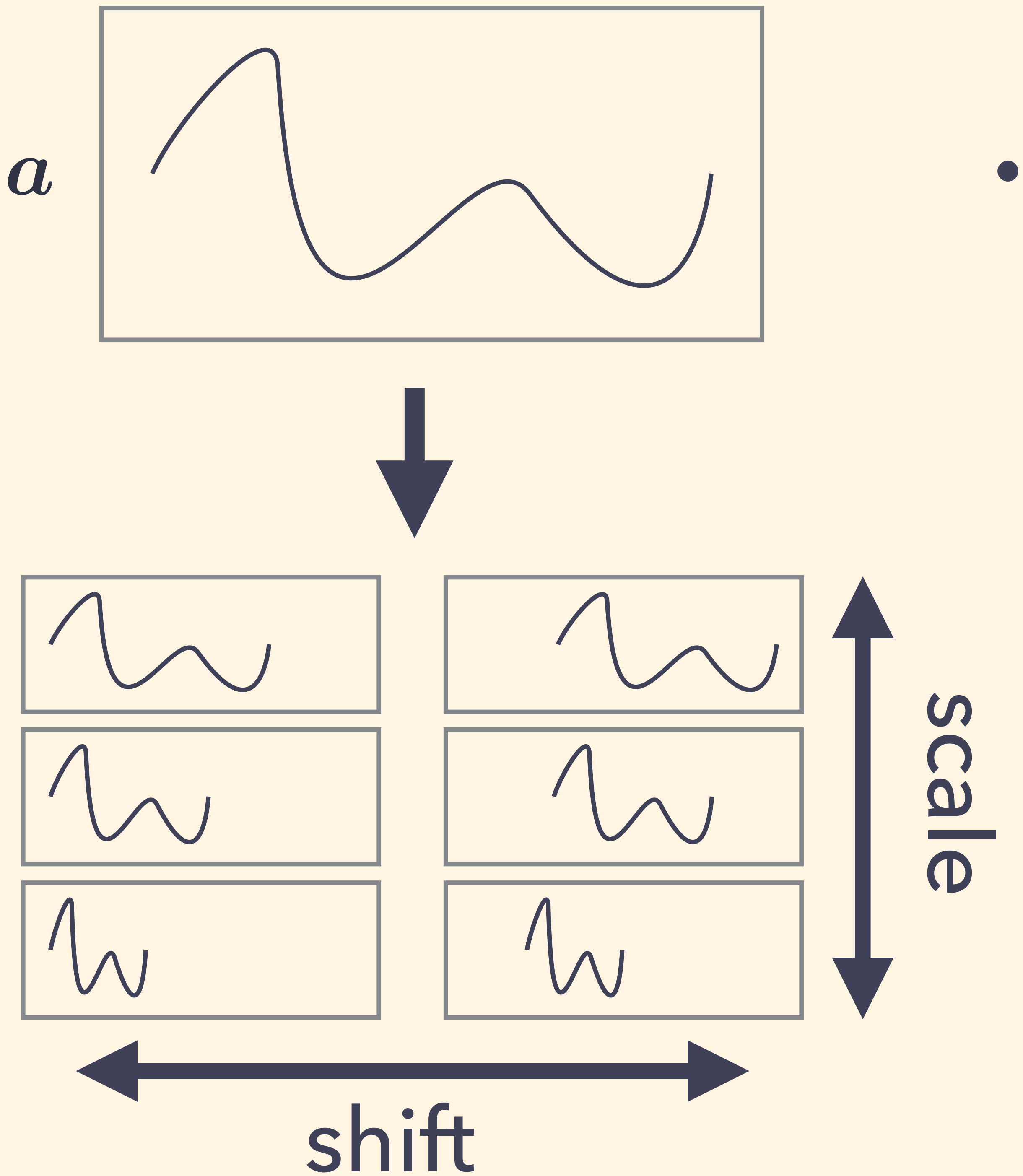
- We assume all atoms of a dictionary is generated from a single vector $\boldsymbol{a} \in \mathbb{R}^n$ which we call **ancestor**
- Atoms are generated by scaling or shifting an ancestor
- <u>An ancestor is an essential feature</u> which generates other features

Introducing Shift and Scaling Structure Into a Dictionary



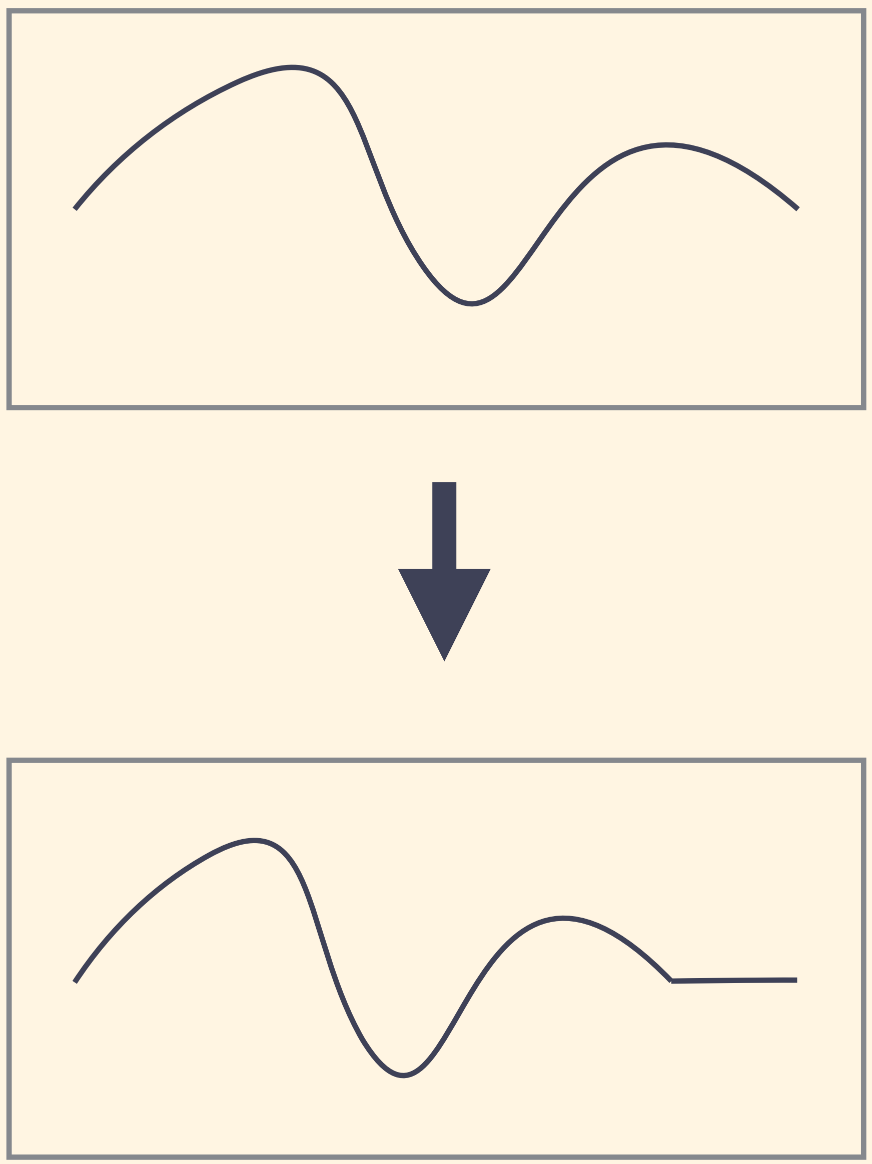
- We can use multiple ancestors $\boldsymbol{a}_l \ (l = 1, 2, \ldots, L)$ to generate a dictionary

Introducing Shift and Scaling Structure Into a Dictionary



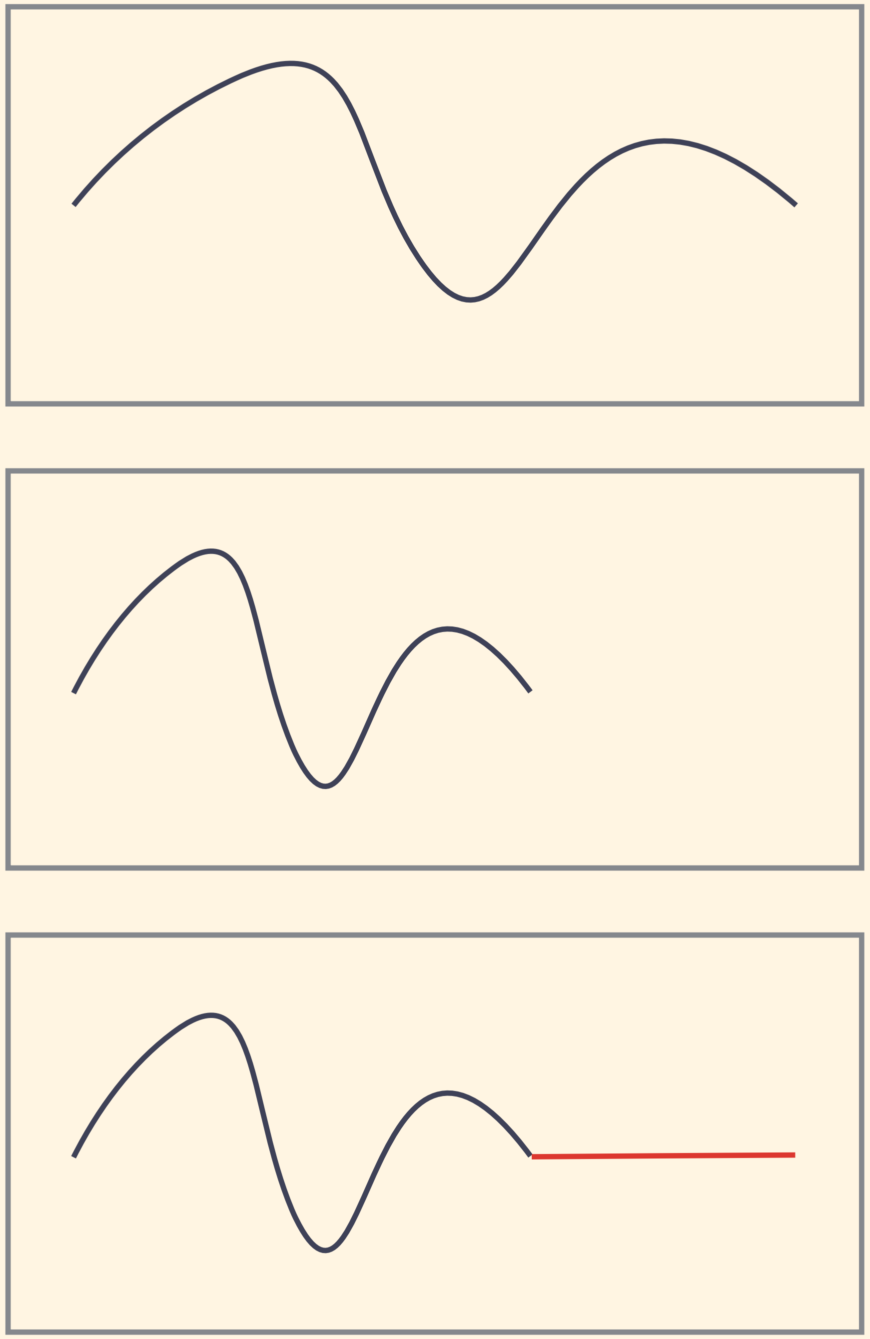
- We can use multiple ancestors $\boldsymbol{a}_l \; (l = 1, 2, \ldots, L)$ to generate a dictionary

Scaling Operation



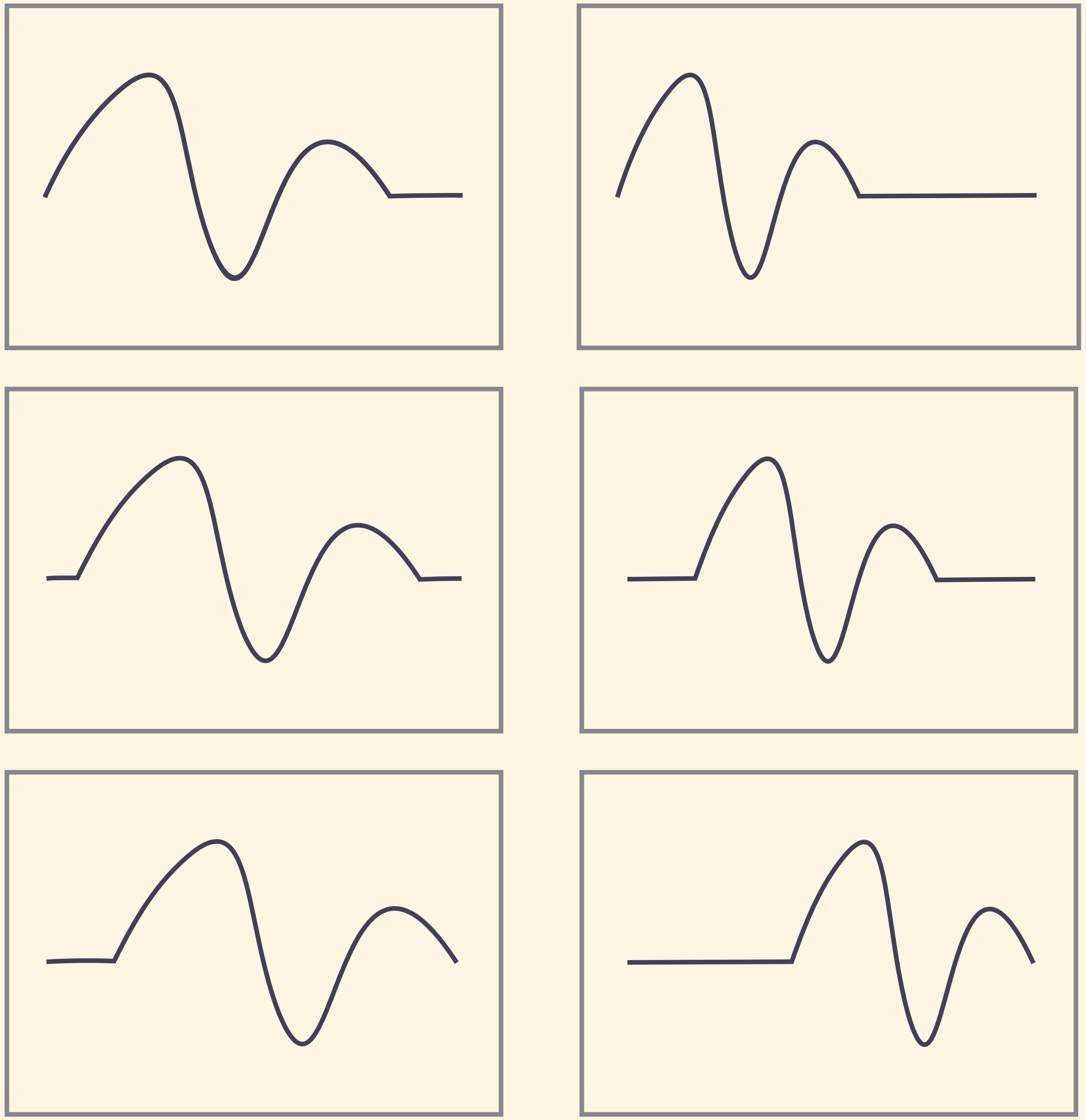
- An ancestor itself is used as an atom of a dictionary
- The scale of the ancestor is changed by scaling operation
- All scaled atoms need to have the same dimensionality to compose a dictionary

Scaling Operation



- We use linear resize operation to shrink the size of the ancestor
- We use zero-padding to keep the dimensionality of resized atoms
- Whole scaling operation (resize and zero-padding) is a linear operation

Shift Operation



- Atoms are shifted by changing the position of zero elements of zero-padded atoms
- Shift operations are also linear operations

Atom generating matrix $F_{p,q}$

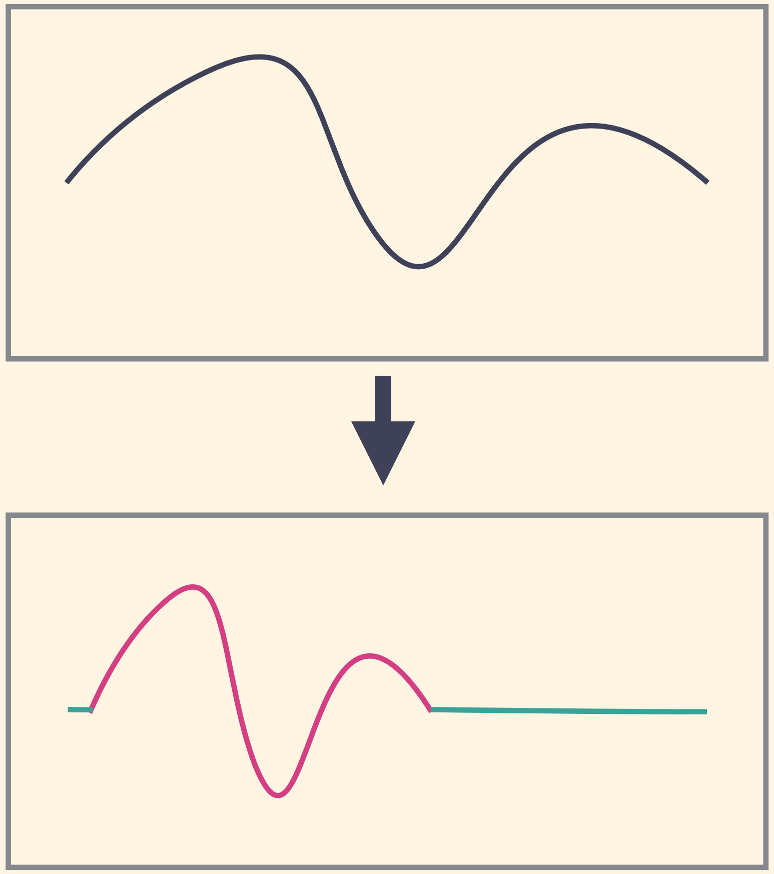
- An atom generating operation is a unique operation composed of a scaling and a shift operation

- An atom generating operator can be written as a matrix as $\boldsymbol{F}_{p,q}$
  p : index of scaling, q : index of shift

Atom generating matrix $\boldsymbol{F}_{p,q}$

- Example of $\boldsymbol{F}_{p,q}$
  - Resize by taking average of two adjacent elements
  - Shift one element by zero-padding
  - zero-pad rest of the elements

$$\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & & & & & & & \\ 0 & 0 & 1/2 & 1/2 & & & & 0 & \\ & & & & \ddots & & & & \\ & & & & & 1/2 & 1/2 & & \\ & 0 & & & & & & 1/2 & 1/2 \\ & & & & 0 & & & & \end{pmatrix}$$



Dictionary Generated From an Ancestor

- Each atom of dictionary is generated from an ancestor $\boldsymbol{a}$ by multiplying atom generating matrix $\boldsymbol{F}_{p,q}$ as $\boldsymbol{F}_{pq}\boldsymbol{a}$
- A dictionary generated from an ancestor is

$$\boldsymbol{D}(\boldsymbol{a}) = \begin{bmatrix} \boldsymbol{F}_{0,0}\boldsymbol{a} & \boldsymbol{F}_{1,0}\boldsymbol{a} & \boldsymbol{F}_{1,1}\boldsymbol{a} & \cdots & \boldsymbol{F}_{P,Q}\boldsymbol{a} \end{bmatrix}$$

- Set of (p, q) is written by Λ

Dictionary Generated From Multiple Ancestors

- When we use multiple ancestors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L$ whole dictionary is generated by concatenating dictionaries $\boldsymbol{D}(\boldsymbol{a}_1), \boldsymbol{D}(\boldsymbol{a}_2), \ldots, \boldsymbol{D}(\boldsymbol{a}_L)$

$$\boldsymbol{D}(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L) = \left[\begin{array}{c|c|c|c} \boldsymbol{D}(\boldsymbol{a}_1) & \boldsymbol{D}(\boldsymbol{a}_2) & \cdots & \boldsymbol{D}(\boldsymbol{a}_L) \end{array} \right]$$

Learning Ancestors

- Problem is to learn a dictionary which has scale and shift structure

Can we learn atoms and their scaled or shifted atoms from a set of signals by dictionary learning?
→ **NO**

Learning Ancestors

- Problem is to learn a dictionary which has scale and shift structure

Can we learn ancestors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_L$
from a set of signals ?
→ **YES**

- Ancestors are essential features which generate other features

Ancestral Atom Learning (AAL)

$$\underset{\{\boldsymbol{x}_j\}_{j=1}^N, \{\boldsymbol{a}_l\}_{l=1}^L}{\text{minimize}} \sum_{j=1}^{N} \left(\|\boldsymbol{y}_j - \sum_{l=1}^{L} \sum_{(p,q)\in\Lambda} \boldsymbol{F}_{p,q} \boldsymbol{a}_l x_j^{pql} \|_2^2 + \lambda \|\boldsymbol{x}_j\|_1 \right)$$

- Find the sparse coefficient vectors $\boldsymbol{x}_j$ $(j = 1, 2, \ldots, N)$ and ancestors $\boldsymbol{a}_l$ $(l = 1, 2, \ldots, L)$ to sparsely approximate the signals $\boldsymbol{y}_j$ $(j = 1, 2, \ldots, N)$
- The problem is not jointly convex with respect to both $\{\boldsymbol{x}_j\}_{j=1}^N$ and $\{\boldsymbol{a}_l\}_{l=1}^L$
- Alternating minimization is used to solve this problem

Algorithm

Initialize ancestors $\boldsymbol{a}_1^{(0)}, \boldsymbol{a}_2^{(0)}, \ldots, \boldsymbol{a}_L^{(0)}$

for $t = 0$ to T

1. Sparse Coding (Lasso)

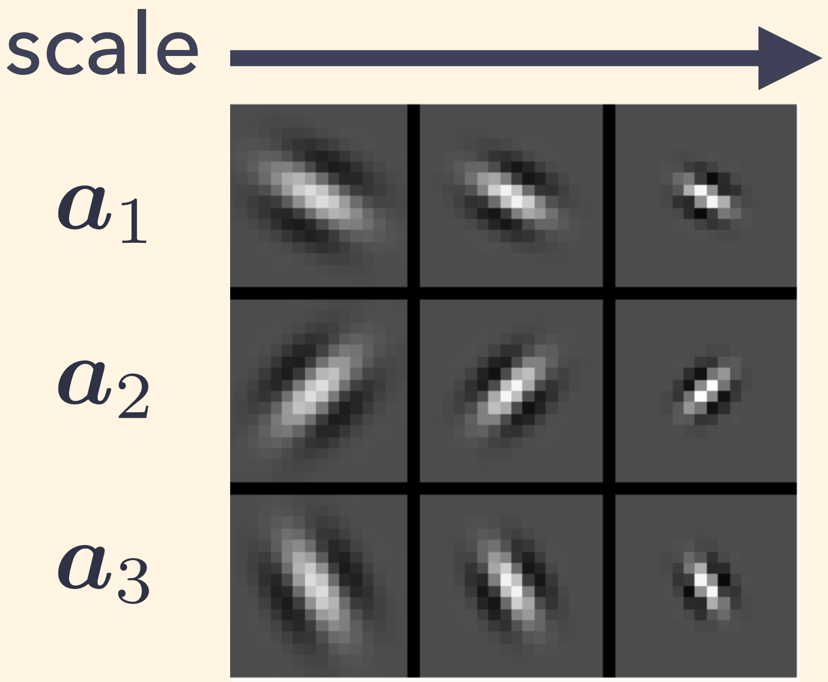
$$\boldsymbol{x}_j^{(t)} = \mathop{\arg\min}_{\boldsymbol{x}_j} \|\boldsymbol{y}_j - \boldsymbol{D}(\boldsymbol{a}_1^{(t)}, \ldots, \boldsymbol{a}_L^{(t)})\boldsymbol{x}_j\|_2^2 + \lambda\|\boldsymbol{x}_j\|_1 \quad (j = 1, \ldots N)$$

2. Ancestor update (Stochastic gradient descent)

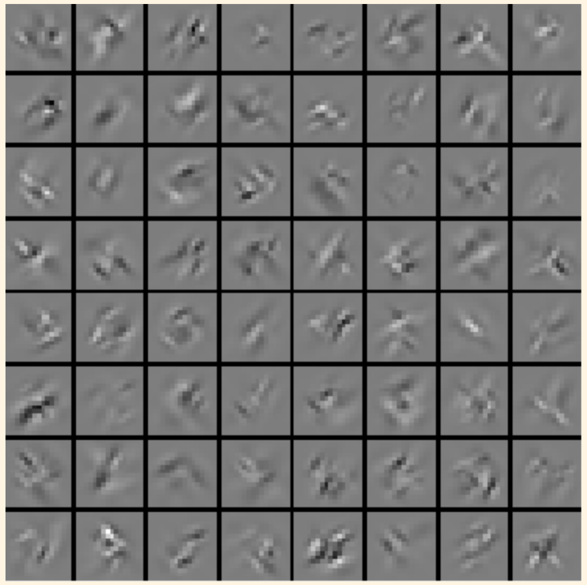
$$\boldsymbol{a}_1^{(t)}, \ldots, \boldsymbol{a}_L^{(t)} = \mathop{\arg\min}_{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L} \sum_{j=1}^{N} \|\boldsymbol{y}_j - \sum_{l=1}^{L} \sum_{(p,q)\in\Lambda} \boldsymbol{F}_{p,q}\boldsymbol{a}_l x_j^{pql(t)}\|_2^2$$

end loop

Experiment with artificial signals



ground-truth ancestors
(three scales)

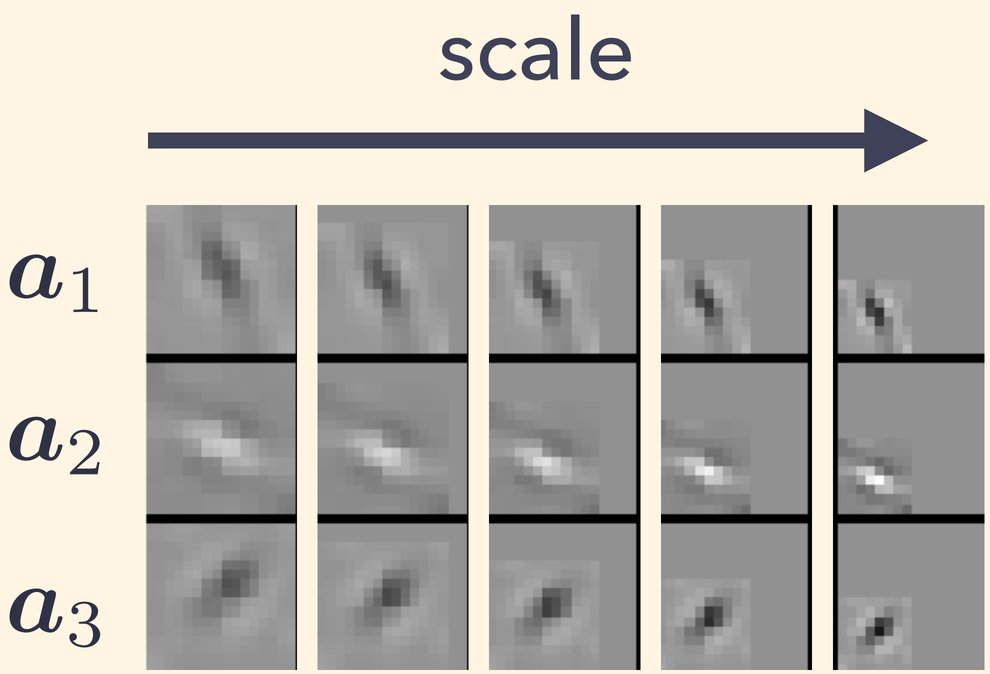


generated signals

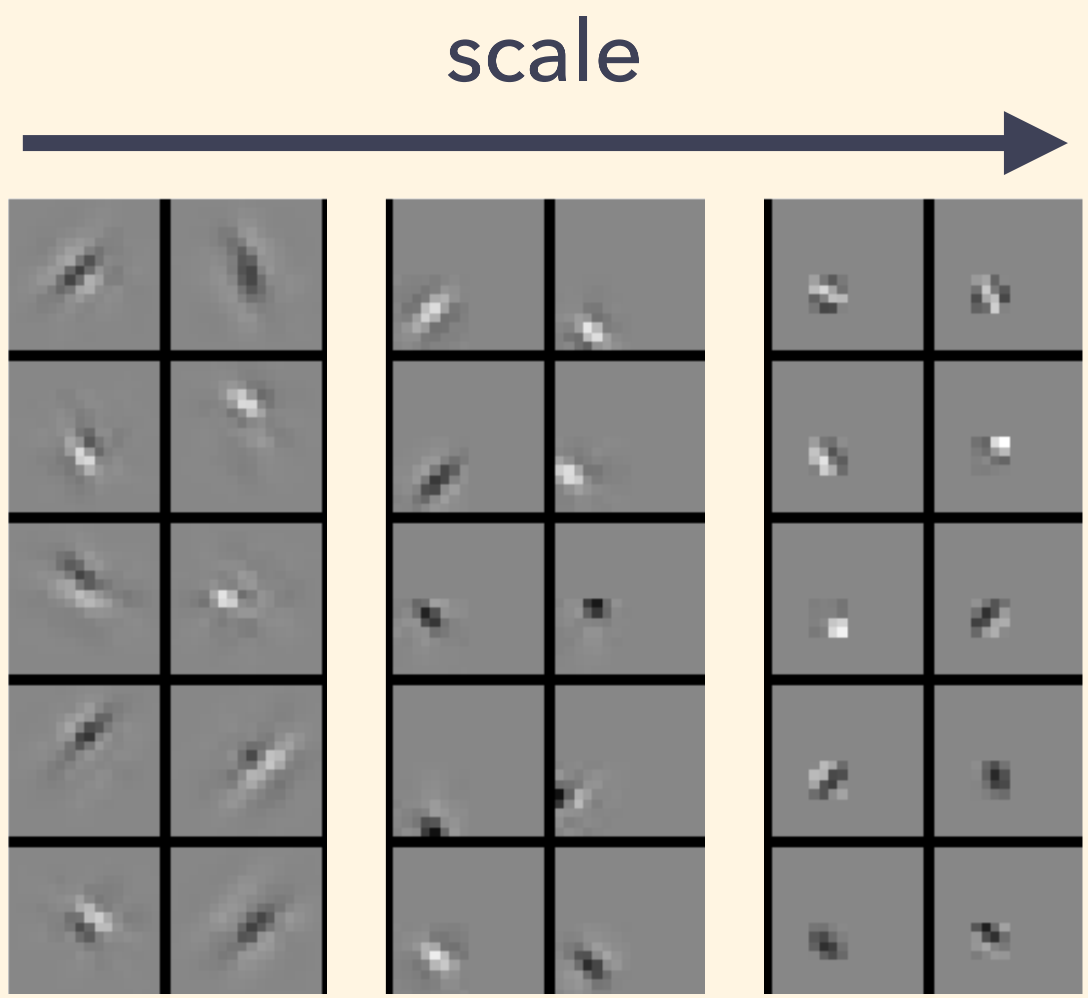
- We use 16x16 pixels 2D Gabor atoms as ground-truth ancestors
- Signals are generated by a linear combination of scaled or shifted ground-truth ancestors

- Generated signals can be approximated by three essential features and their variants
- Can we recover the ground-truth ancestors from signals ?

Results



Ancestral Atom Learning (AAL)

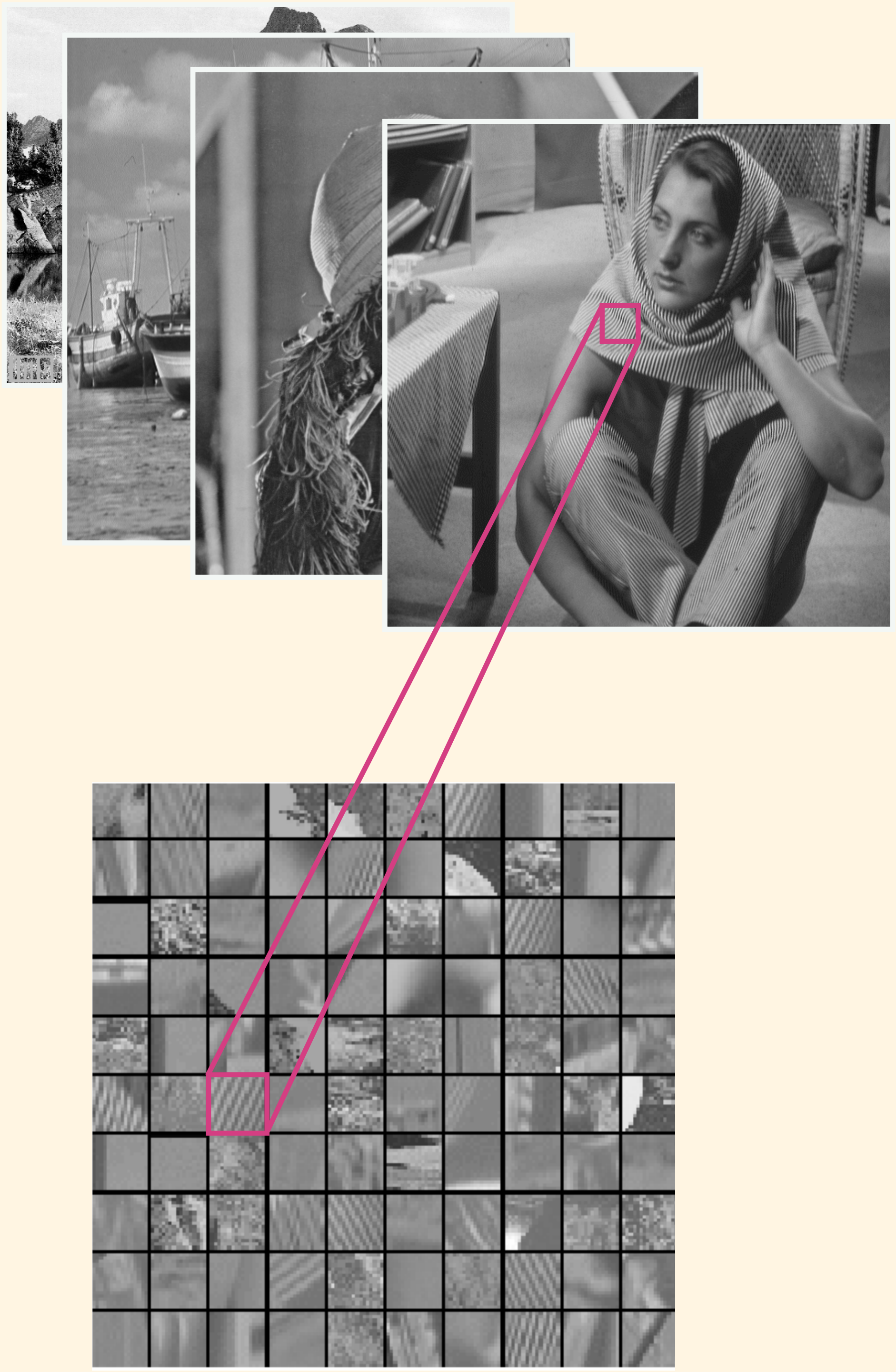


Multi-scale K-SVD

Results

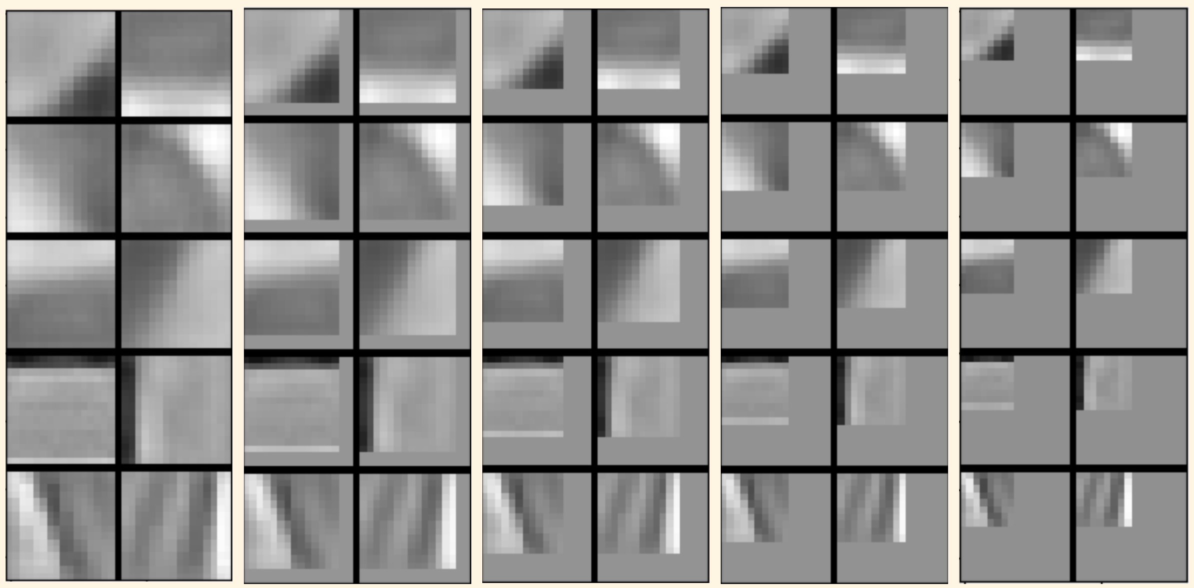
	Orientation	Scale	Reconstruction error
AAL	✓	✓	Slightly higher than Multi-scale K-SVD
Multi-scale K-SVD	can be different from ground-truth	Smaller than ground-truth	✓

Experiments with Natural Images

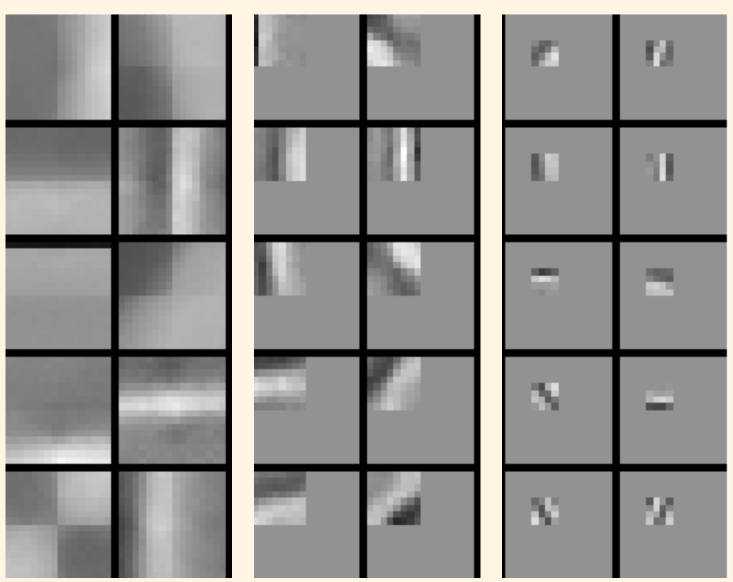


- We extract 16x16 patches from natural images and these patches are used to learn ancestors
- No ground-truth ancestors are known

Experiments with Natural Images



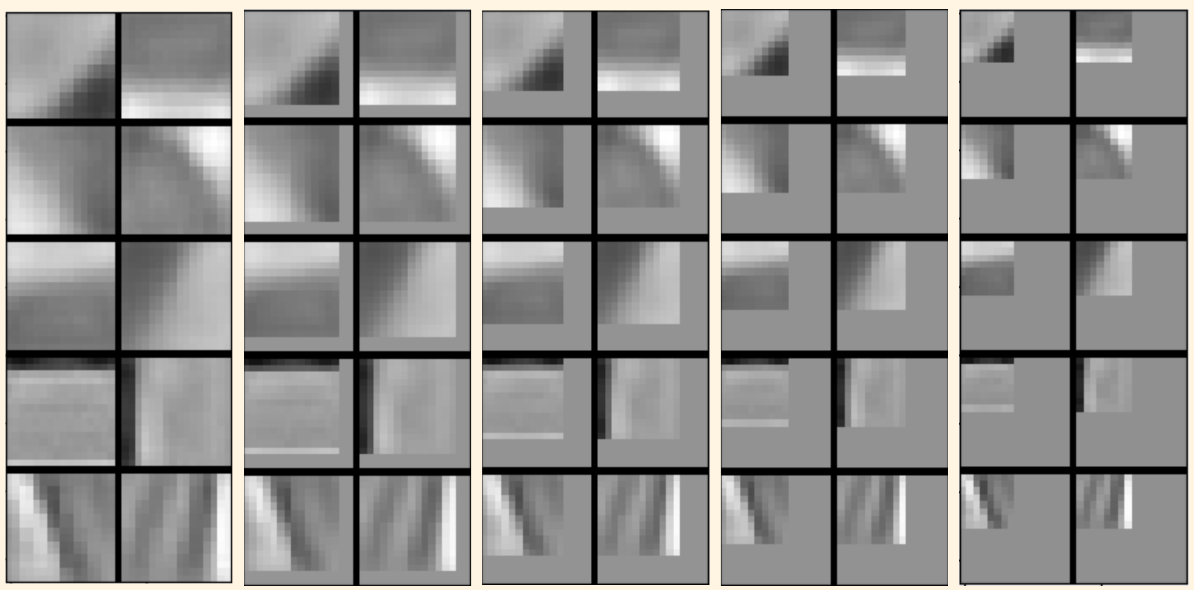
Ancestral Atom Learning



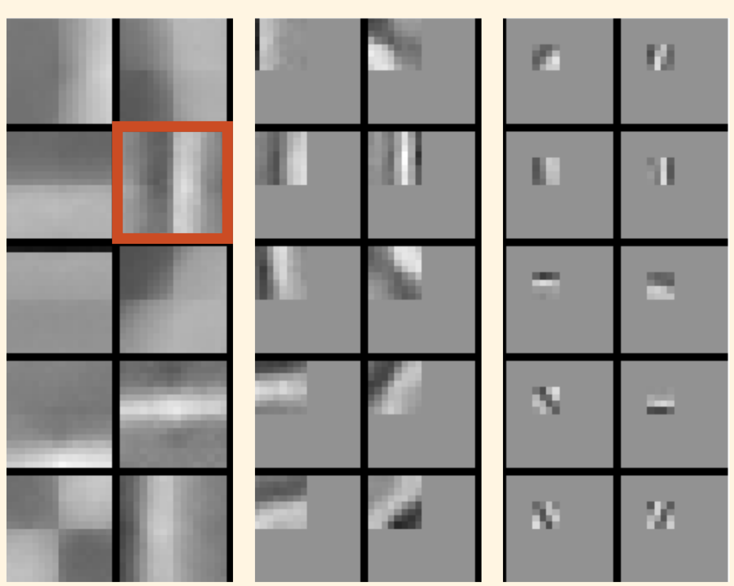
Multi-scale K-SVD

- Edge-like features and texture-like features are learned from signals
- Texture like features of multi-scale K-SVD are only learned at smaller scale
- Artifacts appear in the learned features by multi-scale K-SVD

Experiments with Natural Images



Ancestral Atom Learning



Multi-scale K-SVD

- Edge-like features and texture-like features are learned from signals
- Texture like features of multi-scale K-SVD are only learned at smaller scale
- Artifacts appear in the learned features by multi-scale K-SVD

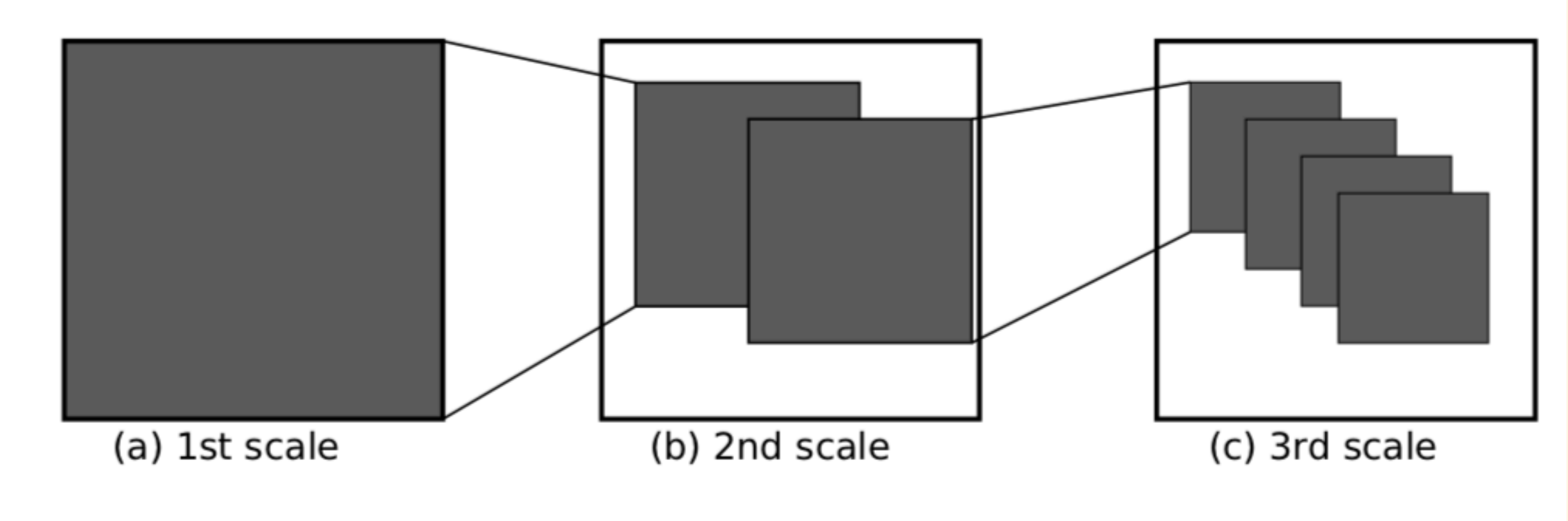
Summary

- We propose a model of a dictionary which have shift and scaling structure
- Shift and scaling structure are introduced by generating atoms from vectors called ancestors
- A simple gradient based algorithm was presented to learn ancestors from signals
- Our proposed method successfully learn features appear at various scales and locations

Appendix

Treating High Dimensional Signals

- We use 2D ancestor when the signal is 2D signal by vectorizing the signals and ancestors
- Scaling and shift is operated along each axis
- 3D or higher signal can be treated in the same way



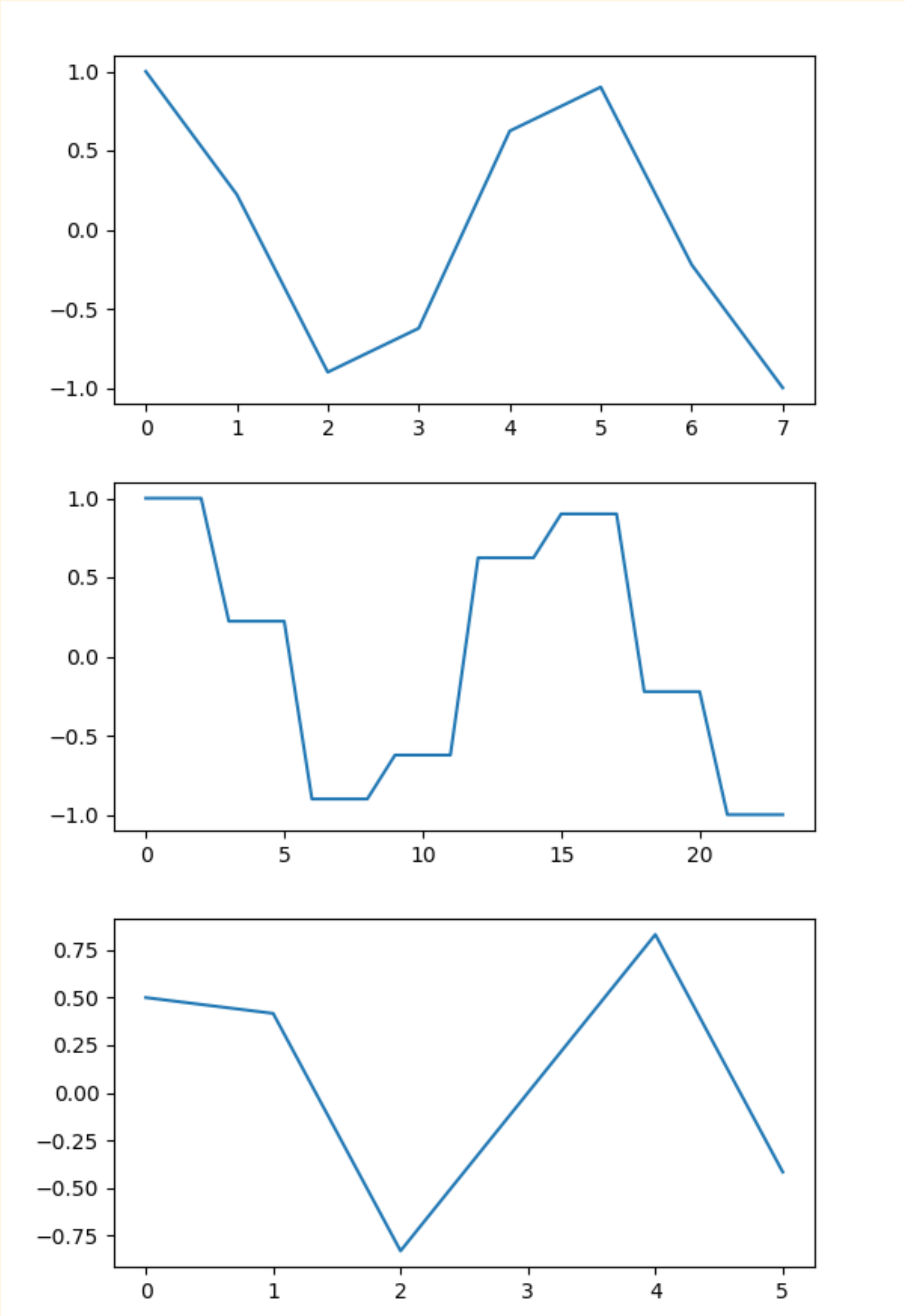
Resize operation (general case)

resize ancestor $\boldsymbol{a} \in \mathbb{R}^n$ to the length n'

1. expand the length to the $\mathrm{lcm}(n, n')$
   (least common multiple of n and n')
   by repeating each elements $\mathrm{lcm}(n, n')/n$ times

2. resize expanded ancestors to the length n'
   by taking average of $\mathrm{lcm}(n, n')/n'$ elements

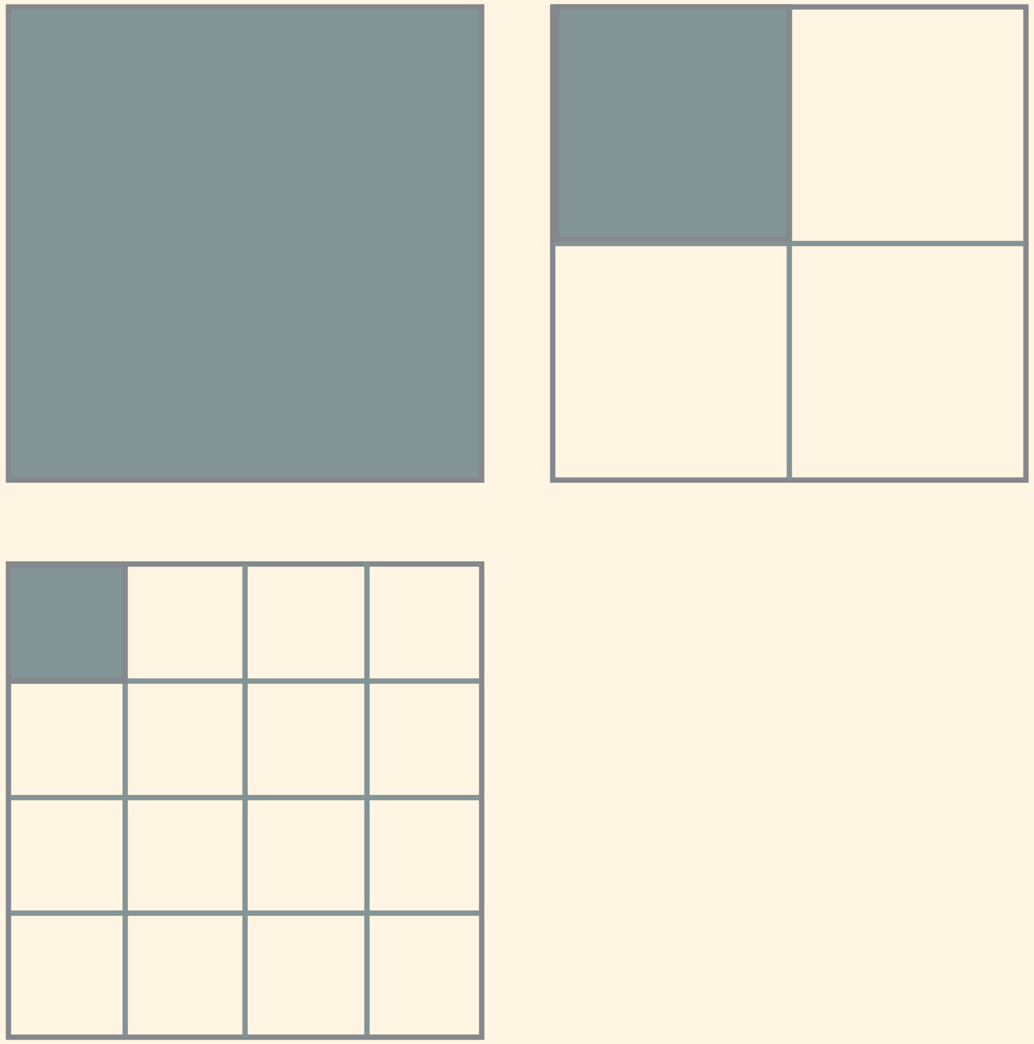
Resize operation (general case)



resize ancestor $\boldsymbol{a} \in \mathbb{R}^8$ to the length 6

1. expand the length to the 24 by repeating each elements 3 times
2. resize expanded ancestors to the length 6 by taking average of adjacent 4 elements

Multiscale K-SVD



- Scale of the features are split into quad-tree structure
- Multiple features are learned for each scale
- The relationship between scales is not considered
- Shifted atoms generated from an atom cannot be overlapped

2D Gabor dictionary

atoms are generated by sampling continuous function

$$g(x, y; \lambda, \theta, \psi \sigma \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right)$$

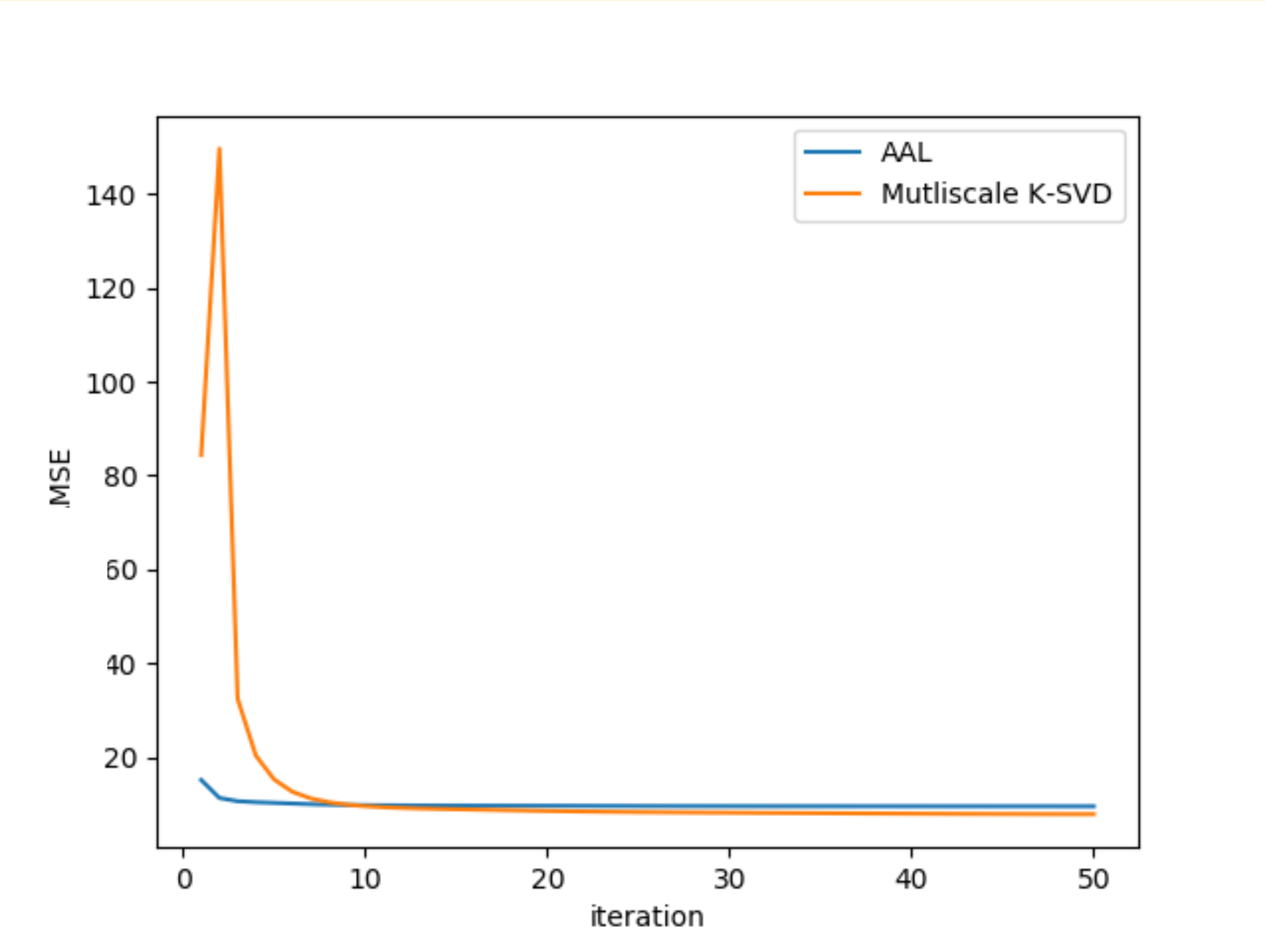
where

$$x' = x\cos\theta + y\sin\theta$$

$$y' = -x\sin\theta + y\cos\theta$$

scale of atom is controlled by σ

RMSE Curve



- In the first 10 iterations, AAL have smaller MSE
- Multi-scale K-SVD have slightly smaller MSE after 10 iterations

Computational Time

- Computational time for 50loops
- Multi-scale K-SVD learns low correlation atoms and the lasso needs small number of iterations
- AAL generate a dictionary which have high correlation therefore lasso take a long time

	AAL	Multi-scale K-SVD
Artificial	1829s	1866s
Natural Image	3h 21m	56m

Experimental Setup

Ancestral Atom Learning

- Number of ancestors : 3 (artificial data), 9 (natural image)
- Amount of shift: 2 for all scales
- Size of ancestors : 16x16, 14x14, 12x12, 10x10, 8x8
- Number of atoms generated from an ancestor: 55=1+4+9+16+25
- Regularization parameter λ : 0.01
- iteration: 50

Experimental Setup

Multi-scale K-SVD

- Number of scales: 3 (artificial, natural image)
- Number of atoms: 10 for each scale
- Size of atoms : 16x16, 8x8, 4x4
- Amount of shift: 0, 8, 4
- Number of atoms for each scale: 10, 40, 160
- Regularization parameter λ : 0.01
- iteration: 50