Bag of Pursuits and Neural Gas for Improved Sparse Coding

Manifold Learning with Sparse Coding

Thomas Martinetz

Institute for Neuro- and Bioinformatics
University of Lübeck

26.8.2010

Natural signals and images

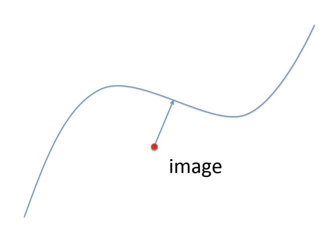
- Natural signals usually occupy only a small fraction within the signal space.
- Example: natural images lie on a submanifold within the high-dimensional image space.
- Knowledge about this submanifold is helpful in many respects.

Example: image reconstruction

90% of the pixels are missing.

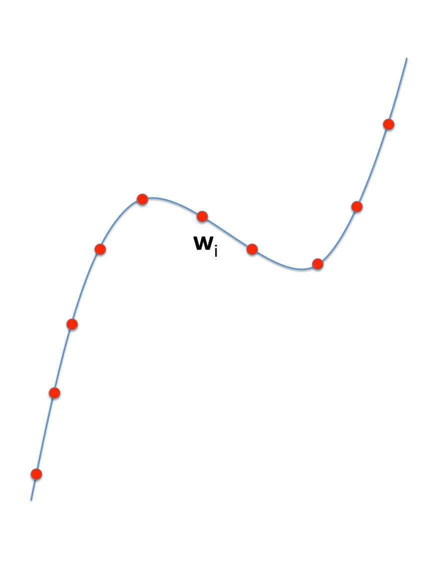
Image dimension 600x400 = 240.000

Reconstruction by projection onto the submanifold.

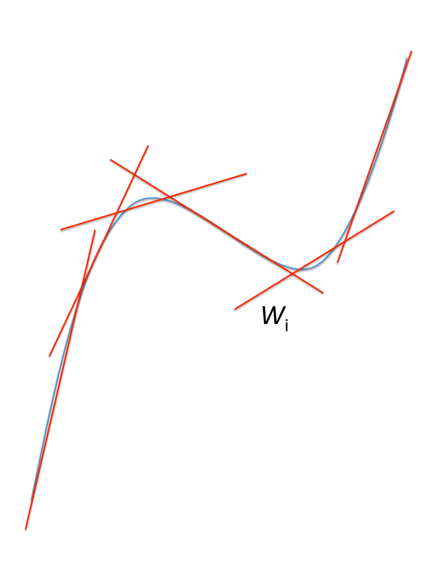


Submanifold dim. $\approx$ 10.000

Submanifold representation



- Submanifold representation by Vector Quantization.
- Each point on the submanifold is represented by its closest reference vector $\mathbf{w}_i \in \mathbb{R}^N$.
- The $\mathbf{w}_i$ can be learned by k-means, Neural Gas or many others.
- Image reconstruction through the $\mathbf{w}_i$ closest to the image.
- Submanifold representation by linear subspaces of zero dimension.

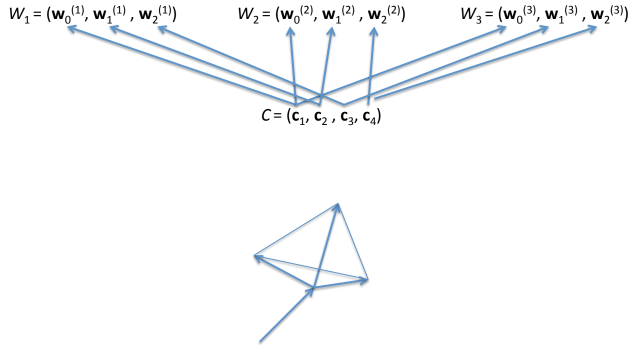


- Submanifold representation by linear subspaces.
- Each linear subspace of dimension K is defined by $W_i \in \mathbb{R}^{N\times(K+1)}$.
- Each point on the submanifold is represented by its closest linear subspace W_i.
- The W_i can be learned similar to k-means or Neural Gas.
- Image reconstruction through the closest point on the closest subspace.

Compact description

- To describe L linear subspaces of dimension K with individual W_i we need $L \times N \times (K+1)$ parameters.
- However, this description can be highly redundant.
- For example, N subspaces of dimension $N-1$ can be described by $\mathcal{O}(N^2)$ instead of N^3 parameters.
- A "K out of M" structure can be much more compact.

$N = 3$, subspace dimension $K = 2$, number of subspaces $L = 3$



- Forming K dimensional subspaces by choosing K vectors out of a set (dictionary) C of M vectors allows to realize

$$L = \binom{M}{K}$$

subspaces.

- Finding the closest subspace to a given $\mathbf{x}$ requires to solve the optimization problem

$$\min_{\mathbf{a}} \|\mathbf{x} - C\mathbf{a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{a}\|_0 = K$$

- **Problem 1**: NP-hard combinatorial optimization problem
- **Problem 2**: How to choose C for a given K?

- The manifold learning problem can be cast into the sparse coding and compressive sensing framework.

(Approximately) solving the NP-hard problem

Greedy Optimization

- Directly tackle the problem by a pursuit method
  - Matching Pursuit
  - Orthogonal Matching Pursuit
  - Optimized Orthogonal Matching Pursuit
- If $\mathbf{x}$ has a sparse enough $(K << N)$ representation, and C fulfills certain properties, the solution provided by the pursuit methods will be the optimal solution (Donoho 2003).

How to choose C (and K)?

- Given data $\mathbf{x}_1, \dots, \mathbf{x}_p$, $\mathbf{x}_i \in \mathbb{R}^N$ (like natural images) which are supposed to lie on an unknown submanifold.
- The goal is to find a C which provides a small average reconstruction error for a K which is as small as possible.

Find $C = (\mathbf{c}_1, \dots, \mathbf{c}_M)$ with $\mathbf{c}_j \in \mathbb{R}^N$ and $\mathbf{a}_i \in \mathbb{R}^M$ minimizing

$$E = \frac{1}{L} \sum_{i=1}^{p} \|\mathbf{x}_i - C\mathbf{a}_i\|_2^2$$

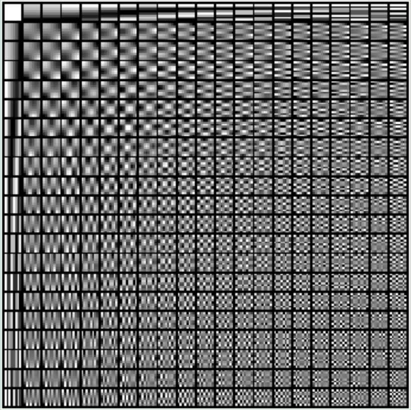
Constraints

- $\mathbf{a}_i$: $\|\mathbf{a}_i\|_0 = K$
- C : $\|\mathbf{c}_j\| = 1$ (without loss of generality)

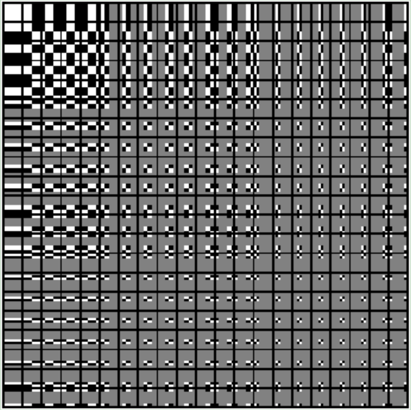
Predefined dictionaries for image data

How to chose C?

Overcomplete 8×8 DCT-Dictionary



Overcomplete 8×8 HAAR-Dictionary



The problem: find

$$\min_C \sum_i \left(\min_a \|\mathbf{x}_i - C\mathbf{a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{a}\|_0 = K \right)$$

Current state-of-the-art solver:

- MOD (Engan et al 1999)
- K-SVD (Aharon et al 2006)

Our new approach:

- Neural-Gas-like soft-competitive stochastic gradient descent.
- Generalization of the Neural Gas to linear subspaces within the sparse coding framework.

What was Neural Gas?

With a randomly chosen data point $\mathbf{x}$ reference vectors for Vector Quantization $\mathbf{w}_i$ are updated according to

$$\Delta \mathbf{w}_{j_l} = \alpha_t e^{-\frac{l}{\lambda_t}} (\mathbf{x} - \mathbf{w}_{j_l}) \qquad 0 = 1, ..., L-1$$

$\mathbf{w}_{j_0}$ is the reference vector closest to $\mathbf{x}$
$\mathbf{w}_{j_1}$ is the reference vector second closest to $\mathbf{x}$
etc.

The update step decreases with the distance rank (reconstruction error) of the reference vectors to the data point $\mathbf{x}$.

- Neural Gas performs soft-competitive stochastic gradient descent on the Vector Quantization error function.
- Neural Gas provides very good and robust solutions to the Vector Quantization problem.

With a randomly chosen data point $\mathbf{x}$ the linear subspaces W_i are updated according to

$$\Delta W_{j_l} = \alpha_t e^{-\frac{l}{\lambda_t}} (\mathbf{x} - W_{j_l} \mathbf{a}_{j_l}) \mathbf{a}_{j_l}^T \qquad l = 0, ..., L-1$$

with

$$\mathbf{a}_{j_l} = \arg\min_{\mathbf{a}} \ \|\mathbf{x} - W_{j_l} \mathbf{a}\|_2^2$$

W_{j_0} is the linear subspace closest to $\mathbf{x}$
W_{j_1} is the linear subspace second closest to $\mathbf{x}$
etc.

The update step decreases with the distance rank (reconstruction error) of the linear subspace to the data point $\mathbf{x}$.

- For a randomly chosen sample **x** determine

$$\mathbf{a}_{j_0} = \arg\min_{\mathbf{a}} \|\mathbf{x} - C\mathbf{a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{a}\|_0 = K$$

and a bag of further good solutions.

- Sort the solutions according to the obtained reconstruction error:

$$\|\mathbf{x}-C\mathbf{a}_{j_0}\| \leq \|\mathbf{x}-C\mathbf{a}_{j_1}\| \leq \cdots \leq \|\mathbf{x}-C\mathbf{a}_{j_l}\| \leq \cdots \leq \|\mathbf{x}-C\mathbf{a}_{j_{L-1}}\|$$

- Update the dictionary by soft-competitive stochastic gradient descent:

$$\Delta C = \alpha_t \sum_{l=0}^{L} e^{-\frac{l}{\lambda_t}} (\mathbf{x} - C\mathbf{a}_{j_l})\mathbf{a}_{j_l}^T$$

For finding a bag of good solutions we developed the so-called "bag of pursuits (BOP)" which

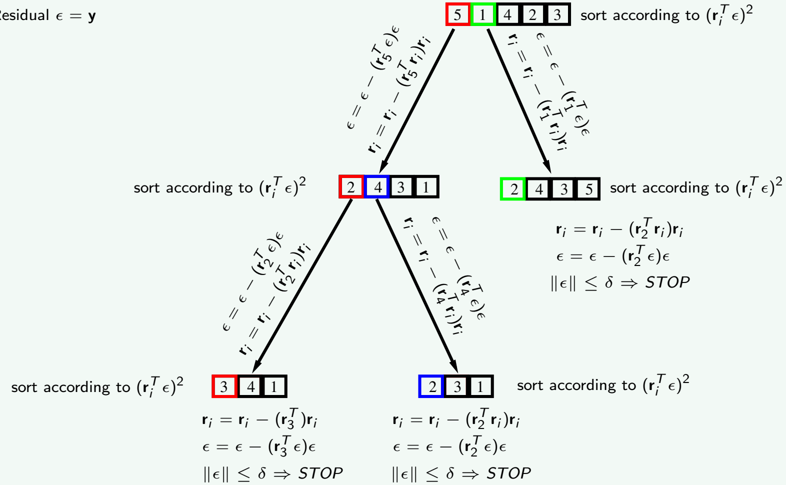
- is derived from Optimized Orthogonal Matching Pursuit
- provides a set of good choices for $\mathbf{a}$ with $\|\mathbf{a}\|_0 = K$ instead of a single solution
- expands the set of solutions in a tree-like fashion

and can be directly combined with the Neural-Gas-like stochastic gradient descent for learning dictionaries.

Bag of Pursuits (BOP)

Dictionary $R = (\mathbf{r}_1, \ldots, \mathbf{r}_5) = D$, $\|\mathbf{r}_i\| = 1$

Residual $\epsilon = \mathbf{y}$



Synthetical experiments

Do we really find the "correct" dictionary?

- Generate synthetical dictionaries $C^{\mathrm{true}} \in \mathbb{R}^{20\times 50}$ and data $\mathbf{x}_1, \ldots, \mathbf{x}_{1500} \in \mathbb{R}^{20}$ that are linear combinations of C^{true}:

$$\mathbf{x}_i = C^{\mathrm{true}}\mathbf{b}_i \ .$$

- Each $\mathbf{b}_i$ has k non-zero entries. The positions of the non-zero entries are chosen according to three different scenarios.

Random dictionary elements

- Chose uniformly k different dictionary elements

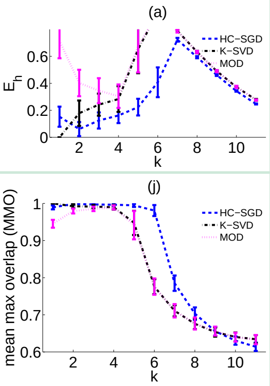
Independent subspaces

- Define $\lfloor 50/k \rfloor$ disjoint groups of k dictionary elements
- Uniformly chose one of the groups

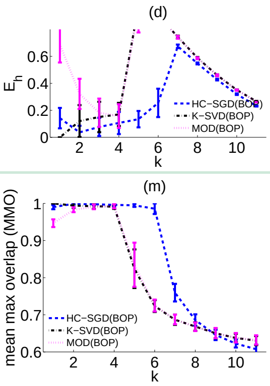
Dependent subspaces

- Uniformly select $k - 1$ dictionary elements.
- Use $50 - k + 1$ groups of dictionary elements where each group consists of the $k - 1$ selected dictionary elements plus one further dictionary element.

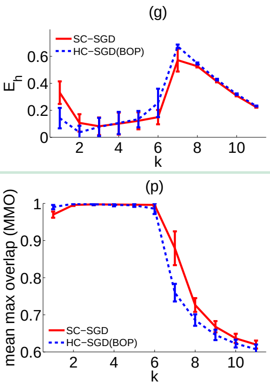
Hard-competitive without BOP



Hard-competitive with BOP

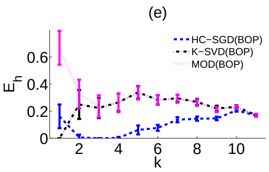


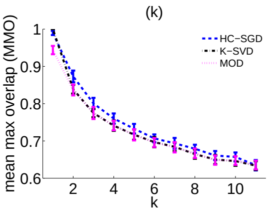
Soft-competitive with BOP



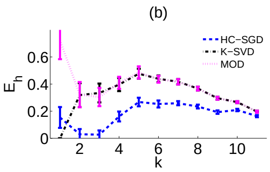
Results
Independent subspaces

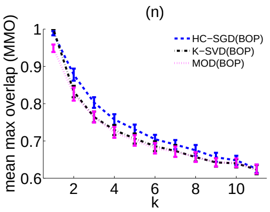
Hard-competitive without BOP



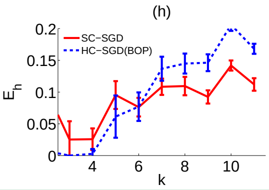


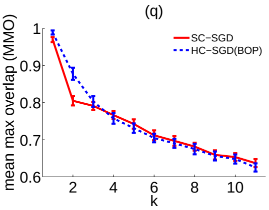
Hard-competitive with BOP



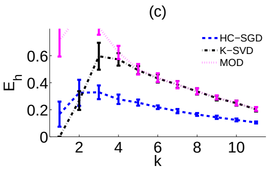


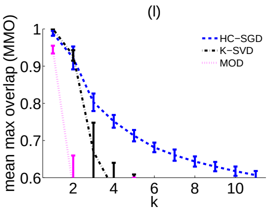
Soft-competitive with BOP



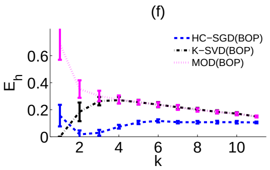


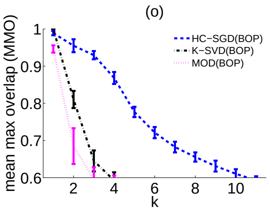
Hard-competitive without BOP



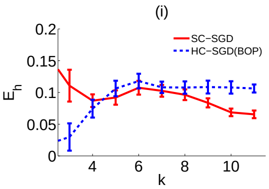


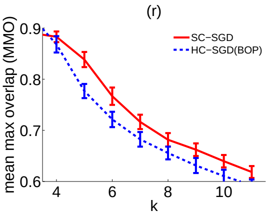
Hard-competitive with BOP





Soft-competitive with BOP





Not whole images are used for learning but 8×8 patches ($N = 64$)

Use random 8×8 patches of this image

... to learn this image specific dictionary C

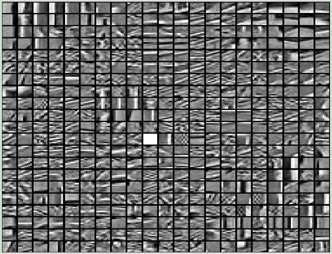
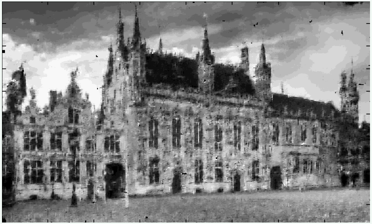


Image reconstruction

- For each 8×8 patch of the image we obtain an estimation by taking the closest point on the closest subspace
- The estimated pixel value at each image position is obtained as the mean value of all estimated patches at that position

overcomplete DCT-dictionary



learned dictionary

overcomplete HAAR-dictionary

original image