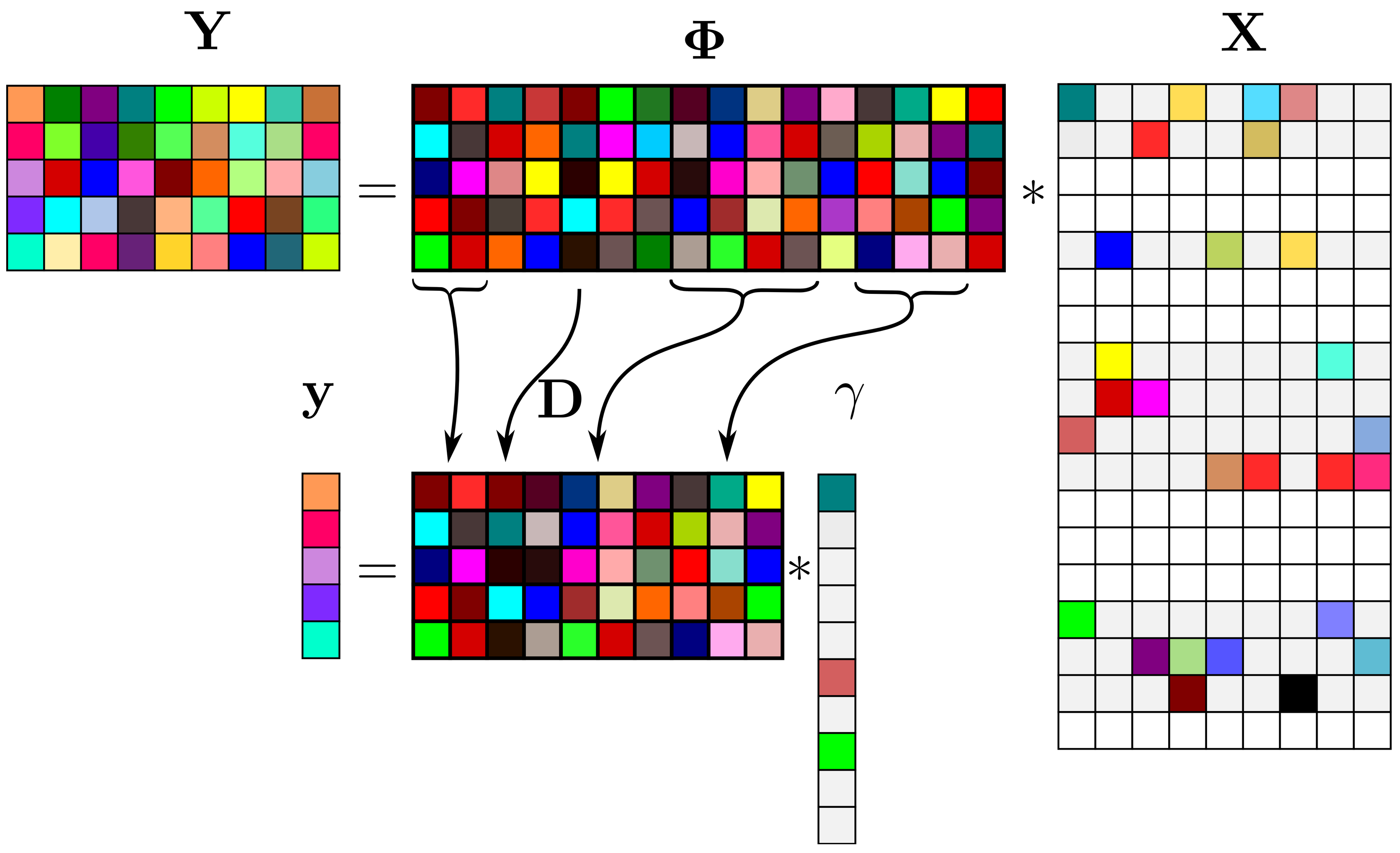
Overcomplete Joint Sparsity Model for Dictionary Selection

Mehrdad Yaghoobi†, Laurent Daudet‡ and Mike E. Davies†

† Institute for Digital Communications (IDCom), the University of Edinburgh, EH9 3JL, UK.
‡ Paris Diderot University / IUF, Institut Langevin, 1, rue Jussieu 75005 Paris, France.
Emails: yaghoobi@ieee.org, laurent.daudet@espci.fr and mike.davies@ed.ac.uk

Abstract — *The problem of dictionary selection for linear sparse approximation will be revisited in this poster. A dictionary for sparsifying a class of signals is often selected based upon the domain knowledge or using some exemplar signals. We present a new exemplar based approach for the dictionary selection, which combines the two approaches. In this framework, a large set of atoms is also given as the mother dictionary and the task is to choose a subset of the atoms, which suits the given exemplars. The new dictionary learning problem is initially formulated as a new type of joint sparsity model, which differs from the standard joint sparsity model. A simple gradient based algorithm will then be presented here to practically solve the optimisation problem. An important advantage of the new formulation is the scalability of the learning algorithm. The new dictionary selection paradigm is here examined with some synthetic experiments.*



Optimal Dictionary Selection Problem

- *Aim:* The aim of optimal dictionary selection is to find the index-set of a *sub-set* of atoms ϕ_i in a large collection of atoms, called the mother dictionary $\mathbf{\Phi} = [\phi_i]_{i \in \mathcal{I}} \in \mathbb{R}^{m \times n}$, which allows us to sparsify a class of signals.
- *Mathematical Formulation:* For any k-sparse signal $\mathbf{y}$ in $\mathbf{\Phi}$, *i.e.* $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$, $\|\mathbf{x}\|_0 \leq k$, we want to have,

$$\mathbf{y} = \mathbf{D}\gamma, \ \|\gamma\| \leq k,$$

where $\mathbf{D} = [\mathbf{d}_j]_{j \in \mathcal{J}}$, $|\mathcal{J}| = p$, is the optimal dictionary and $\mathcal{J}$ is the desired subset of $\mathcal{I}$.
- *Difference with Dictionary Learning:*
  1. The optimal dictionary is a subset of the mother dictionary $\rightarrow$ The atoms can not change in the selection process.
  2. The problem is a discrete subset selection problem $\rightarrow$ significantly easier problem (It is still a *combinatorial problem* [2]).
  3. The optimal dictionary $\mathbf{D}$ has a computationally fast implementation, if the mother dictionary has such a property.

Optimal Dictionary Selection using an Overcomplete Joint Sparsity Model

Problem Formulation

- **Overcomplete Joint Sparsity Model:** Let $\mathbf{\Theta} \in \mathbb{R}^{n \times L}$ be a coefficient matrix. The $(k,p)-$(overcomplete) joint sparse matrices lie on the intersection of $\mathcal{K}$ and $\mathcal{P}$, where

$$\mathcal{K} := \left\{ \mathbf{\Theta} \in \mathbb{R}^{n \times L} : \ \|\theta_l\|_0 \leq k, \forall l \in [1, L] \right\} \text{ and}$$

$$\mathcal{P} := \left\{ \mathbf{\Theta} \in \mathbb{R}^{n \times L} : \ \|\mathbf{\Theta}\|_{0,\infty} \leq p, \right\}.$$

- **Dictionary Selection Formulation:** Consider the system of approximate equations $\mathbf{Y} \approx \mathbf{\Phi}\mathbf{\Theta}$ for the given mother dictionary $\mathbf{\Phi}$ and a training matrix $\mathbf{Y}$. We are seeking for a $\mathbf{\Theta}$ which is p-joint sparse and also k-sparse on each column. We can then use the overcomplete joint sparsity model to find the index set $\mathcal{J}$ using,

$$\min_{\mathbf{\Theta}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{\Theta}\|_F^2, \ \text{s.t.} \ \mathbf{\Theta} \in \mathcal{K} \cap \mathcal{P},$$

where the indices corresponding to the non-zero rows, specify $\mathcal{J}$.

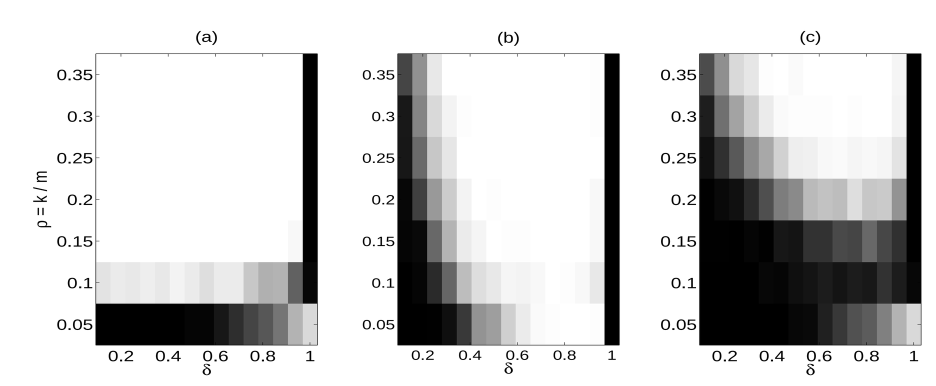
Suggested Algorithm

initialisation: $\mathbf{X}^{[0]}$, $S = \operatorname{supp}\left(\mathcal{P}_{\mathcal{K}}\left(\mathcal{P}_{\mathcal{P}}\left(\mathbf{\Phi}^T\mathbf{Y}\right)\right)\right)$, $\rho < 1$, $\beta < 1$, $\epsilon \ll 1$, $t = 0$, $K \geq 1$ and $i = 0$
while Not Converged **do**
  $\mathbf{G} = 2\mathbf{\Phi}^T\left(\mathbf{\Phi}\mathbf{X}^{[i]} - \mathbf{Y}\right)$ (Gradient), $\mu = \frac{1}{2}\frac{\mathbf{G}_S^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{G}_S}{\mathbf{G}_S^T\mathbf{G}_S}$ (Step-Size)
  $\mathbf{Z} = \mathcal{P}_{\mathcal{K}}\left(\mathcal{P}_{\mathcal{P}}\left(\mathbf{X}^{[i]} - \mu\mathbf{G}\right)\right)$
  while $\mu > \frac{\rho}{2}\frac{\|\mathbf{X}^{[i]} - \mathbf{Z}\|_F^2}{\|\mathbf{\Phi}(\mathbf{X}^{[i]} - \mathbf{Z})\|_F^2}$ **do**
    $\mu = \beta.\mu$, $\mathbf{Z} = \mathcal{P}_{\mathcal{K}}\left(\mathcal{P}_{\mathcal{P}}\left(\mathbf{X}^{[i]} - \mu\mathbf{G}\right)\right)$
  end while
  $i = i + 1$, $\mathbf{X}^{[i]} = \mathbf{Z}$, $S = \operatorname{supp}\left(\mathbf{X}^{[i]}\right)$
end while
output: $\mathbf{X}^{[i-1]}$

Simulations and Summary

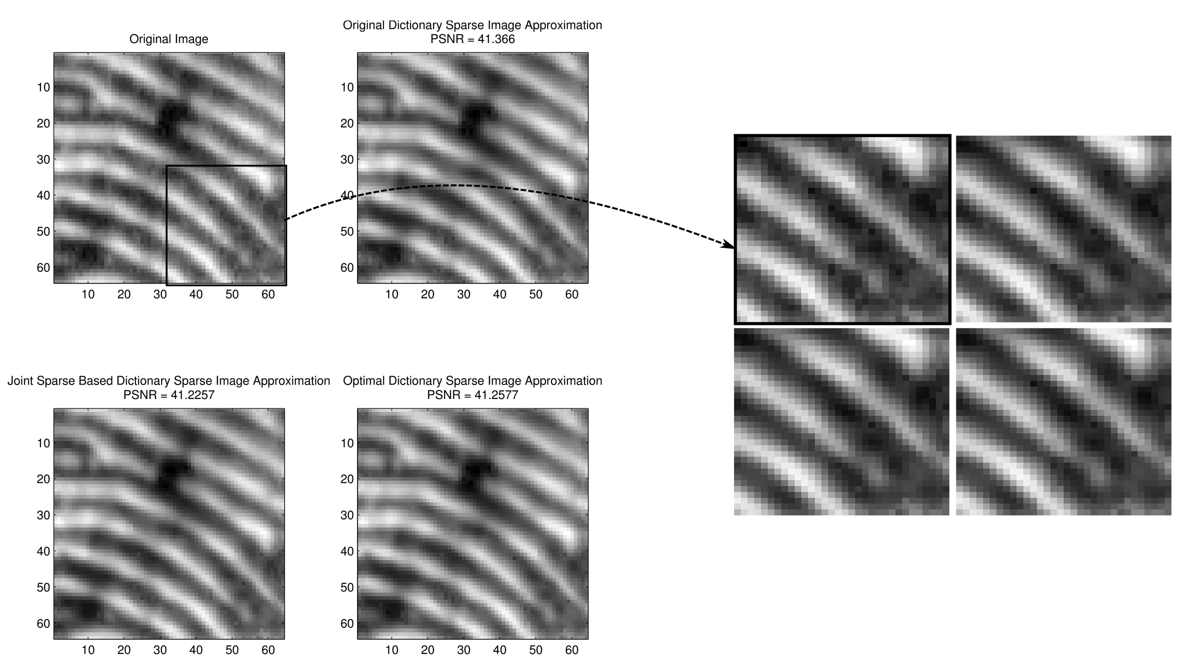
Synthetic Dictionary Recovery

- Selecting $\mathbf{D} \in \mathbb{R}^{20 \times p}$ with a randomly generated mother dictionary $\mathbf{\Phi} \in \mathbb{R}^{20 \times 80}$, when $\mathbf{Y} \in \mathbb{R}^{20 \times 320}$.
- The sparsity k and the size of target dictionary p are changing to generate the phase transition plots, $\rho = \frac{k}{m}$ and $\delta = \frac{p}{n}$, averaged over 100 trials. (Black = Exact Dictionary Recovery)
- Three different settings were used to recover the dictionary:
  1. k-*sparsity:* $\mathcal{K}$ was used as the admissible set. The indices corresponding to the largest p row norms specify $\mathcal{J}$ (a).
  2. p-*joint sparsity:* $\mathcal{P}$ was used to recover $\mathcal{J}$ (b).
  3. (k,p)-*(overcomplete) joint sparsity:* intersection of $\mathcal{K}$ and $\mathcal{P}$ was used as the admissible set (c).



Dictionary Selection for Finger Print Image Patches

- Selecting a sub-dictionary of discrete Curvelet dictionary, for the 64×64 image patches, using a set of finger print images (L = 64) as the image exemplars. $\mathbf{\Phi}$ is 2.59 times overcomplete which we want to shrink its size to half, $\mathbf{D} \in \mathbb{R}^{4096 \times 5260}$.
- k was selected to be $1025 \approx 0.1n$.
- Three settings were used here to denoise a finger print image, using a k-sparse approximation technique:
  1. using the original Curvelet dictionary $\mathbf{\Phi}$ (top right),
  2. using D_p, selected by p-joint sparsity model (bottom left),
  3. using $D_{(k,p)}$, selected by (k,p)-overcomplete joint sparsity (bottom right).
- Despite the large size of the dictionary selection problem, the simulation took less than 2 minutes for each setting.
- The PSNR of denoised image using Curvelet dictionary, is the highest, while using $\mathbf{D}_{(k,p)}$ provides the second PSNR. The denoised image using $\mathbf{D}_{(p,k)}$ is visually less distorted.



Summary

- A new signal model was presented, which can be used for the dictionary selection problem.
- An optimisation problem was introduced, which finds the optimal dictionary.
- A gradient projection based algorithm was introduced to (approximately) solve the problem.
- The new framework is a stable formulation, under some mild condition on the null-space of the mother dictionary, see [1].
- The implementation of selected dictionary is computationally fast, if the mother dictionary has such an implementation.
- The dictionary can be selected in a large size setting, where it is computationally very difficult to learn a dictionary.

Acknowledgement

This work was supported by EU FP7, FET-Open grant number 225913 and EPSRC grant EP/J015180/1. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership in Engineering and Mathematics.

[1] M. Yaghoobi, L. Daudet, and M. E. Davies, *"Optimal Dictionary Selection Using an Overcomplete Joint Sparsity Model"*, submitted, http://arxiv.org/abs/1212.2834.
[2] A. Krause and V. Cevher, *"Submodular Dictionary Selection for Sparse Representation"*, International Conference on Machine Learning (ICML), 2010.