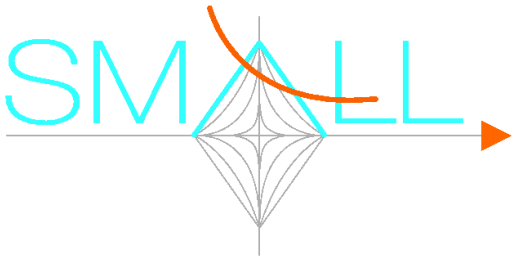
Dictionary Learning for Sparse Representations: A Pareto Curve Root Finding Approach

Mehrdad Yaghoobi, Mike E. Davies

LVA/ICA2010, Spetember 29^{th}, 2010



IDCom, School of Engineering and Electronics,
The University of Edinburgh, UK

Outline

Sparse Coding

Generative model

$$\overbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}}^{\mathbf{y}} = \overbrace{\begin{bmatrix} d_{1,1} & & d_{1,k} & & d_{1,N} \\ d_{2,1} & & d_{2,k} & & d_{2,N} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{d,1} & & d_{d,k} & & d_{d,N} \end{bmatrix}}^{\mathbf{D}} \overbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_N \end{bmatrix}}^{\mathbf{x}} + \overbrace{\begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_d \end{bmatrix}}^{\nu}$$

- sparse approximation: $\widehat{\mathbf{x}} = \arg\min_x \|\mathbf{x}\|_0, \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2 \leq \xi,$

Convex Relaxed:

- Basis Pursuit Denoising, $\widehat{\mathbf{x}} = \operatorname{argmin}_x \|\mathbf{x}\|_1, \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2 \leq \xi,$
- LASSO, $\widehat{\mathbf{x}} = \arg\min_x \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2, \quad \|\mathbf{x}\|_1 \leq \tau,$
- Tikhonov Regularized, $\widehat{\mathbf{x}} = \operatorname{argmin}_x \|\mathbf{x}\|_1 + \lambda\|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2,$

Dictionary Selection Methods

- *Concatenation of orthonormal bases:* Let $\mathcal{O}$ be the set of all orthonormal dictionaries in $\mathbb{R}^{d\times d}$. $\mathcal{D} = \{\mathbf{D}_i\}_{i\in\mathcal{I}}, \forall i \in \mathcal{I}, \mathbf{D}_i \in \mathcal{O}$ is given. A dictionary $\mathbf{D}$ in $\mathbb{R}^{d\times d|\mathcal{I}|}$ is generated using,

$$\mathbf{D} = [\mathbf{D}_1 \cdots \mathbf{D}_i \cdots \mathbf{D}_{|\mathcal{I}|}].$$

- *Dictionary design subject to a certain property:* These properties include, but not restrict to, Restricted Isometry Property (RIP), minimum coherence μ and minimum cumulative coherence $\mu_1(m)$.

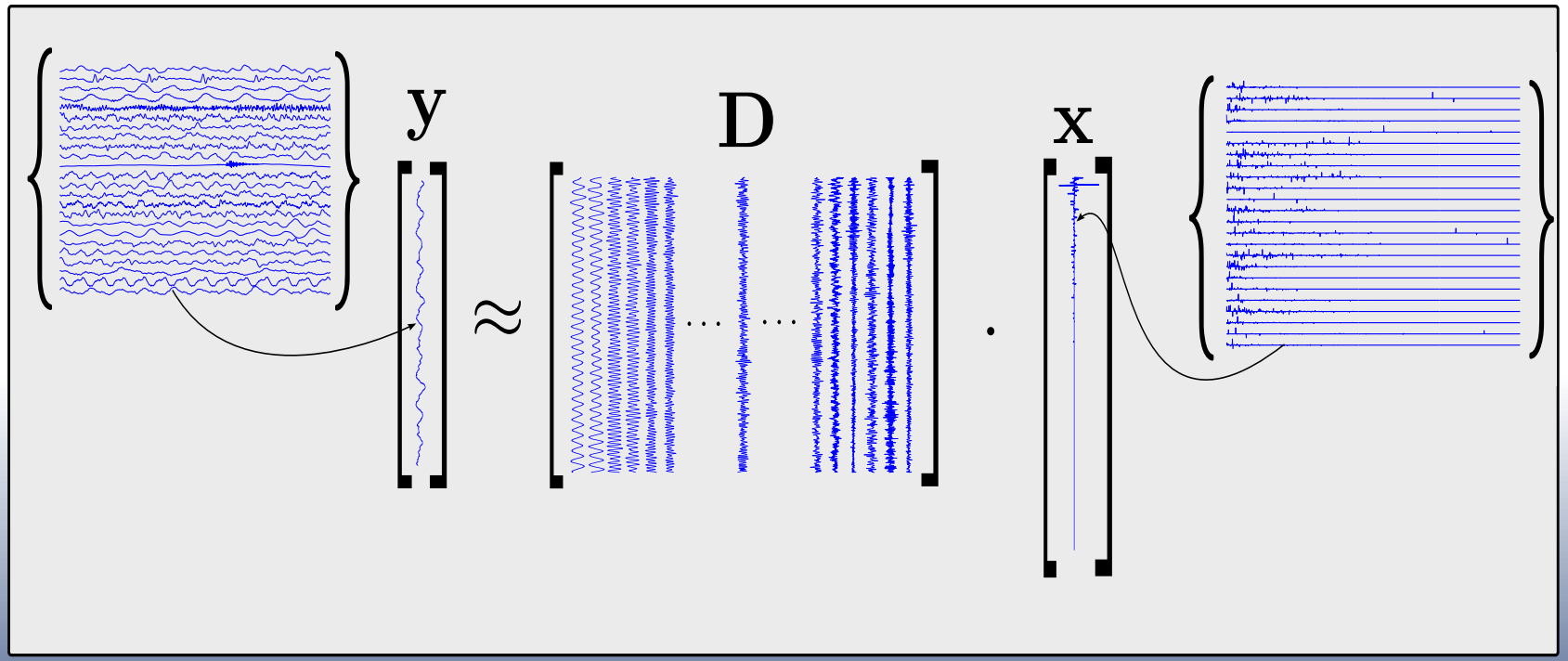
- *Dictionary learning using a set of training samples:* The goal is to find a dictionary such that it provides sparser coding for the given class of signals.

Dictionary Learning for Sparse Approximations

Definition

Let a set of training samples $\mathscr{L} = \{\mathbf{y}_i\}_{i \in \mathcal{I}}$ be given. Find a dictionary $\mathbf{D} \in \mathbb{R}^{d \times N}$ such that any training sample $\mathbf{y}_i$ has a sparse approximate representation $\mathbf{x}_i \in \mathbf{R}^N$ as follows,

$$\mathbf{y}_i \approx \mathbf{D}\mathbf{x}_i.$$



Current Dictionary Learning Methods

Minimizing a Sparse Regularized Objective

- Minimizing a joint variable objective based on the coefficient vectors and dictionary.

$$(\widehat{\mathbf{D}}, \widehat{\mathbf{X}}) = \text{argmin}_{\mathbf{D}\in\mathcal{D},\mathbf{X}}\, \mathcal{J}(\mathbf{X}) + \lambda\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$$

- It is a non-convex constrained optimization problem which has been minimized, using different block-relaxation techniques, to find a local minimum.
- $\mathcal{J}(\cdot)$ is often non-differentiable. The Projected Gradient method can not be used in its standard form.
- When $\lambda \gg 0$, the algorithm converges very slowly in practice. This formulation is thus not suitable for dictionary learning for high fidelity approximations, *i.e.* $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\| \approx 0$.

Current Dictionary Learning Methods

Minimizing a LASSO-type Problem

- Minimizing a biquadratic objective subject to some constraints on the coefficient vectors.

$$(\widehat{\mathbf{D}}, \widehat{\mathbf{X}}) = \text{argmin}_{\mathbf{D}\in\mathcal{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ s. t. } \mathcal{J}(\mathbf{X}) \leq \tau$$

- The biquadratic objective is an analytic function.
- When the projection onto $\mathcal{J}(\mathbf{X}) \leq \tau$ is possible, the Projected Gradient method is a good candidate to find the learned dictionary. Here we choose $\mathcal{J}(\cdot) = \|\cdot\|_1 = \sum_{i,j} |\{\cdot\}_{i,j}|$.
- When the smallest τ which provides $\mathbf{Y} \approx \mathbf{D}\mathbf{X}$, τ^*, is available, this framework can be used in dictionary learning for high fidelity sparse approximations.

Projected Gradient Method

- **Projected Gradient** method is an iterative algorithm for minimizing a differentiable objective $\psi(\theta)$ subject to a convex constraint $\theta \in \mathbf{\Theta}$.
- Each iteration of Projected Gradient method includes an update of current solution $\theta^{[n]}$ in the negative gradient direction and projection onto the constraint set $\mathbf{\Theta}$, $\mathcal{P}_{\mathbf{\Theta}}(\cdot)$,

$$\theta^{[n+1]} = \mathcal{P}_{\mathbf{\Theta}}\left(\theta^{[n]} - \alpha \left.\frac{\partial}{\partial\theta}\psi\right|_{\theta^{[n]}}\right)$$

- When the objective is twice differentiable, if the gradient step size α is bounded by $\alpha < \frac{\|\theta^{[n]}\|_F^2}{\left|\left.\frac{\partial^2}{\partial\theta^2}\phi\right|_{\theta^{[n]}}\right|}$, the iterative algorithm is converging monotonically.

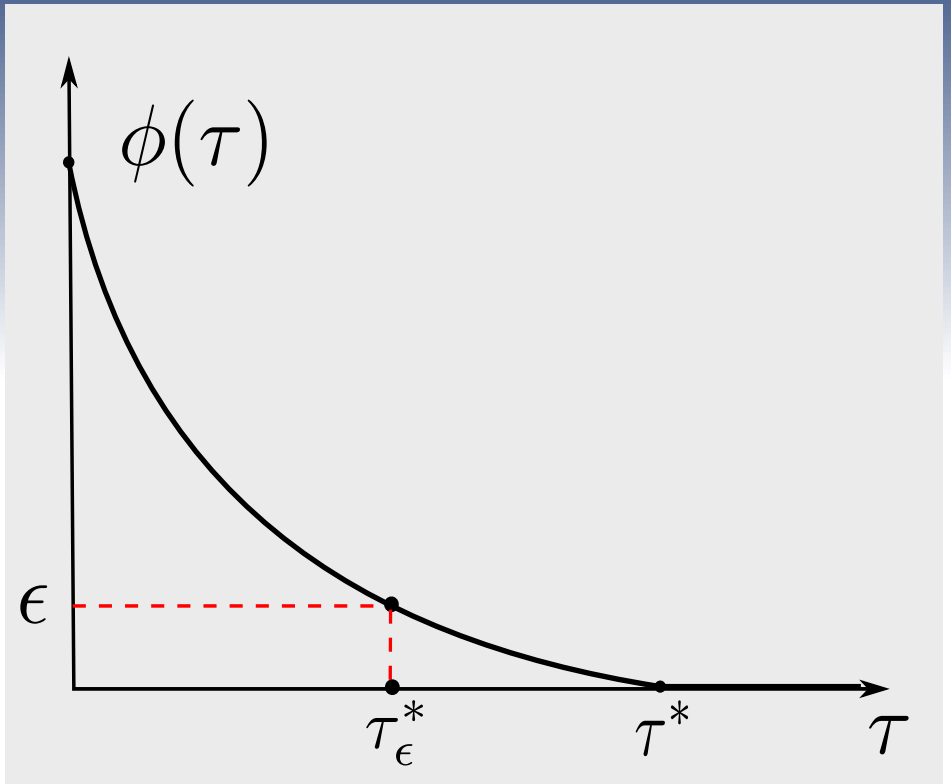
Pareto Curve Root Finding for Sparse Representations

When the optimal τ^* is not known, a Pareto curve root finding approach can be used.

$$\phi(\tau) = \min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2, \text{ s.t.}$$

$$\mathbf{X} \in \mathcal{B}_1^\tau := \{\mathbf{X} | \|\mathbf{X}\|_1 \leq \tau\},$$

$$\mathbf{D} \in \mathcal{D} := \{\mathbf{D} | \|\mathbf{d}_i\|_2 \leq 1\},$$



- $\phi(\tau)$ is non-increasing and continuous.
  - When $\mathbf{D}$ is fixed, *i.e.* sparse coding, $\phi(\tau)$ is convex and τ^* or τ_ϵ^* can be found using Newton's method.
  - In dictionary learning problem, a root finding algorithm, *e.g.* line-search and bisection, can be used.

Dictionary Learning with Root Finding by Line Search

A descriptive explanation of the proposed method:

1. For a given ϵ and precision δ, start with an initial small $\tau \leq \tau^*$.
2. Solve dictionary learning problem with the fixed τ, using projected gradient method and starting from the last solution of $\mathbf{X}$ and $\mathbf{D}$ to find a new local minima.
3. If $\|\mathbf{Y} - \mathbf{DX}\|_F^2 > \epsilon$, choose a larger τ.
4. If $\|\mathbf{Y} - \mathbf{DX}\|_F^2 \leq \epsilon$, choose a smaller τ and update $\mathbf{X}$ and $\mathbf{D}$ with the last parameters satisfy condition of step 3.
5. If $\left|\|\mathbf{Y} - \mathbf{DX}\|_F^2 - \epsilon\right| > \delta$ go to step 2.
6. Terminate

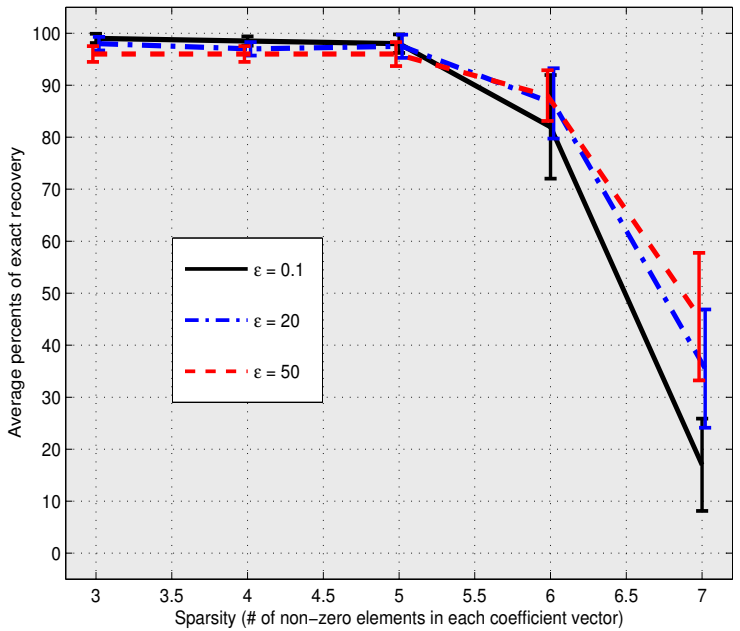
Synthetic Experiment: Dictionary Recovery Setting

- Recovering the generative dictionary by solving,

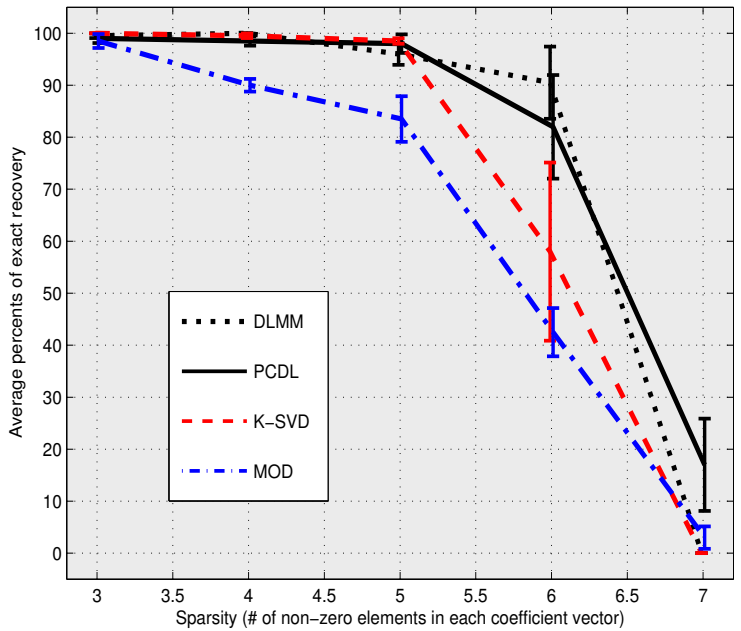
$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{X}\|_1$$
$$\mathbf{Y} = \mathbf{D}\mathbf{X}, \quad \forall i \ \|\mathbf{d}_i\|_2 = 1.$$

- A normalized random dictionary $\mathbf{D} \in \mathbf{R}^{20\times 40}$ have been used to generate $L = 1280$ K-sparse, $K \in [3, 7]$, synthetic sparse signals with randomly selecting the location of the non-zero coefficients, where each coefficient had a magnitude in $[.2, 1]$.
- We call an atom "recovered atom", if its Euclidean distance to one atom of the original dictionary is less than $.1\sqrt{2}$.
- The algorithm was terminated after 500 iterations.
- The recovery results are plotted by averaging over 5 trials.

Recovery Result



Comparing recovery results using different fidelities, $\epsilon = \{0.1, 20, 50\}$, of the representations.

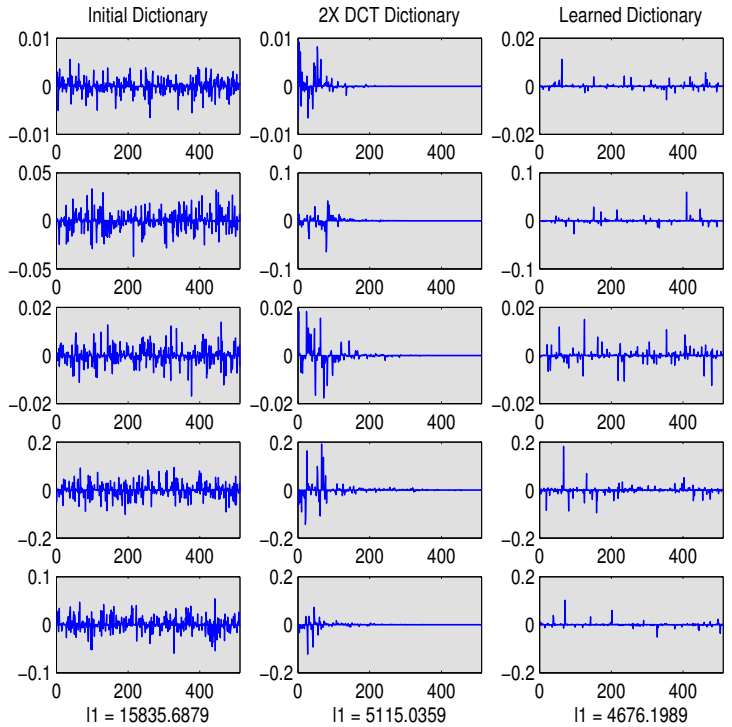


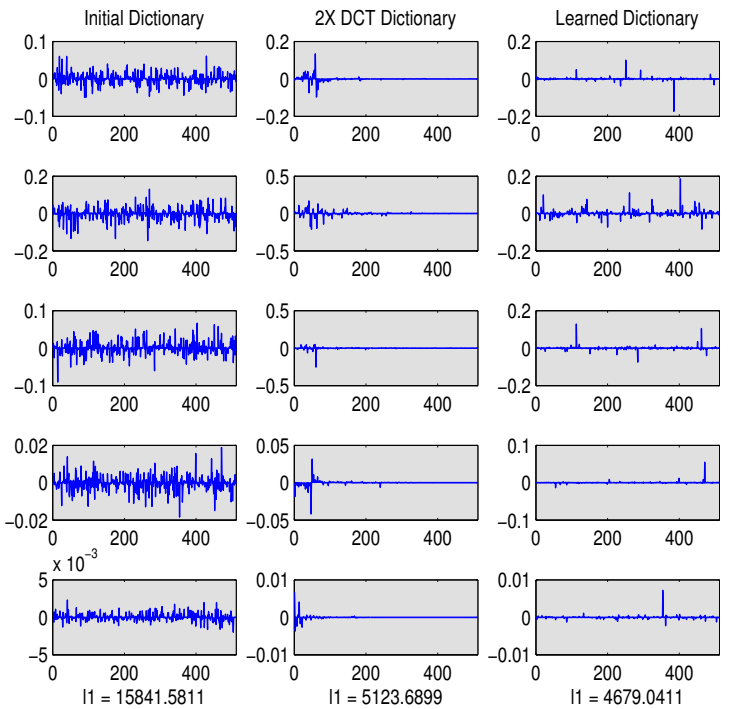
Comparing recovery results with other dictionary learning methods, where $\epsilon = 0.1$ and $\lambda = 0.4$.

Sparse Audio Coding

- High fidelity sparse coding of polyphonic music is a challenging problem, as it may not have sparse representation with available dictionaries.
- We investigate the problem of dictionary learning for high fidelity audio representations. Some classic music, recorded from BBC radio 3, have been used in this experiment.
- A random dictionary $\mathbf{D} \in \mathbf{R}^{256\times512}$ and $L = 16384$ blocks of audio signals have respectively been selected as the initial dictionary and the training signal matrix $\mathbf{Y}$.

Sparse Audio Coding Results





- Sparse representations of two audio blocks using initial, two times overcomplete DCT and learned dictionaries.
- Randomly selected 5 audio signals from training samples and out of training samples are respectively plotted on left and right panels.

Conclusion and Future Work

Conclusion

- A new LASSO-type dictionary learning problem has been introduced.
- A stable algorithm, based on Projected Gradient method, was presented to iteratively solve the problem.
- A line-search method was presented to help us find the *correct* parameter.
- The simulation results were compared to other dictionary learning methods for dictionary recovery, which shows the new method has advantages in less sparse dictionary recovery.
- The learned dictionary for audio signals shows a better performance in high fidelity sparse audio approximation.

Future Work

- A more efficient method for the inner loop dictionary learning step, based on Spectral Projected Gradient method.
- A more efficient root finding technique.
- Investigating the properties of Pareto curve, *e.g.* differentiability.

Thanks for your attention

Any questions?