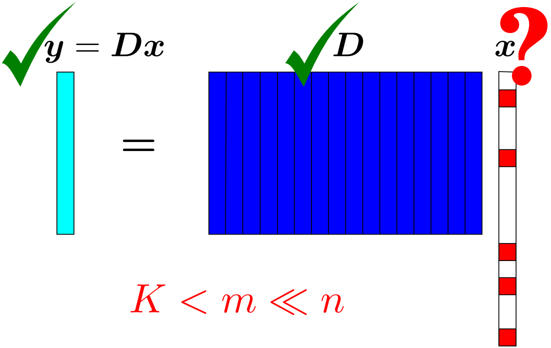
Dictionary Learning for Sparse Representations
Algorithms and Applications

Wei Dai, Boris Mailhé, & Wenwu Wang

Imperial College London
Queen Mary University of London
University of Surrey

May 2013

Sparse Representation

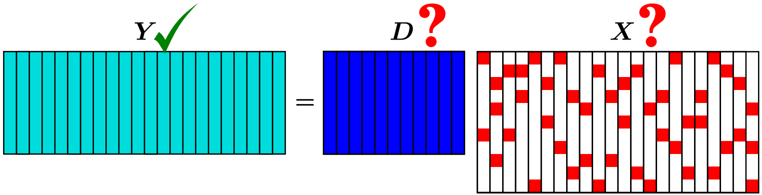


Build Good Dictionaries

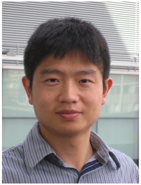
- Predefined dictionaries:
  - DCT/Wavelet dictionaries: image compression.
  - Time-frequency dictionaries: audio presentation.

- Dictionaries learned directly from the data:
  - Denoising, inpainting, $\cdots$
  - Compressed sensing: imperfect calibration.
  - Spectrum surveillance: off-grid frequencies.
  - Blind source separation: unknown dictionaries.
  - Machine learning: feature selection.

Dictionary Learning



The Speakers

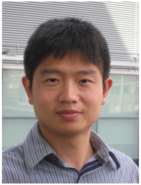


Dr. Wei Dai, Lecturer
Electrical and Electronic Engineering
Imperial College London
wei.dai1@imperial.ac.uk

Dr. Boris Mailhe, Postdoc RA
School of Electric Engineering and Computer Science
Queen Mary University of London
boris.mailhe@eecs.qmul.ac.uk

Dr. Wenwu Wang, Senior Lecturer
Department of Electronic Engineering
University of Surrey
w.wang@surrey.ac.uk

Outline



Part I

Dictionary learning: an optimization framework

Part II

Dictionary learning: extensions and toolbox

Part III

Dictionary learning: applications and final comments

Part I: Outline

Dictionary learning: an optimization framework

- Two stage procedure
  - Sparse coding
  - Dictionary update
- Dictionary update
  - MOD
  - K-SVD
  - SimCO
- Singularity issue
  - How to address the singularity issue

Acknowledgment

Imperial College London

Xiaochen Zhao
Guangyu Zhou

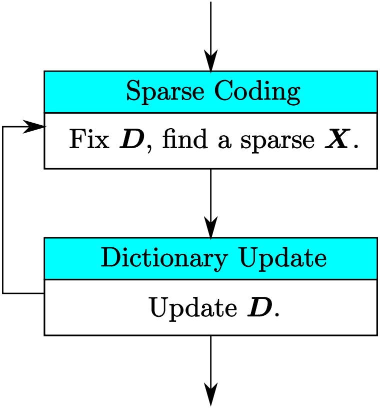
University of Surrey

Tao Xu
Wenwu Wang

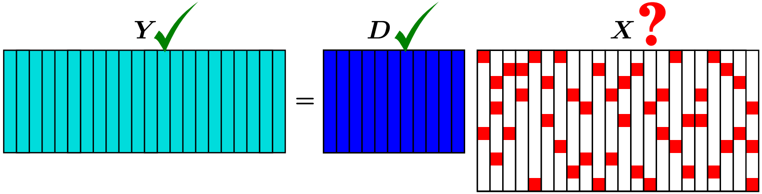
Supported by

- UK MOD University Defence Research Center in Signal Processing
- Electrical and Electronic Department, Imperial College London
- Center for Vision Speech and Signal Processing, University of Surrey
- China Scholarship Council
- Engineering and Physical Sciences Research Council of the UK (Grants EP/H050000/1 and EP/H012842/1)

A Two Stage Procedure

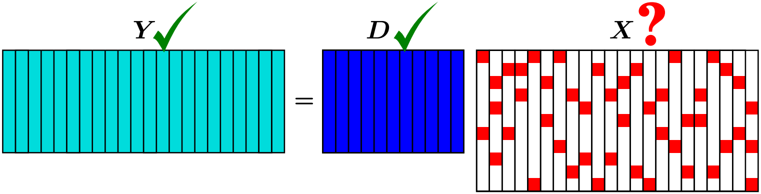


Sparse Coding



$$\min \ \|\boldsymbol{X}\|_0 \ \text{s.t.} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \leq \epsilon.$$

Sparse Coding



$$\min \ \|\boldsymbol{X}\|_0 \ \text{s.t.} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \leq \epsilon.$$

Greedy algorithms:

- OMP Y. Pati, et al. 1993; J. Tropp 2004
- Subspace pursuit (SP) W. Dai and O. Milenkovic 2009 CoSaMP D. Needell and J. Tropp 2009
- IHT T. Blumensath and M. Davies 2009

Sparse Coding: Other Approaches

ℓ_1-approach: Candes, et al. 2005; Candes, et al. 2006; Donoho 2006

- $$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_1 \text{ s.t. } \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \leq \epsilon.$$
- $$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1.$$

Bayesian approach:

- Relevance vector machine (RVM) M. Tipping 2001
- Bayesian compressed sensing (BCS) S. Ji, et al. 2008

Dictionary Update: the Formulation

- Constraints:
  - Fixed sparsity pattern

$$\begin{aligned} \Omega &= \{(i,j) : \ \boldsymbol{X}_{i,j} \neq 0\}, \\ \mathcal{X}_\Omega &= \{\boldsymbol{X} : \ \boldsymbol{X}_{i,j} = 0, \ \forall (i,j) \in \Omega^c\}. \end{aligned}$$

  - Unit norm codewords

$$\mathcal{D} = \left\{\boldsymbol{D} : \ \left\|\boldsymbol{D}_{:,j}\right\|_2 = 1, \ \forall j \in [d]\right\}.$$

- Dictionary Update:

$$\min_{\boldsymbol{D}\in\mathcal{D}, \ \boldsymbol{X}\in\mathcal{X}_\Omega} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

The MOD Method K. Engan and S. Husoy 1999

$$\min_{\boldsymbol{D}\in\mathcal{D},\, \boldsymbol{X}\in\mathcal{X}_\Omega} \ \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

MOD: least squares

1. Fix $\boldsymbol{D}$, solve for $\boldsymbol{X}$:

$$\min_{\boldsymbol{X}\in\mathcal{X}_\Omega} \ \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

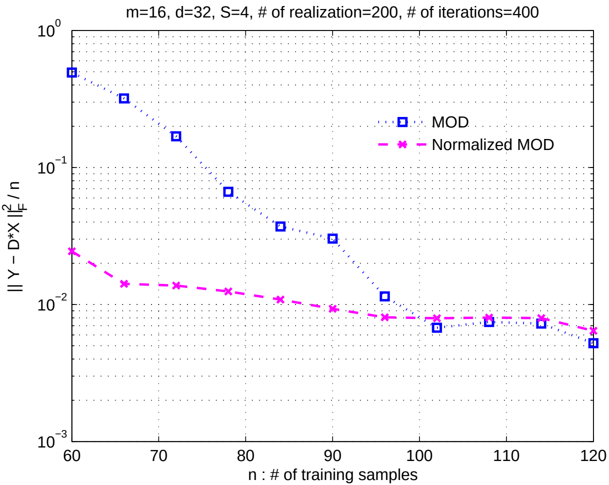
2. Fix $\boldsymbol{X}$, solve for $\boldsymbol{D}$:

$$\min_{\boldsymbol{D}} \ \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

3. (Optional) Normalization:

$$\boldsymbol{D}_{:,i} = \boldsymbol{D}_{:,i} / \|\boldsymbol{D}_{:,i}\|_2 .$$

Normalization Matters

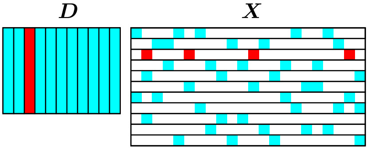


The K-SVD Method M. Aharon, et al. 2006

$$\min_{\boldsymbol{D}\in\mathcal{D},\ \boldsymbol{X}\in\mathcal{X}_\Omega} \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

For each column:

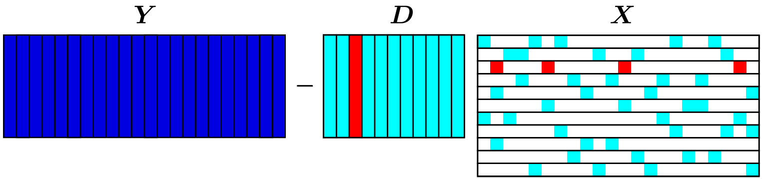
Update: this column in $\boldsymbol{D}$ & the corresponding row in $\boldsymbol{X}$.



Fix: other columns in $\boldsymbol{D}$ & the corresponding rows in $\boldsymbol{X}$.

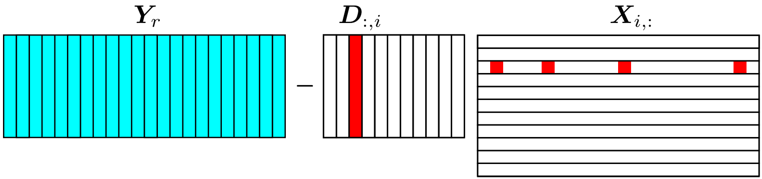
K-SVD: Details

$$\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|^2$$
$$= \|\boldsymbol{Y} - \boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2$$



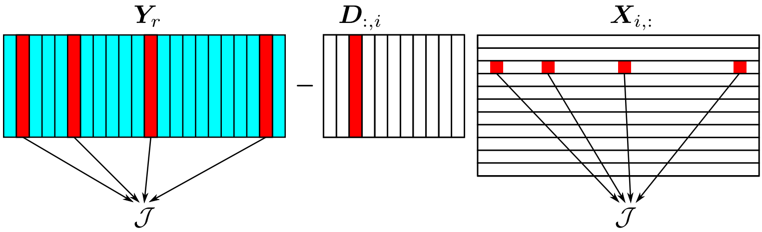
K-SVD: Details

$$\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|^2$$
$$= \|\boldsymbol{Y} - \boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2$$
$$= \|\boldsymbol{Y}_r - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2$$



K-SVD: Details

$$
\begin{aligned}
&\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|^2 \\
&= \|\boldsymbol{Y} - \boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2 \\
&= \|\boldsymbol{Y}_r - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2 \\
&= \left\|(\boldsymbol{Y}_r)_{:,\mathcal{J}} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,\mathcal{J}}\right\|^2 + c
\end{aligned}
$$



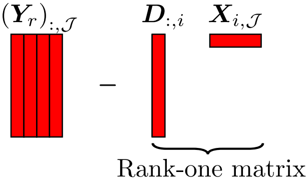
K-SVD: Details

$$\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|^2$$

$$= \|\boldsymbol{Y} - \boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2$$

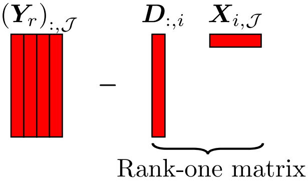
$$= \|\boldsymbol{Y}_r - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2$$

$$= \left\|(\boldsymbol{Y}_r)_{:,\mathcal{J}} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,\mathcal{J}}\right\|^2 + c$$



K-SVD: Details

$$
\begin{aligned}
&\left\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\right\|^2 \\
&= \left\|\boldsymbol{Y}-\boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:}-\boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\right\|^2 \\
&= \left\|\boldsymbol{Y}_r-\boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\right\|^2 \\
&= \left\|\left(\boldsymbol{Y}_r\right)_{:,\mathcal{J}}-\boldsymbol{D}_{:,i}\boldsymbol{X}_{i,\mathcal{J}}\right\|^2 + c
\end{aligned}
$$



SVD: optimal rank-one matrix approximation.

$$
\begin{aligned}
\boldsymbol{A} &= \sum \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \\
&\approx \lambda_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T
\end{aligned}
$$

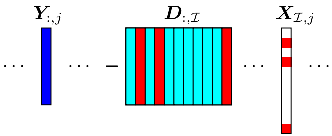
The SimCO Formulation W. Dai, et al. 2012

$$\min_{\boldsymbol{D}\in\mathcal{D},\ \boldsymbol{X}\in\mathcal{X}_\Omega} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$

The SimCO Formulation W. Dai, et al. 2012

$$
\begin{aligned}
& \min_{\boldsymbol{D}\in\mathcal{D},\, \boldsymbol{X}\in\mathcal{X}_\Omega} \left\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\right\|_F^2 \\
\Rightarrow \quad & \min_{\boldsymbol{D}\in\mathcal{D}} \underbrace{\min_{\boldsymbol{X}\in\mathcal{X}_\Omega} \left\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\right\|_F^2}_{f(\boldsymbol{D})} \\
= \quad & \min_{\boldsymbol{D}\in\mathcal{D}} f\left(\boldsymbol{D}\right)
\end{aligned}
$$

$\boldsymbol{X}$ is a function of $\boldsymbol{D}$: $\boldsymbol{X}\left(\boldsymbol{D}\right)$



$$\boldsymbol{X}_{\mathcal{I},j}\left(\boldsymbol{D}\right) = \boldsymbol{D}_{:,\mathcal{I}}^{\dagger}\boldsymbol{Y}_{:,j}$$

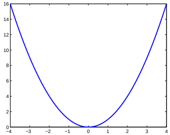
The Objective Function $f(\boldsymbol{D})$

$$f(\boldsymbol{D}) = \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}(\boldsymbol{D})\|_F^2, \text{ where } \boldsymbol{X}(\boldsymbol{D}) = \boldsymbol{D}^\dagger \boldsymbol{Y}.$$

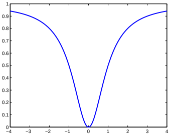
- Simultaneous Update:
  - Update $\boldsymbol{D} \Rightarrow \boldsymbol{X}(\boldsymbol{D})$ is also updated.
- Not convex in $\boldsymbol{D}$.
  Example:

$$\left\| \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ d \end{bmatrix} x \right\|_2^2$$

$x = 1$



$x = \boldsymbol{d}^\dagger \boldsymbol{y}$



Update the Dictionary

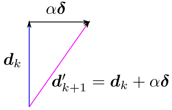
$$\min_{\boldsymbol{D}\in\mathcal{D}} f(\boldsymbol{D}) \text{ where } \mathcal{D} = \left\{\boldsymbol{D} \in \mathbb{R}^{m\times d} : \text{ unit columns}\right\}.$$

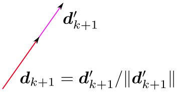
Update the Dictionary

$$\min_{\boldsymbol{D}\in\mathcal{D}} f(\boldsymbol{D}) \text{ where } \mathcal{D} = \left\{\boldsymbol{D} \in \mathbb{R}^{m\times d} : \text{ unit columns}\right\}.$$

Two ways to ensure $\boldsymbol{D} \in \mathcal{D}$:

Option 1:



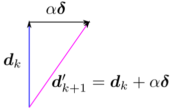


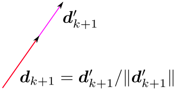
Update the Dictionary

$$\min_{\boldsymbol{D}\in\mathcal{D}} f(\boldsymbol{D}) \text{ where } \mathcal{D} = \left\{ \boldsymbol{D} \in \mathbb{R}^{m\times d} : \text{ unit columns} \right\}.$$

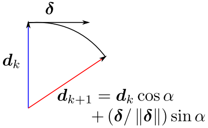
Two ways to ensure $\boldsymbol{D} \in \mathcal{D}$:

Option 1:





Option 2: (our choice) A. Edelman, et al. 1998



Connections to MOD and K-SVD

- MOD: a special case of SimCO.
  - A Gauss-Newton method to solve SimCO.

- K-SVD: also a special case of SimCO.

$$\min_{\boldsymbol{D}} \min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$
$$\Downarrow$$
$$\min_{\boldsymbol{D}_{:,i}} \min_{\boldsymbol{X}_{i,:}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$

Connections to MOD and K-SVD

- MOD: a special case of SimCO.
  - A Gauss-Newton method to solve SimCO.
- K-SVD: also a special case of SimCO.

$$\min_{\boldsymbol{D}} \min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$

$$\Downarrow$$

$$\min_{\boldsymbol{D}_{:,i}} \min_{\boldsymbol{X}_{i,:}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$

Performance: the Ideal Case

The ideal scenario:

- No noise: $\boldsymbol{Y} = \boldsymbol{D}_{\text{true}}\boldsymbol{X}_{\text{true}}$.
- True sparsity pattern is known.

Expect $\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$.

Performance: the Ideal Case

The ideal scenario:

- No noise: $\boldsymbol{Y} = \boldsymbol{D}_{\mathrm{true}} \boldsymbol{X}_{\mathrm{true}}$.
- True sparsity pattern is known.

Expect $\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X} = \mathbf{0}$.

However,

- No algorithm is guaranteed to find a global minimizer.

Performance: the Ideal Case

The ideal scenario:

- No noise: $\boldsymbol{Y} = \boldsymbol{D}_{\text{true}} \boldsymbol{X}_{\text{true}}$.
- True sparsity pattern is known.

Expect $\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$.

However,

- No algorithm is guaranteed to find a global minimizer.

Reason:

- Most failures are due to singular points.
  - $\nabla f(\boldsymbol{D}) \nrightarrow \boldsymbol{0}$.

Singular Points: Illustrative Examples

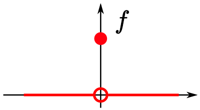
$$f\left(\boldsymbol{D}\right)=\min_{\boldsymbol{X}\in\mathcal{X}}\ \left\Vert \boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\right\Vert _{F}^{2}.$$

Singular Points: Illustrative Examples

$$f(\boldsymbol{D}) = \min_{\boldsymbol{X} \in \mathcal{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

An artificial example

$$\begin{aligned} f(d) &= \min_x \|1 - d \cdot x\|^2 \\ &= \begin{cases} 0 & \text{if } d \neq 0 \\ 1 & \text{if } d = 0 \end{cases} . \end{aligned}$$

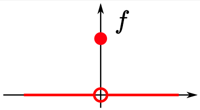


Singular Points: Illustrative Examples

$$f\left(\boldsymbol{D}\right)=\min_{\boldsymbol{X}\in\mathcal{X}}\ \left\Vert \boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\right\Vert _{F}^{2}.$$

A more realistic example

$$\begin{aligned} f\left(\epsilon\right) & =\min_{\boldsymbol{x}}\ \left\Vert \underbrace{\left[\begin{array}{c} 0\\ 1 \end{array}\right]}_{\boldsymbol{y}}-\underbrace{\left[\begin{array}{cc} 1 & \sqrt{1-\epsilon^{2}}\\ 0 & \epsilon \end{array}\right]}_{\boldsymbol{D}(\epsilon)}\left[\begin{array}{c} x_{1}\\ x_{2} \end{array}\right]\right\Vert ^{2}\\ & =\begin{cases} 0 & \text{if }\epsilon\neq0\\ 1 & \text{if }\epsilon=0 \end{cases}. \end{aligned}$$



Singular Points: a More Concrete Example

Given $\boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0.7 & 0 \\ 0 & 1 & 0.7 & 0 \\ 0 & 0 & -0.1 & 1 \\ 0 & 0 & -0.1 & 1 \end{bmatrix}$ and $\boldsymbol{X} = \begin{bmatrix} ? & 0 & 0 & ? \\ 0 & ? & 0 & ? \\ 0 & 0 & ? & ? \end{bmatrix}$,

find $\boldsymbol{D}$ and $\boldsymbol{X}$ such that $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$.

Singular Points: a More Concrete Example

Given $\boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0.7 & 0 \\ 0 & 1 & 0.7 & 0 \\ 0 & 0 & -0.1 & 1 \\ 0 & 0 & -0.1 & 1 \end{bmatrix}$ and $\boldsymbol{X} = \begin{bmatrix} ? & 0 & 0 & ? \\ 0 & ? & 0 & ? \\ 0 & 0 & ? & ? \end{bmatrix}$,

find $\boldsymbol{D}$ and $\boldsymbol{X}$ such that $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$.

Optimal solution:

$\boldsymbol{D}_{\text{opt}} = \begin{bmatrix} 1 & 0 & 0.7 \\ 0 & 1 & 0.7 \\ 0 & 0 & -0.1 \\ 0 & 0 & -0.1 \end{bmatrix}$ and $\boldsymbol{X}_{\text{opt}} = \begin{bmatrix} 1 & 0 & 0 & 7 \\ 0 & 1 & 0 & 7 \\ 0 & 0 & 1 & -10 \end{bmatrix}$.

Singular Points: a More Concrete Example

Given $\boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0.7 & 0 \\ 0 & 1 & 0.7 & 0 \\ 0 & 0 & -0.1 & 1 \\ 0 & 0 & -0.1 & 1 \end{bmatrix}$ and $\boldsymbol{X} = \begin{bmatrix} ? & 0 & 0 & ? \\ 0 & ? & 0 & ? \\ 0 & 0 & ? & ? \end{bmatrix}$,

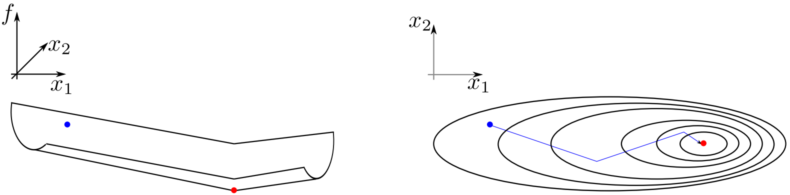
find $\boldsymbol{D}$ and $\boldsymbol{X}$ such that $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$.

Our analysis shows

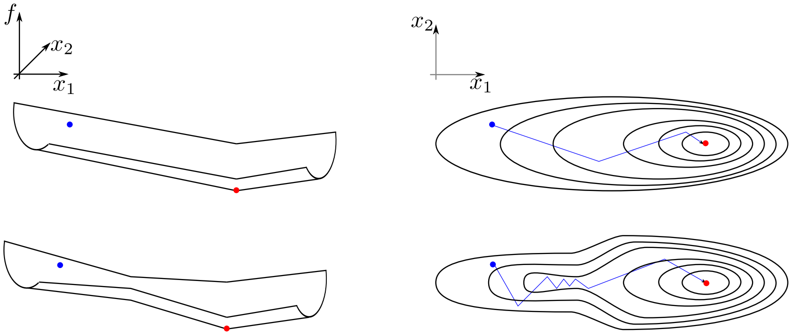
Assume $\boldsymbol{D}(\epsilon) = \begin{bmatrix} 1 & 0 & \sqrt{(1-2\epsilon^2)/2} \\ 0 & 1 & \sqrt{(1-2\epsilon^2)/2} \\ 0 & 0 & \epsilon \\ 0 & 0 & \epsilon \end{bmatrix}$ with $\epsilon_0 = 0.1$.

Benchmark algorithms: $\epsilon_k \to 0$ $(\epsilon^* = -0.1)$.

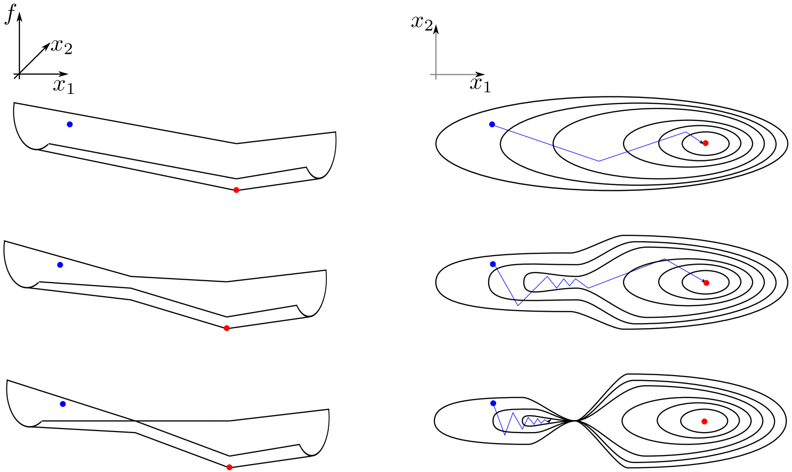
Effects of Singular Points



Effects of Singular Points



Effects of Singular Points



Handle the Singularity: Regularization?

Regularization:

$$f_r(\boldsymbol{D}) = \min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 + \mu \|\boldsymbol{X}\|_F^2.$$

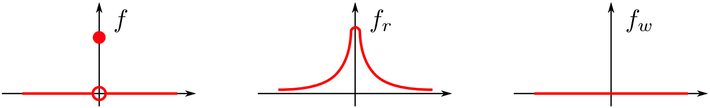
- Continuous.
  - Improve the empirical performance.

Handle the Singularity: Regularization?

Regularization:

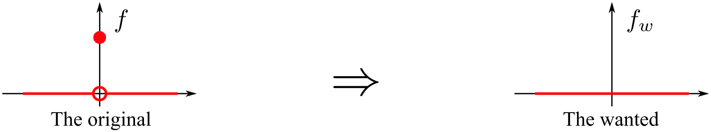
$$f_r(\boldsymbol{D}) = \min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 + \mu \|\boldsymbol{X}\|_F^2.$$

- Continuous.
  - Improve the empirical performance.
- Does not solve the singularity problem:

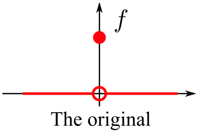


The original The regularized The wanted

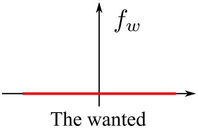
Handle the Singularity Issue: a Modulation Function



Handle the Singularity Issue: a Modulation Function

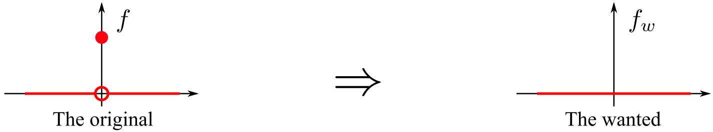


$\Rightarrow$

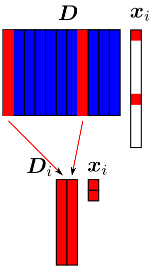


$$f(\boldsymbol{D}) = \min_{\boldsymbol{X}} \; \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$

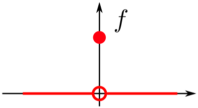
Handle the Singularity Issue: a Modulation Function



$$f(\boldsymbol{D}) = \min_{\boldsymbol{X}} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2$$
$$= \sum_i \min_{\boldsymbol{x}_i} \ \|\boldsymbol{y}_i - \boldsymbol{D}\boldsymbol{x}_i\|_2^2$$

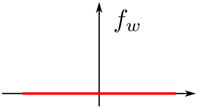


Handle the Singularity Issue: a Modulation Function



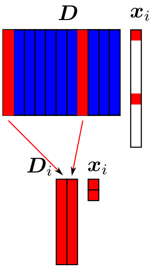
The original

$\Rightarrow$

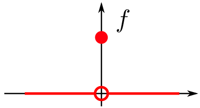


The wanted

$$
\begin{aligned}
f(\boldsymbol{D}) &= \min_{\boldsymbol{X}} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \\
&= \sum_i \min_{\boldsymbol{x}_i} \ \|\boldsymbol{y}_i - \boldsymbol{D}\boldsymbol{x}_i\|_2^2 \\
&= \sum_i \underbrace{\min_{\boldsymbol{x}_i} \ \|\boldsymbol{y}_i - \boldsymbol{D}_i\boldsymbol{x}_i\|_2^2}_{f_i(\boldsymbol{D}_i)}.
\end{aligned}
$$

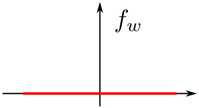


Handle the Singularity Issue: a Modulation Function



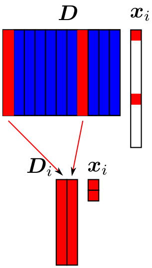
The original

$\Rightarrow$



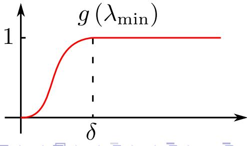
The wanted

$$
\begin{aligned}
f(\boldsymbol{D}) &= \min_{\boldsymbol{X}} \ \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 \\
&= \sum_i \min_{\boldsymbol{x}_i} \ \|\boldsymbol{y}_i - \boldsymbol{D}\boldsymbol{x}_i\|_2^2 \\
&= \sum_i \underbrace{\min_{\boldsymbol{x}_i} \ \|\boldsymbol{y}_i - \boldsymbol{D}_i\boldsymbol{x}_i\|_2^2}_{f_i(\boldsymbol{D}_i)}.
\end{aligned}
$$



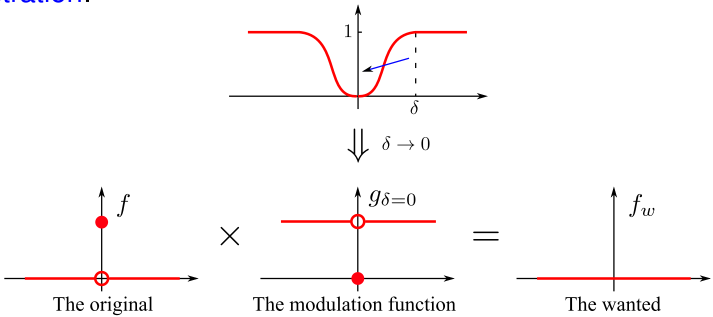
$$\tilde{f}(\boldsymbol{D}) = \sum_i \ f_i(\boldsymbol{D}_i) \cdot g_\delta(\lambda_{\min}(\boldsymbol{D}_i)).$$

- Singular points $\Leftrightarrow \exists i$ s.t. $\lambda_{\min}(\boldsymbol{D}_i) = 0$.
- g_δ is double differentiable.



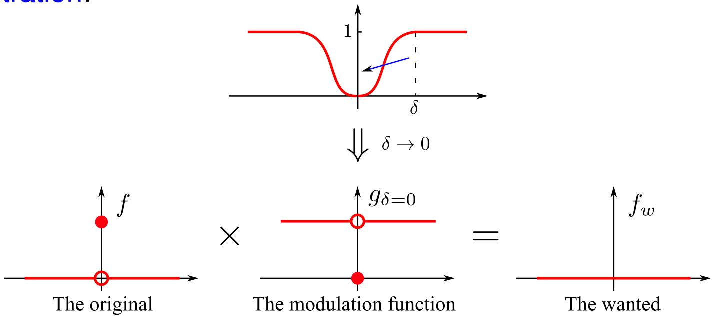
Effect of the Modulation Function

1-d illustration:

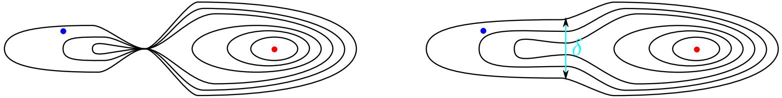


Effect of the Modulation Function

1-d illustration:



2-d illustration:



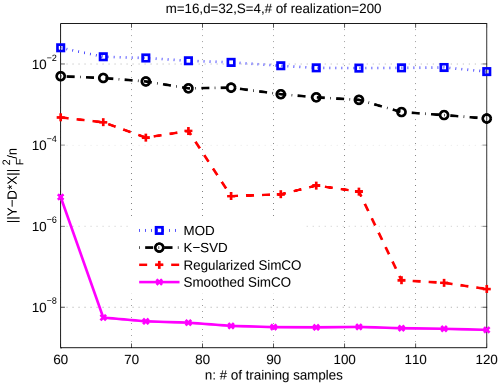
Effect of the Modulation Function: Theoretic Results

Theorem

- *When $\delta > 0$, $\tilde{f}$ is continuous.*
- *When $\delta \to 0$, $\tilde{f}$ is the best possible lower semi-continuous approximation of f.*
  - *$\tilde{f}$ and f differ only at singular points.*
  - *The lower level sets of $\tilde{f}$ are the closure of the lower sets of f.*

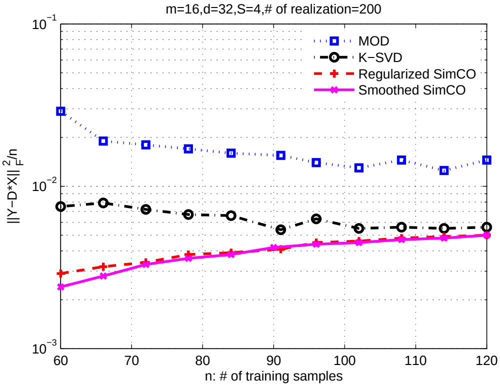
Empirical Performance 1

- The true sparsity pattern Ω_{true} is given.
- Noiseless case.



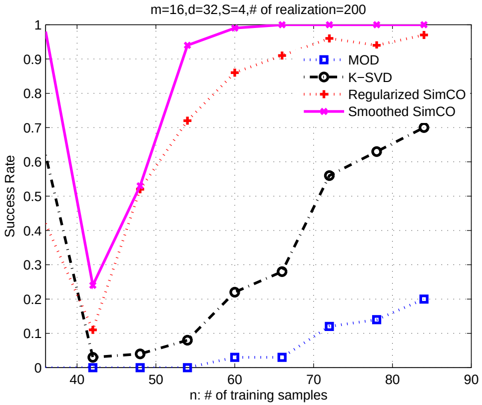
Empirical Performance 2

- The true sparsity pattern Ω_{true} is given.
- Noisy case.



Empirical Performance 3

- The true sparsity pattern Ω_{true} is given.
- Noiseless case.
- Success rate.



Implementation: A Newton CG Method

Gradient descent: slow convergence.

Newton CG: fast convergence.

Implementation: A Newton CG Method

Gradient descent: slow convergence.

Newton CG: fast convergence.

$$f_i = \min_{\boldsymbol{x}_i} \ \left\|\boldsymbol{y}_i - \boldsymbol{D}_i \boldsymbol{x}_i\right\|^2$$
$$= \left\|\boldsymbol{y}_i - \boldsymbol{D}_i \boldsymbol{x}_i^*\right\|^2 \text{ where } \boldsymbol{x}_i^* = \boldsymbol{D}_i^\dagger \boldsymbol{y}_i.$$

- Newton method: $\nabla \boldsymbol{D}_i^\dagger$.
- Newton CG: directional derivative of $\boldsymbol{D}_i^\dagger$.

Directional Derivatives

Gradient:

$$\tilde{f}(\boldsymbol{D}) : \ \mathbb{R}^{m \times n} \to \mathbb{R}$$

$$\nabla \tilde{f} = [\partial f / \partial \boldsymbol{D}_{i,j}] \in \mathbb{R}^{m \times n}$$

$$\nabla^2 \tilde{f} = \left[\partial^2 f / \partial \boldsymbol{D}_{i,j} \partial \boldsymbol{D}_{k,\ell}\right] \in \mathbb{R}^{(m \cdot n) \times (m \cdot n)}$$

Consider $\dim(\boldsymbol{D}) = 64 \times 128$:

$$\dim\left(\nabla^2 \tilde{f}\right) \approx 8000 \times 8000 \approx 64,000,000.$$

Directional Derivatives

Gradient:

$$\tilde{f}\left(\boldsymbol{D}\right): \ \mathbb{R}^{m\times n} \to \mathbb{R}$$

$$\nabla \tilde{f} = \left[\partial f / \partial \boldsymbol{D}_{i,j}\right] \in \mathbb{R}^{m\times n}$$

$$\nabla^2 \tilde{f} = \left[\partial^2 f / \partial \boldsymbol{D}_{i,j} \partial \boldsymbol{D}_{k,\ell}\right] \in \mathbb{R}^{(m\cdot n)\times(m\cdot n)}$$

Consider $\dim\left(\boldsymbol{D}\right) = 64 \times 128$:

$$\dim\left(\nabla^2 \tilde{f}\right) \approx 8000 \times 8000 \approx 64,000,000.$$

Directional gradient:

- $\nabla_{\boldsymbol{\eta}} \tilde{f} \triangleq \lim_{t\to 0} \frac{\tilde{f}(\boldsymbol{D}+t\boldsymbol{\eta}) - \tilde{f}(\boldsymbol{D})}{t} \in \mathbb{R}$.
- $\nabla_{\boldsymbol{\eta}} \nabla \tilde{f} = \lim_{t\to 0} \frac{\nabla \tilde{f}\big|_{\boldsymbol{D}+t\boldsymbol{\eta}} - \nabla \tilde{f}\big|_{\boldsymbol{D}}}{t} \in \mathbb{R}^{m\times n}$.

Complexity is highly reduced.

Weighting: Make the Complexity Further Lower

Smoothed objective function:

$$\tilde{f} = \sum_i f_i\left(\boldsymbol{D}\right) g_\delta\left(\lambda_{\min}\left(\boldsymbol{D}_i\right)\right)$$

Compared to $f = \sum_i f_i$:

$g,\ \nabla g,\ \nabla_{\boldsymbol{\eta}} \nabla g$ require extra computations.

Weighting: Make the Complexity Further Lower

Smoothed objective function:

$$\tilde{f} = \sum_i f_i\left(\boldsymbol{D}\right) g_\delta\left(\lambda_{\min}\left(\boldsymbol{D}_i\right)\right)$$

Compared to $f = \sum_i f_i$:

$g,\ \nabla g,\ \nabla_{\boldsymbol{\eta}} \nabla g$ require extra computations.

Weighted objective function:

At the k^{th} optimization iteration:

$$\hat{f} = \sum_i f_i\left(\boldsymbol{D}\right) \cdot g_\delta\left(\lambda_{\min}\left(\boldsymbol{D}_i^{(k)}\right)\right)$$

$$= \sum_i f_i\left(\boldsymbol{D}\right) \cdot w_i^{(k)}$$

$w_i^{(k)}$: a constant in the k^{th} iteration.

- A Newton method similar to MOD ($w_i \equiv 1$).
- Mitigate the singular issue.

A Summary

- Dictionary learning.
  - MOD
  - K-SVD
  - SimCO
- Singularity problem
  - A modulation function to smooth the objective function

Key References

- Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition." IEEE Asilomar Conference on Signals, Systems and Computers, pp. 40-44., 1993.
- J. A. Tropp. "Greed is good: Algorithmic results for sparse approximation." IEEE Transactions on Information Theory, vol. 50, no. 10 (2004): 2231-2242.
- W. Dai, and O. Milenkovic. "Subspace pursuit for compressive sensing signal reconstruction." IEEE Transactions on Information Theory, vol. 55, no. 5 (2009): 2230-2249.
- D. Needell, and J. A. Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples." Applied and Computational Harmonic Analysis 26, no. 3 (2009): 301-321.
- T. Blumensath, and M. E. Davies. "Iterative hard thresholding for compressed sensing." Applied and Computational Harmonic Analysis, vol. 27, no. 3 (2009): 265-274.

Key References

- E. J. Candès, J. Romberg, and T. Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." IEEE Transactions on Information Theory, vol. 52, no. 2 (2006): 489-509.
- E. J. Candès, and T. Tao. "Decoding by linear programming." IEEE Transactions on Information Theory, vol. 51, no. 12 (2005): 4203-4215.
- D. L. Donoho. "Compressed sensing." IEEE Transactions on Information Theory, vol. 52, no. 4 (2006): 1289-1306.
- M. E. Tipping. "Sparse Bayesian learning and the relevance vector machine." The Journal of Machine Learning Research (2001): 211-244.
- S. Ji, Y. Xue, and L. Carin. "Bayesian compressive sensing." IEEE Transactions on Signal Processing, vol. 56, no. 6 (2008): 2346-2356.

Key References

- K. Engan, S. O. Aase, and J. H. Husoy. "Method of optimal directions for frame design." IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 2443-2446, 1999.
- M. Aharon, M. Elad, and A. Bruckstein. "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation." IEEE Transactions on Signal Processing, vol. 54, no. 11 (2006): 4311-4322.
- M. Elad, and M. Aharon. "Image denoising via sparse and redundant representations over learned dictionaries." IEEE Transactions on Image Processing, vol. 15, no. 12 (2006): 3736-3745.

Key References

- W. Dai, T. Xu, and W. Wang. "Simultaneous codeword optimization (SimCO) for dictionary update and learning. " IEEE Transactions on Signal Processing, vol. 60, no. 12 (2012): 6340-6353.
- A. Edelman, T. A. Arias, and S. T. Smith. "The geometry of algorithms with orthogonality constraints." SIAM journal on Matrix Analysis and Applications, vol. 20, no. 2 (1998): 303-353.
- X. Zhao, G. Zhou, and W. Dai. "Smoothed SimCO for dictionary learning: handling the singularity issue." IEEE International Conference on Acoustics, Speech, and Signal Processing, accepted, 2013.

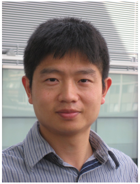
Dictionary Learning for Sparse Representations
Algorithms and Applications

Wei Dai, Boris Mailhé, & Wenwu Wang

Imperial College London
Queen Mary University of London
University of Surrey

May 2013

The Speakers



Dr. Wei Dai, Lecturer
Electrical and Electronic Engineering
Imperial College London
wei.dai1@imperial.ac.uk

Dr. Boris Mailhé, Postdoc RA
School of Electric Engineering and Computer Science
Queen Mary University of London
boris.mailhe@eecs.qmul.ac.uk

Dr. Wenwu Wang, Senior Lecturer
Department of Electronic Engineering
University of Surrey
w.wang@surrey.ac.uk

Thanks

Prof. M. D. Plumbley
QMUL

Dr D. Barchiesi
QMUL

Dr R. Gribonval
INRIA

Prof. P. Vandergheynst
EPFL

Dr F. Bimbot
CNRS

- EPSRC Project EP/G007144/1 Machine Listening using Sparse Representations
- EU FET-Open project FP7-ICT-225913-SMALL

1. Global optimization ideas for dictionary learning
    - Online methods
    - Step size influence, LGD
2. Structured dictionary learning
    - Shift-invariant dictionary learning
    - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
    - Overview
    - Toolbox contents

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Non-convexities in dictionary learning

- Dictionary learning:

$$\left(\hat{\mathbf{D}}, \hat{\mathbf{X}}\right) = \min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$$
$$\text{s. t. } \|\mathbf{x}_{\mathbf{y}}\|_0 \leq K, \ \forall \mathbf{y} \in \mathbf{Y}$$
$$\text{and } \|\mathbf{d}\|_2 = 1, \ \forall \mathbf{d} \in \mathbf{D}$$

- 2 sources of non-convexity:
  - the ℓ_0 constraint,
  - the matrix product $\mathbf{DX}$ where both $\mathbf{D}$ and $\mathbf{X}$ are variables.
  - (the ℓ_2 normalization turns out to be convex.)
- Ideas for dictionary update: use stochastic updates to find a global minimum.

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

ODL [Mairal10] and RLS-DLA [Skretting10]

- Stochastic gradient: use approached gradients to avoid local minima.
- Online processing: at iteration i, only the first i data points are available.

$$f_{[1,i]}(\mathbf{D}, \mathbf{X}_{[1,i]}) = \left\| \mathbf{Y}_{[1,i]} - \mathbf{D}\mathbf{X}_{[1,i]} \right\|_F^2$$

- Real-time: the complexity of an iteration must be constant over time.

for $i = 1$ to I **do**
  $\mathbf{x}_i$ = decomp($\mathbf{y}_i$, $\mathbf{D}$)
  $\mathbf{D}$ = dict_update($\mathbf{Y}_{[1,i]}$, $\mathbf{X}_{[1,i]}$)
  normalize($\mathbf{D}$)
end for

- How to perform the dictionary update with constant complexity?

Online dictionary updates

$$f_{[1,i]}(\mathbf{D}, \mathbf{X}_{[1,i]}) = \left\| \mathbf{Y}_{[1,i]} - \mathbf{D}\mathbf{X}_{[1,i]} \right\|_F^2$$

- Successive optimal step gradient descent (ODL):

$$\begin{aligned} \mathbf{d} &\leftarrow \mathbf{d} + \frac{1}{\left\| \mathbf{x}_{[1,i]}^d \right\|_2^2} \left(\mathbf{Y}_{[1,i]} - \mathbf{D}\mathbf{X}_{[1,i]} \right) \mathbf{x}_{[1,i]}^d{}^* \\ &= \mathbf{d} + \frac{1}{\left\| \mathbf{x}_{[1,i]}^d \right\|_2^2} \left(\mathbf{Y}_{[1,i]} \mathbf{x}_{[1,i]}^d{}^* - \mathbf{D}\mathbf{X}_{[1,i]} \mathbf{x}_{[1,i]}^d{}^* \right) \end{aligned}$$

- Least-squares solution (RLS-DLA):

$$\begin{aligned} \mathbf{D} &\leftarrow \mathbf{Y}_{[1,i]} \mathbf{X}_{[1,i]}{}^\dagger \\ &= \mathbf{Y}_{[1,i]} \mathbf{X}_{[1,i]}{}^* \left(\mathbf{X}_{[1,i]} \mathbf{X}_{[1,i]}{}^* \right)^{-1} \end{aligned}$$

Constant complexity updates

$$\mathbf{A}^{(i)} = \mathbf{X}_{[1,i]}\mathbf{X}_{[1,i]}{}^* \qquad \mathbf{B}^{(i)} = \mathbf{Y}_{[1,i]}\mathbf{X}_{[1,i]}{}^*$$

- Computing $\mathbf{A}^{(i)}$ and $\mathbf{B}^{(i)}$ in constant time:

$$\mathbf{A}^{(i)} = \mathbf{A}^{(i-1)} + \mathbf{x}_i\mathbf{x}_i{}^* \qquad \mathbf{B}^{(i)} = \mathbf{B}^{(i-1)} + \mathbf{y}_i\mathbf{x}_i{}^*$$

- ODL:

$$\mathbf{d} \leftarrow \mathbf{d} + \frac{1}{a^{(i)}_{\mathbf{d},\mathbf{d}}}\left(\mathbf{b}^{(i)}_{\mathbf{d}} - \mathbf{D}\mathbf{a}^{(i)}_{\mathbf{d}}\right)$$

- RLS-DLA:

$$\mathbf{D} \leftarrow \mathbf{B}^{(i)}\mathbf{A}^{(i)}{}^{-1}$$

Forgetting factor

- the signal $\mathbf{y}_i$ is used in all iterations from i to I.
- Early selected signals carry more weight than late ones.
- Fix: decrease the influence of the past data over time

$$\mathbf{A}^{(i)} = \beta_i \mathbf{A}^{(i-1)} + \mathbf{x}_i {\mathbf{x}_i}^* \qquad \mathbf{B}^{(i)} = \beta_i \mathbf{B}^{(i-1)} + \mathbf{y}_i {\mathbf{x}_i}^*$$

with $0 < \beta_i < 1$.

$\mathbf{A} \leftarrow 0$, $\mathbf{B} \leftarrow 0$
for $i = 1$ to I **do**
  $\mathbf{x}_i$ = decomp($\mathbf{y}_i$, $\mathbf{D}$)
  $\mathbf{A} \leftarrow \beta_i \mathbf{A} + \mathbf{x}_i {\mathbf{x}_i}^*$
  $\mathbf{B} \leftarrow \beta_i \mathbf{B} + \mathbf{y}_i {\mathbf{x}_i}^*$
  $\mathbf{D}$ = dict_update($\mathbf{A}$, $\mathbf{B}$)
  normalize($\mathbf{D}$)
end for

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Fixed points of dictionary learning algorithms [Mailhé13]

Consider K-SVD, MOD and Olshausen-Field in a fixed support context.

- Olshausen-Field [Olshausen97]: fixed step gradient descent.
- MOD [Engan99]: least-squares dicitonary update (pseudo-inverse).
- K-SVD [Aharon05]: joint atom/coefficient update by SVD.

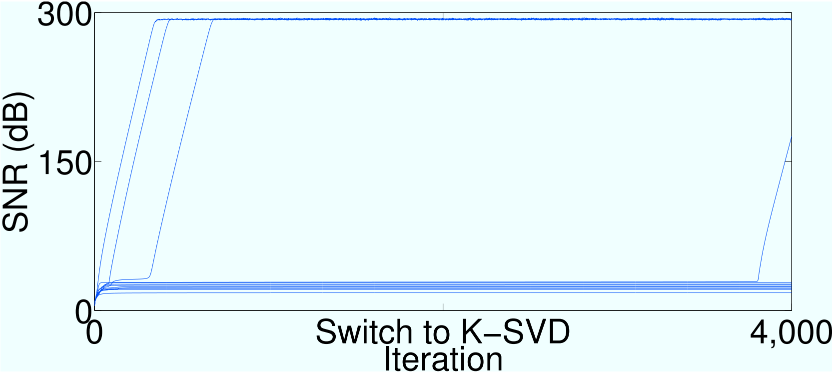
+ least-squares coefficients update.

Theorem (Mailhé13)

The set of the fixed points of K-SVD with an oracle support is strictly included in the set of the fixed points of MOD and gradient-based methods with an oracle support.

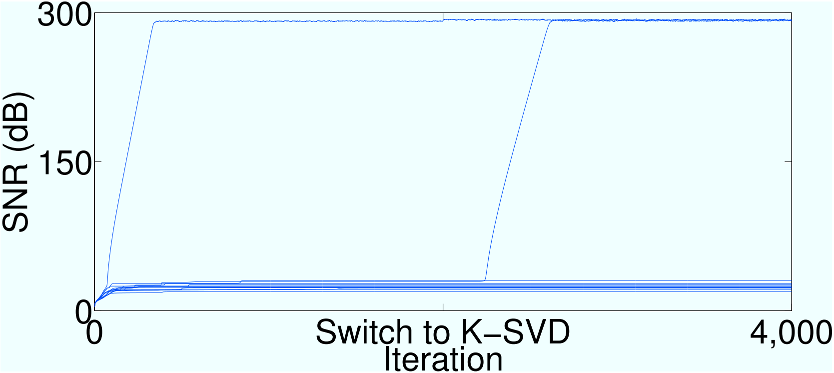
Can we use Olshausen-Fields or MOD to initialize K-SVD?

K-SVD with data initialization



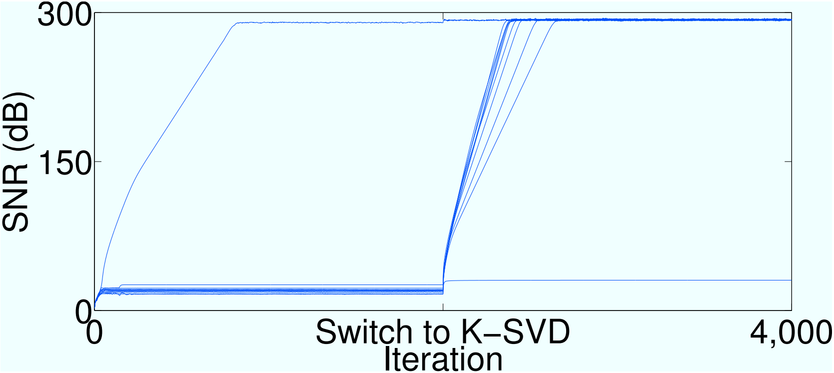
- 20% success
- Some very long plateaux

MOD, then K-SVD



- 4 % success
- Lots of plateaux

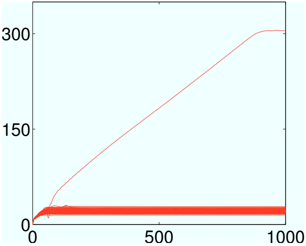
Olshausen-Field, then K-SVD



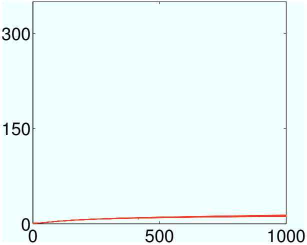
- 98 % success
- Some non-monotonicities: was the step size too large?

Goldilocks and the fixed step gradient descent

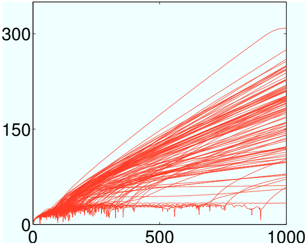
Let α be the step size.



$\alpha = 0.1$: too large :-(



$\alpha = 0.01$: too small :-(



$\alpha = 0.05$: just right :-)

- With the right step, gradient descent outperforms both MOD and K-SVD
- The "right" step must be larger than the optimal step to avoid local minima
- Can we estimate the step automatically?

Large step Gradient "Descent" (LGD) [Mailhé12]

- Maximal exploration principle:

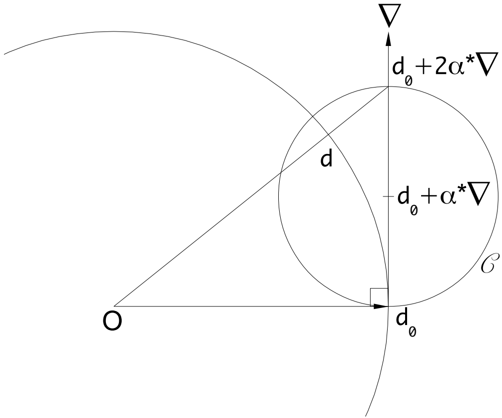
$$\mathbf{d} \leftarrow \underset{\mathbf{d}}{\text{argmax}} \; \|\mathbf{d} - \mathbf{d}_0\|_2^2$$
$$\text{s. t. } f(\mathbf{D}, \mathbf{X}) \leq f(\mathbf{D}_0, \mathbf{X})$$

- Gradient "descent" update with twice the optimal step size:

$$\mathbf{d} \leftarrow \mathbf{d} + \frac{2}{\|\mathbf{x}^{\mathbf{d}}\|_2^2} \mathbf{R}\mathbf{x}^{\mathbf{d}*}$$

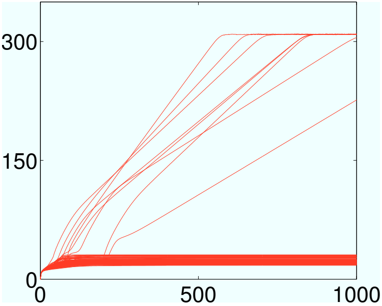
- Followed by renormalization.

Monotonicity proof sketch

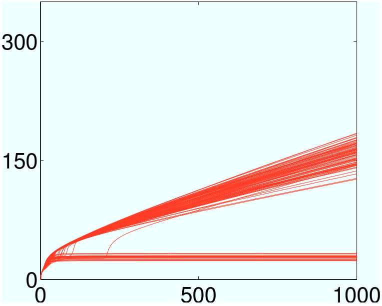


- With OMP, the gradient is orthogonal to the atom.
- The atom level set is circular.
- Normalization strictly decreases the error.

Results



Optimal step: 8% success



LGD: 88% success

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Shift-invariant dictionary learning

- Training data: one long signal $\mathbf{y}$ of length L.
- $\mathbf{D}$ of size $N \times M$ with $N \ll L$.
- $\mathcal{T} = \{\mathbf{T}_t \mid t \in [1, L]\}$

$$\mathbf{T}_t = \begin{pmatrix} \mathbf{0}_{t \times N} \\ \mathbf{Id}_N \\ \mathbf{0} \end{pmatrix}$$

- Learning problem:

$$\min_{\mathbf{D},\mathbf{X}} \left\| \mathbf{y} - \sum_{t=1}^{L} \mathbf{T}_t \mathbf{D} \mathbf{x}_t \right\|_2^2$$

$$\text{s. t. } \sum_{t=1}^{L} \|\mathbf{x}_t\|_0 \leq K \text{ and } \|\mathbf{d}\|_2 = 1, \forall \mathbf{d} \in \mathbf{D}.$$

Shift-invariant dictionary learning

- Sparse decomposition: the dictionary structure allows for faster implementations [Mallat93, Krstulovic05, Mailhé11]
- Dictionary update:
  - the gradient is still known [Blumensath06, Mailhé08]:

$$\nabla_{\mathbf{D}} = -2 \sum_{t=1}^{L} \mathbf{T}_t^* \mathbf{r} \mathbf{x}_t^*$$

  - closed form solution for one atom with fixed coefficients [Skretting06]:

$$\mathbf{T}_{\mathbf{d}} = \sum_{t=1}^{L} \mathbf{T}_t x_{t,\mathbf{d}} \qquad \mathbf{d} \leftarrow \mathbf{d} + \mathbf{T}_{\mathbf{d}}^{\dagger} \mathbf{r}$$

  - no closed form solution for K-SVD and MOD: overlaps between different shifts of the same atom invalidate the standard equations [Mailhé08].

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Learning low-coherence dictionaries

- Coherence

$$\mu(\mathbf{D}) = \max_{(\mathbf{d}_i,\mathbf{d}_j)\in\mathbf{D}^2, i\neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|$$

- Hard formulation (see [Ramirez09] for soft version):

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{DX}\|_F^2 \\ &\text{s.t. } \|\mathbf{x_y}\|_0 \leq K, \forall \mathbf{y} \in \mathbf{Y} \\ &\text{and } \mu(\mathbf{D}) \leq \bar{\mu} \text{ and } \|\mathbf{d}\|_2 = 1, \forall \mathbf{d} \in \mathbf{D} \end{aligned}$$

- Sparse approximation: same as before!

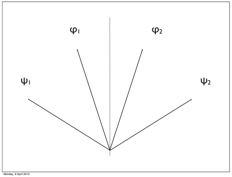
INK-SVD [Mailhé12-2] and IPR [Barchiesi13]

- Principle: add a dictionary decorrelation step to the learning, after the dictionary update.
- Decorrelation: projection on the (non-convex) set of low coherence dictionaries:

$$min_{\mathbf{D}} \|\mathbf{D} - \mathbf{D}_0\|_F^2$$
$$\text{s. t. } \mu(\mathbf{D}) \leq \bar{\mu} \text{ and } \|\mathbf{d}\|_2 = 1, \forall \mathbf{d} \in \mathbf{D}.$$

INK-SVD decorrelation

- Decorrelate atoms pair by pair.
- For a pair $(\mathbf{d}_1, \mathbf{d}_2)$, the projection (ψ_1, ψ_2) is the symmetric rotation of the atoms.



- Disjoint pairs can be decorrelated in parallel.

while $\mu(\mathbf{D}) > \bar{\mu}$ **do**
  $E =$ disjoint pairs in $\mathbf{D}$ with correlation higher than $\bar{\mu}$
  for $\forall(\mathbf{d}_i, \mathbf{d}_j) \in E$ **do**
    decorrelate_pair $(\mathbf{d}_i, \mathbf{d}_j)$
  end for
end while

IPR decorrelation

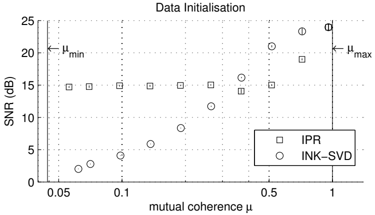
- Decorrelation in 2 steps:
  - decorrelate the Gram matrix $\mathbf{D}_0^*\mathbf{D}_0$,
  - factorize it back.
- Gram matrix decorrelation:
  - enforce low coherence and normalization: threshold the off-diagonal terms to $\bar{\mu}$ and the diagonal terms to 1,
  - enforce rank N s.d.p.: keep the N largest positive eigenvalues only.
- Factorization: find one factorization $\mathbf{D}_1$ and rotate it to minimize the error:

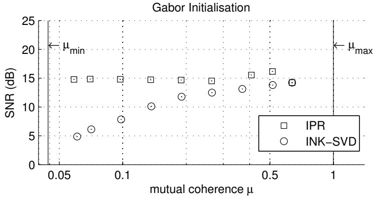
$$\mathbf{W} = \min_{\mathbf{W}\in\mathcal{O}(N)} \|\mathbf{Y} - \mathbf{W}\mathbf{D}_1\mathbf{X}\|_F^2$$

- Closed form solution:

$$\mathbf{D}_1\mathbf{X}\mathbf{Y}^* = \mathbf{U}\Delta\mathbf{V}^*$$
$$\mathbf{W} = \mathbf{V}\mathbf{U}^*$$

Results





Dictionary learning typically learns coherent dictionaries, even when there are much less coherent ones with the same error.

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Dictionary learning software: SMALLbox [Damnjanovic10]

SMALLbox is a dictionary learning benchmarking toolbox proposing a common API for dictionary learning problems, a few implementations and wrappers to third-party toolboxes.

- Coded in MATLAB
- Separation between problems and algorithms
- Integration of third-party code
- Add-on structure to plug more problems and algorithms

`http://code.soundsoftware.ac.uk/projects/smallbox`

Workflow

Problem creation:

- `create_problem`: preprocess signals to form a training set

Sparse representation:

- `SMALL_init_solver`: create a sparse solver structure
- `SMALL_solve(problem, solver)`: apply a solver to a problem

Dictionary learning:

- `SMALL_init_DL`: create a dictionary learning algorithm structure
- `SMALL_learn(problem, DL)`: apply a dictionary learning algorithm to a problem

The final signal reconstruction is called automatically by `SMALL_solve` and `SMALL_learn`.

APIs

- `problem`:
  - `A`: the (initial) dictionary
  - `b`: the signal(s)
  - `@reconstruct`: the synthesis function from the sparse coefficients to the signal
  - `p`: the number of atoms to learn
- `solver`:
  - `toolbox`: the toolbox name
  - `name`: the algorithm name in `toolbox`
  - `param`: a structure of parameters
  - `solution`: the output sparse coefficients
  - `reconstructed`: the output reconstructed signal
- `DL`:
  - `toolbox`: the toolbox name
  - `name`: the algorithm name in `toolbox`
  - `param`: a structure of parameters
  - `D`: the learnt dictionary

Outline of the second part

1. Global optimization ideas for dictionary learning
   - Online methods
   - Step size influence, LGD
2. Structured dictionary learning
   - Shift-invariant dictionary learning
   - Low-coherence dictionaries
3. Dictionary learning software: SMALLbox
   - Overview
   - Toolbox contents

Problems

In SMALLbox:

- Music transcription
- Audio declipping
- Audio denoising
- Image denoising

Third party:

- Sparco `http://www.cs.ubc.ca/labs/scl/sparco/`

Sparse solvers

In SMALLbox:

- MP
- OMP for Gabor dictionaries
- CGP

Third-party:

- Sparselab (ℓ_1, IRLS, greedy) `http://sparselab.stanford.edu/`
- SPGL1 (ℓ_1, group sparsity) `http://www.cs.ubc.ca/~mpf/spgl1/`
- Sparsify (greedy, IHTs) `http://users.fmrib.ox.ac.uk/~tblumens/sparsify/sparsify.html`
- GPSR (ℓ_1) `http://www.lx.it.pt/~mtf/GPSR/`
- Alps (IHTs) `http://lions.epfl.ch/ALPS`

General convex optimization toolboxes

- CVX `http://cvxr.com/cvx/`
- UNLocBox `http://unlocbox.sourceforge.net/`

Dictionary learning algorithms

In SMALLbox:

- twoStepDL: gradient descent (Olshausen-Fields, LGD), MOD, K-SVD, INK-SVD, with a modular sparse solver choice
- Recursive Least Squares (RLS)

Third-party:

- KSVD-box: KSVD, KSVDS (double sparsity) `http://www.cs.technion.ac.il/~ronrubin/software.html`
- SPAMS (Online Dictionary Learning + structure) `http://spams-devel.gforge.inria.fr/`

Add-ons

- Create a new problem: just write the `create_problem` and `reconstruct` functions.
- New solvers/DL algorithms must be registered so that `SMALL_solve` and `SMALL_learn` find them. This is done by editing the `SMALL_solve_config_local.m` and `SMALL_learn_config_local.m` files.

References

[Aharon] Aharon, M., Elad, M., & Bruckstein, A. (2005). K-SVD: Design of dictionaries for sparse representation. Proceedings of SPARS, 5, 9-12.

[Barchiesi13] Barchiesi, D., & Plumbley, M. D. (2013). Learning incoherent dictionaries for sparse approximation using iterative projections and rotations. IEEE Transactions on Signal Processing, 61(8), 2055-2065.

[Blumensath06] Blumensath, T., & Davies, M. (2006). Sparse and shift-invariant representations of music. Audio, Speech, and Language Processing, IEEE Transactions on, 14(1), 50-57.

[Damnjanovic10] Damnjanovic, I., Davies, M. E., & Plumbley, M. D. (2010). SMALLbox-an evaluation framework for sparse representations and dictionary learning algorithms. In Latent Variable Analysis and Signal Separation (pp. 418-425). Springer Berlin Heidelberg.

[Engan99] Engan, K., Aase, S. O., & Hakon Husoy, J. (1999). Method of optimal directions for frame design. In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on (Vol. 5, pp. 2443-2446). IEEE.

[Krstulovic05] Krstulovic, S., & Gribonval, R. (2006, May). MPTK: Matching pursuit made tractable. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Vol. 3, pp. III-III). IEEE.

[Mailhé08] Mailhé, B., Lesage, S., Gribonval, R., Bimbot, F., & Vandergheynst, P. (2008). Shift-invariant dictionary learning for sparse representations: extending K-SVD. In 16th EUropean SIgnal Processing COnference (EUSIPCO'08).

[Mailhé11] Mailhé, B., Gribonval, R., Vandergheynst, P., & Bimbot, F. (2011). Fast orthogonal sparse approximation algorithms over local dictionaries. Signal Processing, 91(12), 2822-2835.

References

[Mailhé12] Mailhé, B., & Plumbley, M. D. (2012). Dictionary learning with large step gradient descent for sparse representations. In Latent Variable Analysis and Signal Separation (pp. 231-238). Springer Berlin Heidelberg.

[Mailhé12-2] Mailhé, B., Barchiesi, D., & Plumbley, M. D. (2012, March). INK-SVD: Learning incoherent dictionaries for sparse representations. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 3573-3576). IEEE.

[Mailhé13] Mailhé, B., & Plumbley, M. D. (2013). Fixed points of dictionary learning algorithms for sparse representations. Submitted to IEEE Transactions on Information Theory.

[Mairal10] Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research, 11, 19-60.

[Mallat93] Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. Signal Processing, IEEE Transactions on, 41(12), 3397-3415.

[Olshausen97] Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by VI?. Vision research, 37(23), 3311-3326.

[Skretting06] Skretting, K., Husøy, J. H., & Aase, S. O. (2006). General design algorithm for sparse frame expansions. Signal processing, 86(1), 117-126.

[Skretting10] Skretting, K., & Engan, K. (2010). Recursive least squares dictionary learning algorithm. Signal Processing, IEEE Transactions on, 58(4), 2121-2130.

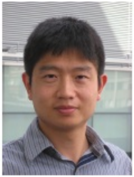
Dictionary Learning for Sparse Representations
Algorithms and Applications

Wei Dai, Boris Mailhé, & Wenwu Wang

Imperial College London
Queen Mary University of London
University of Surrey

May 2013

The Speakers



Dr. Wei Dai, Lecturer
Electrical and Electronic Engineering
Imperial College London
wei.dai1@imperial.ac.uk

Dr. Boris Mailhe, Postdoc RA
School of Electric Engineering and Computer Science
Queen Mary University of London
boris.mailhe@eecs.qmul.ac.uk

Dr. Wenwu Wang, Senior Lecturer
Department of Electronic Engineering
University of Surrey
w.wang@surrey.ac.uk

Acknowledgement of Financial Supports

- Engineering and Physical Science Research Council [Grant nos EP/H012842/1 & EP/H050000/1] (UK)
- Ministry of Defence (UK)
- Defence Science and Technology Laboratory (UK)
- MoD UDRC in Signal Processing (UK)
- BBC Audio Research Partnership (UK)
- China Scholarship Council (China)
- Centre for Vision Speech and Signal Processing (Surrey, UK)

Acknowledgement of Co-Workers

- University of Surrey
  - Josef Kittler
  - Philip Jackson
  - Mark Barnard
  - Tao Xu
  - Qingju Liu
  - Piotr Koniusz
- Loughborough University
  - Jonathon Chambers
  - Syed Mohsen Naqvi
- Imperial College London
  - Wei Dai
  - Guangyu Zhou
  - Xiaochen Zhao

Outline for the Third Part

1. Underdetermined blind speech separation Xu, et al. 2013; Dai, et al., 2012
2. Image separation and denoising Zhao, et al., 2013; Dai, et al., 2012
3. Audio-visual source separation Liu, et al., 2012; Q. Liu, et al., 2013
4. Multi-speaker tracking Barnard, et al., 2012; Barnard, et al., 2013

Underdetermined Blind Speech Separation (BSS)

- Instantaneous noiseless BSS model:

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{S}$$

where both the mixing matrix $\boldsymbol{A}$ and source signals $\boldsymbol{S}$ are unknown:

- Expanded form:

$$\begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_M \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_N \end{pmatrix}$$

- Underdetermined BSS:
  - when $M < N$, e.g. four sources and two mixtures.

Reformulating Underdetermined BSS

- Interpretation: Xu and Wang, 2009, 2010, 2011

$$\underbrace{\begin{pmatrix} z_1(1) \\ \vdots \\ z_1(T) \\ \vdots \\ \vdots \\ z_M(1) \\ \vdots \\ z_M(T) \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} \Lambda_{11} & \cdots & \Lambda_{1N} \\ \vdots & \ddots & \vdots \\ \Lambda_{M1} & \cdots & \Lambda_{MN} \end{pmatrix}}_{\mathbf{M}} \underbrace{\begin{pmatrix} s_1(1) \\ \vdots \\ s_1(T) \\ \vdots \\ \vdots \\ s_N(1) \\ \vdots \\ s_N(T) \end{pmatrix}}_{\mathbf{f}}$$

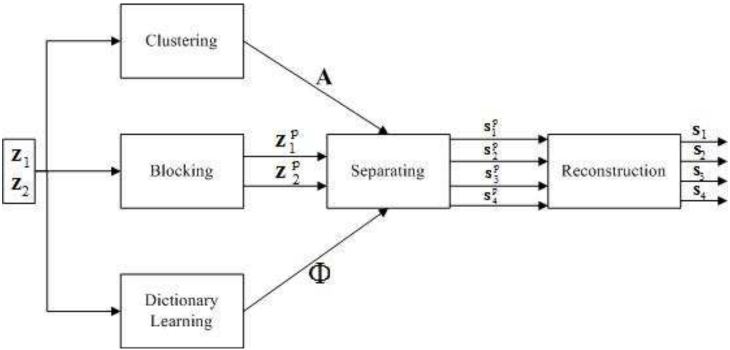
- Links to sparse signal recovery:

$$\mathbf{b} = \mathbf{M}\boldsymbol{\Phi}\mathbf{y}$$

where $\boldsymbol{\Phi}$ is a dictionary to sparsify $\mathbf{f}$.

A Multi-Stage Algorithm for Underdetermined BSS

A typical two-mixture-four-source case:



Learning Dictionary from Data

- The dictionary can be learned from either sources (STD) or mixtures (MTD). Xu and Wang, 2011; Xu, et al., 2013
- Algorithms discussed in the previous two parts of this tutorial, such as K-SVD and SimCO can be used to obtain the dictionaries. Aharon, 2006; Dai, et al., 2012

$$\underbrace{\begin{pmatrix} s_1(1)\\ \vdots \\ s_1(T)\\ s_2(1)\\ \vdots\\ s_2(T)\\ s_3(1)\\ \vdots\\ s_3(T)\\ s_4(1)\\ \vdots\\ s_4(T)\end{pmatrix}}_{\mathbf{f}} = \underbrace{\begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & D_3 & \\ & & & D_4\end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} y_1(1)\\ \vdots \\ y_1(T)\\ y_2(1)\\ \vdots\\ y_2(T)\\ y_3(1)\\ \vdots\\ y_3(T)\\ y_4(1)\\ \vdots\\ y_4(T)\end{pmatrix}}_{\mathbf{y}}$$

Experiments on TIMIT dataset

- A pool of 12 speech signals from the TIMIT database, sampled at 10 kHz, and trimmed to 5 seconds.
- In each random test, a group of 4 speech signals is randomly picked from the pool to generate the mixtures.
- For each comparison, 50 random tests have been performed.
- Performance measured by SDR, SIR, and SAR. Vincent, et al., 2006

Results on TIMIT data

- Comparison between predefined v.s. learned dictionaries:

	STD	MTD	DCT	STFT	MDCT
SDR	7.85	5.32	6.87	6.00	5.14
SIR	12.43	8.94	10.86	9.37	9.33
SAR	10.36	8.80	9.86	10.19	8.58

- Comparison between SimCO, K-SVD and GAD:

	SimCO	K-SVD	GAD
SDR	5.32	3.99	2.93
SIR	8.94	6.25	6.19
SAR	8.80	9.35	7.08

Experiments on SiSEC 2008 data

- The sources are available for comparison, which are sampled at 16 kHz, with length 10 seconds.
- The method (Gowreesunker and Tewfik, 2008, 2009) whose results were reported in the evaluation campaign is used as a baseline. This algorithm uses peak picking on threshold histogram to estimate the mixing matrix and achieves separation using coefficient space partitioning with K-SVD trained dictionary.
- Following algorithms are used in each stage of our proposed multistage algorithm: K-means clustering for the estimation of the mixing matrix, BP for signal recovery, and SimCO trained dictionary using the MTD strategy, and blocking for improving computational efficiency.

Results on SiSEC 2008 data

- Male speech mixtures:

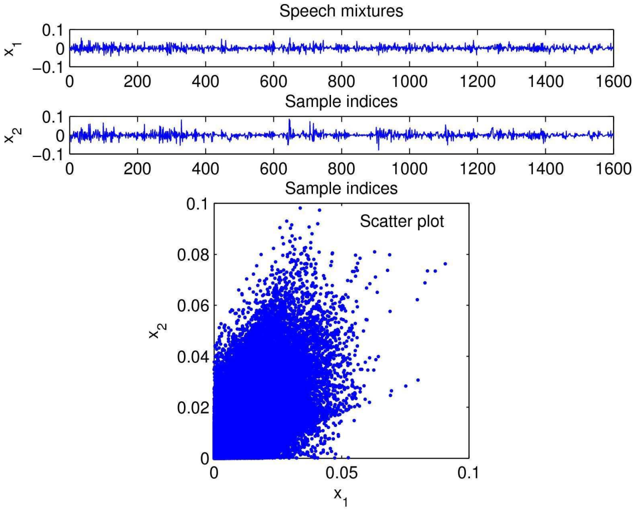
	Proposed method	Gowreesunker and Tewfik	STFT method
SDR	4.38	2.73	4.77
SIR	7.53	8.15	7.99
SAR	9.02	5.93	9.23

- Female speech mixtures:

	Proposed method	Gowreesunker and Tewfik	STFT method
SDR	4.04	3.80	4.51
SIR	6.19	8.58	6.86
SAR	9.73	6.60	9.78

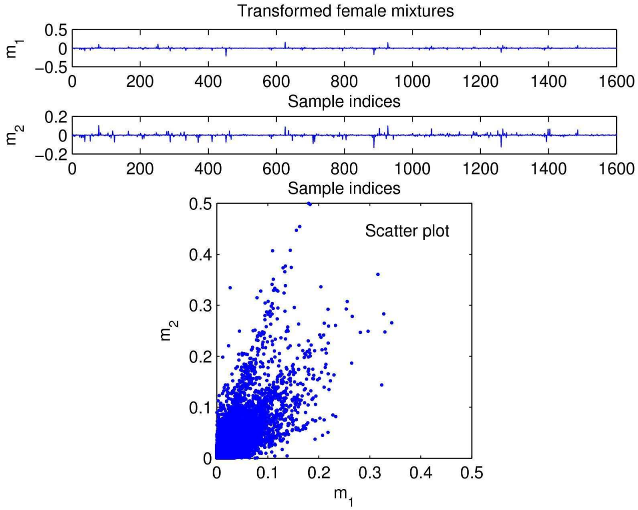
Scatter Plots

Mixtures:



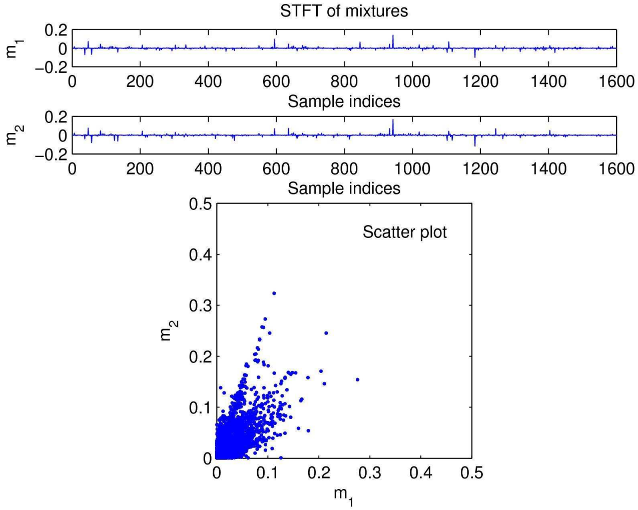
Scatter Plots

Transformed coefficients using SimCO:



Scatter Plots

Transformed coefficients using STFT:



Sound Demonstrations

- Two speech mixtures (x1, x2), four sources (s1-s4), and four estimated sources (es1-es4)

s1	s2	s3	s4
	x1	x2	
es1	es2	es3	es4

Image Separation and Denoising

- Cost function for joint dictionary learning and source separation:

$$\min_{\boldsymbol{A},\boldsymbol{S},\boldsymbol{D},\boldsymbol{X}} \lambda \left\|\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{S}\right\|_F^2 + \left\|\mathcal{P}^{\dagger}\left(\boldsymbol{D}\boldsymbol{X}\right) - \boldsymbol{S}^T\right\|_F^2, .$$

- Joint optimisation: Zhao, et al., 2013
  - Dictionary learning stage

$$\min_{\boldsymbol{D},\boldsymbol{X}} \left\|\boldsymbol{D}\boldsymbol{X} - \left(\mathcal{P}\boldsymbol{S}\right)^T\right\|_F^2,$$

  - Mixture learning stage

$$\min_{\boldsymbol{A},\boldsymbol{S}} \lambda \left\|\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{S}\right\|_F^2 + \left\|\mathcal{P}^{\dagger}\left(\boldsymbol{D}\boldsymbol{X}\right) - \boldsymbol{S}^T\right\|_F^2.$$

Proposed Joint DL and BSS Algorithm

Input: Observations $\boldsymbol{Z}$, patch size n, number of dictionary codewords d, regularization parameters λ and μ, and total number of iterations l_{max}.

Output: Dictionary $\boldsymbol{D}$, sparse coefficients $\boldsymbol{X}$, separated images $\boldsymbol{S}$, and estimated mixing matrix $\boldsymbol{A}$.

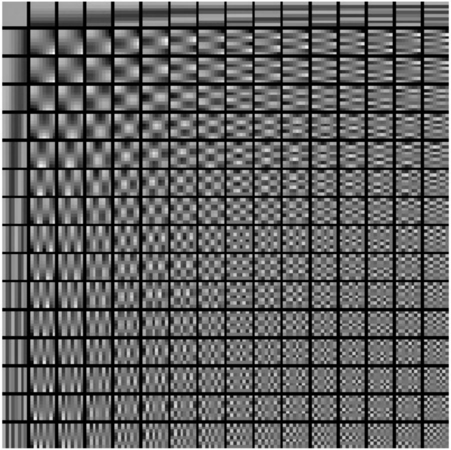
1. Set $\boldsymbol{D}$ to over-complete DCT dictionaries.
2. Set a random column-normalized matrix $\boldsymbol{A}$.
3. Compute $\boldsymbol{S} = \boldsymbol{A}^{\dagger}\boldsymbol{Z}$.
4. **For** $k = 1, 2, \ldots, l_{max}$ **repeat** $(6) - (10)$.
5. $\boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{X}} \left\| \boldsymbol{D}\boldsymbol{X} - (\mathcal{R}\boldsymbol{S})^T \right\|_F^2.$
6. $\boldsymbol{D}, \boldsymbol{X} \leftarrow \arg\min_{\boldsymbol{D}\in\mathcal{U}_{m,d}, \boldsymbol{X}\in\Omega} \left\| \boldsymbol{D}\boldsymbol{X} - (\mathcal{R}\boldsymbol{S})^T \right\|_F^2 + \mu \left\| \boldsymbol{X} \right\|_F^2.$
7. Let $\tilde{\boldsymbol{Z}} = \begin{bmatrix} \sqrt{\lambda}\boldsymbol{Z}^T & \boldsymbol{R}^T \end{bmatrix}^T$, $\tilde{\boldsymbol{A}} = \begin{bmatrix} \sqrt{\lambda}\boldsymbol{A}^T & \boldsymbol{I} \end{bmatrix}^T$.
8. Compute$\boldsymbol{S} = \tilde{\boldsymbol{A}}^{\dagger}\tilde{\boldsymbol{Z}}$.
9. $\boldsymbol{A} \leftarrow \arg\min_{\boldsymbol{A}\in\mathcal{U}_{r,s}} \left\| \tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{A}}\boldsymbol{S} \right\|_F^2.$

Simulations

Mixtures:

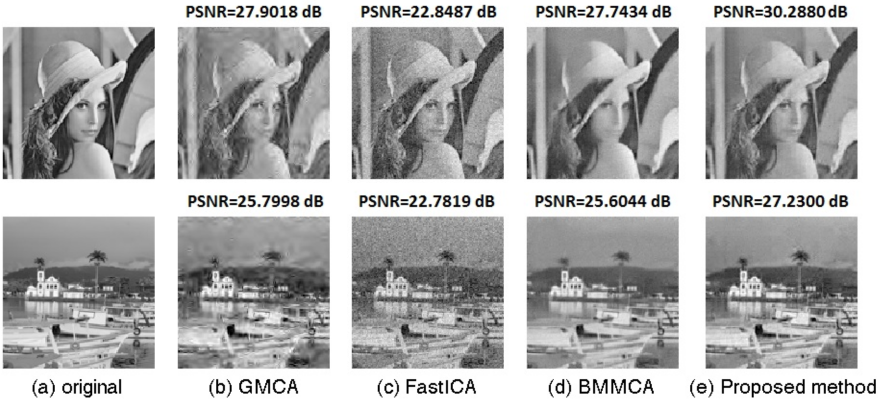
Simulations

Learned dictionary:



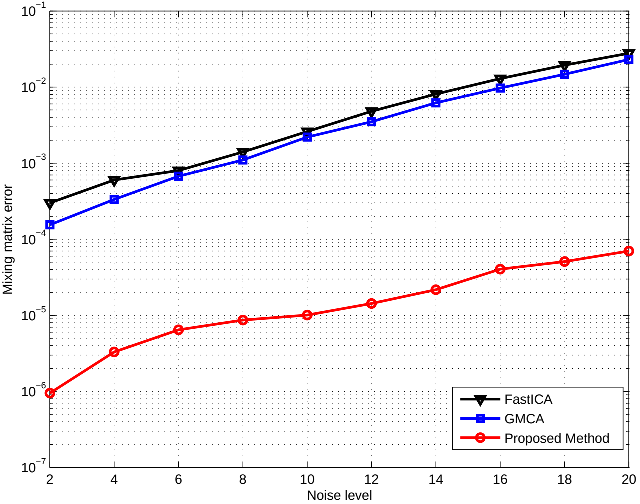
Simulations

Separation results: Zhao, et al., 2013; Elad, et al., 2006; Abolghasemi, et al., 2012



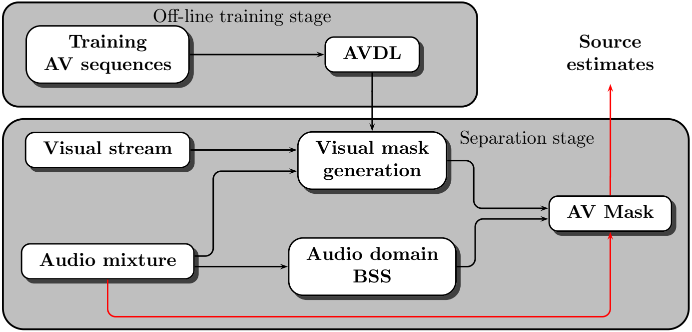
Simulations

Estimation errors:



Audio-visual Blind Source Separation (AV-BSS)

Source separation system based on audio-visual dictionary learning (AVDL): Liu, et al., 2012, Liu, et al., 2013



Audio-Visual Dictionary Learning

- Audio-visual sequence:

$$\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$$

$$\boldsymbol{\psi}^a = (\psi^a(m,\omega)) \in \mathbb{R}^{\tilde{M}\times\tilde{W}},$$
$$\boldsymbol{\psi}^v = (\psi^v(y,x,l)) \in \mathbb{R}^{\tilde{Y}\times\tilde{X}\times\tilde{L}}.$$

- Audio-visual atom:

$$\boldsymbol{\phi}_k^a \in \mathbb{R}^{\tilde{M}\times\tilde{W}},$$
$$\boldsymbol{\phi}_k^v = (\phi_k^v(y,x,l)) \in \mathbb{R}^{Y\times X\times L}$$

Signal Model

- Generative model: Liu, et al., 2013; Monaci, et al., 2007; Casanovas, et al., 2010

$$(\psi^a;\psi^v) \approx \sum_{k=1}^{K} \sum_{\breve{y}=1,\breve{x}=1,\breve{l}=1}^{Y_s,X_s,L_s} \begin{pmatrix} c_{k\breve{y}\breve{x}\breve{l}}\phi_k^a(m - m_{k\breve{y}\breve{x}\breve{l}}) \\ b_{k\breve{y}\breve{x}\breve{l}}\phi_k^v(y-\breve{y}, x-\breve{x}, l-\breve{l}) \end{pmatrix}$$

  where

$$m_{k\breve{y}\breve{x}\breve{l}} \in \left\{ \left\lceil (f_s^a/f_s^v)(\breve{l}-1) \right\rceil + 1, \ldots, \left\lceil (f_s^a/f_s^v)\breve{l} \right\rceil \right\}$$

- Parameters to learn:

$$\Omega = \{\mathbf{C}, \mathbf{B}, \mathbf{M}\},$$

  where

$$\mathbf{C} = (c_{k\breve{y}\breve{x}\breve{l}}), \mathbf{B} = (b_{k\breve{y}\breve{x}\breve{l}}), \mathbf{M} = (m_{k\breve{y}\breve{x}\breve{l}}) \in \mathbb{R}^{K\times Y_s \times X_s \times L_s}$$

Coding and Learning in AVDL

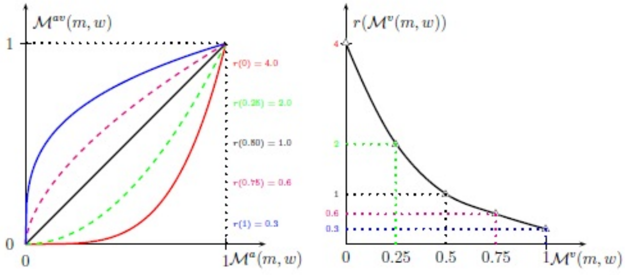
- Given a dictionary, sparse coding algorithms (such as matching pursuit) can be used to find the coding parameters, according to the signal model and a pre-defined matching criterion.
- Given the parameter set, the dictionary atoms are updated to fit the signal model. We used the K-SVD and K-means to update the audio and visual atoms respectively.

Integrating AVDL with Audio-Domain BSS

- Probabilistic time-frequency masking based binaural speech separation method is used to estimate a soft mask. Mandel, et al., 2010

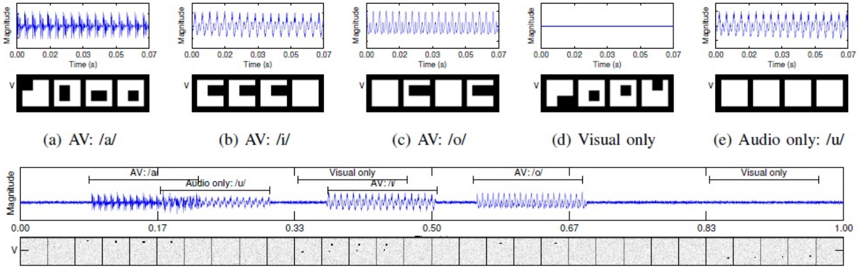
- This soft mask is then modified using the following power-law transformation where the visual information is incorporated:

$$\mathcal{M}^{av}(m,\omega) = \mathcal{M}^{a}(m,\omega)^{r(\mathcal{M}^{v}(m,\omega))},$$



Synthetic Examples

Original AV atoms and the synthesized AV sequence (with noise): Liu, et al., 2013; Monaci, et al., 2007

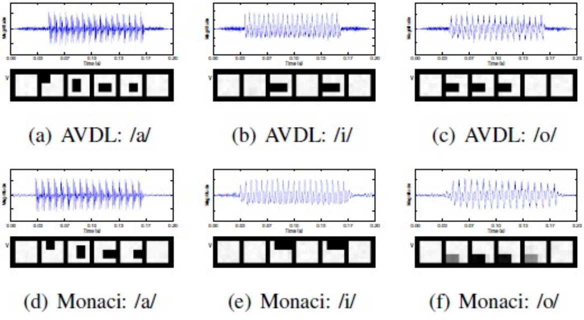


(a) AV: /a/ (b) AV: /i/ (c) AV: /o/ (d) Visual only (e) Audio only: /u/

(f) The generated AV synthetic sequence (only one second data is shown)

Synthetic Examples

Learned AV atoms (additive noise):

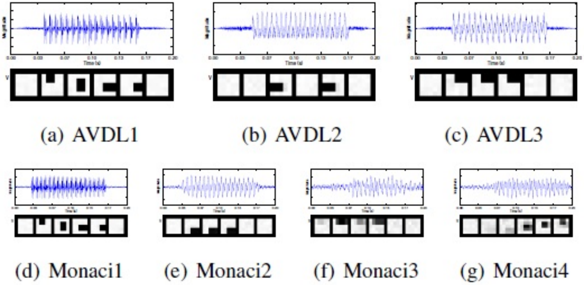


(a) AVDL: /a/ (b) AVDL: /i/ (c) AVDL: /o/

(d) Monaci: /a/ (e) Monaci: /i/ (f) Monaci: /o/

Synthetic Examples

Learned AV atoms (convolutive noise):

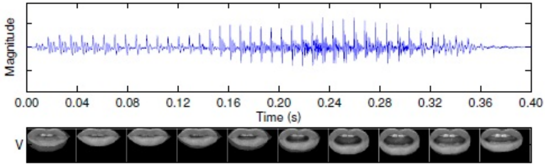


(a) AVDL1 (b) AVDL2 (c) AVDL3

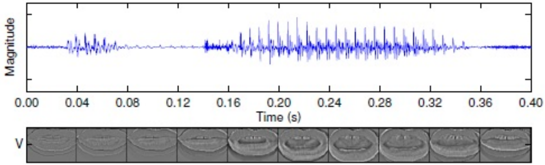
(d) Monaci1 (e) Monaci2 (f) Monaci3 (g) Monaci4

Real Speech Example

Learned AV atoms:



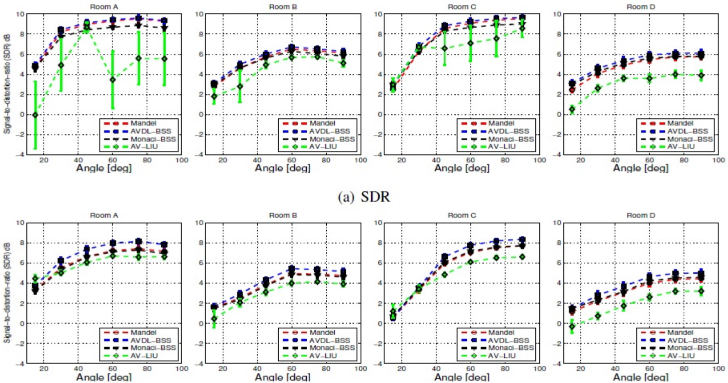
(a) AVDL



(b) Monaci

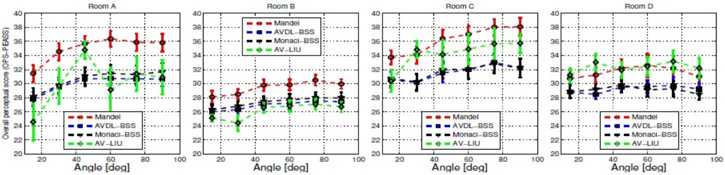
Separation Performance

SDR measurements: Liu, et al., 2012; Liu, et al., 2013

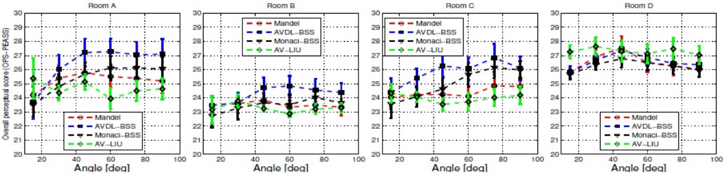


Separation Performance

PEASS measurements: Emiya, et al., 2011

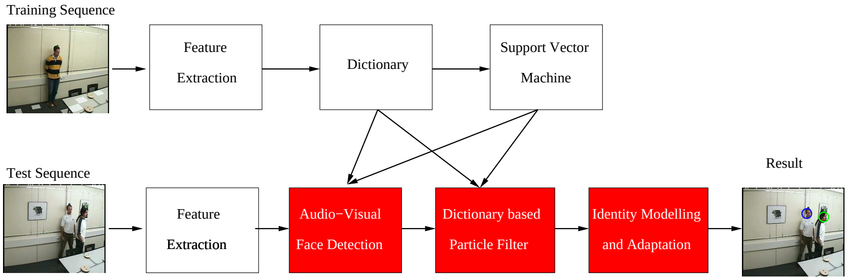


(a) PEASS



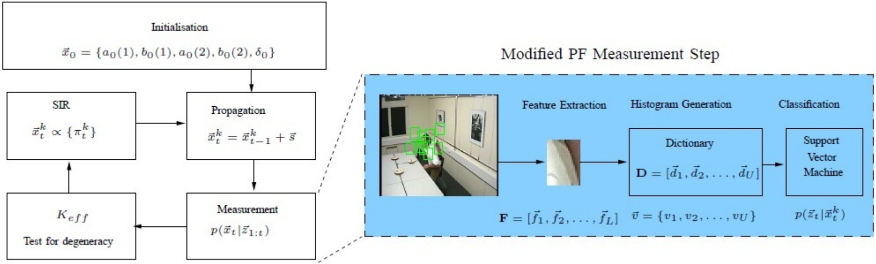
Multi-Speaker Tracking

Overall tracking system (including training and testing phases): Barnard, et al., 2012; Barnard, et al., 2013



Dictionary Based Particle Filter

Particle filter tracking algorithm with modified measurement step:



Modified Measurement Step

Using SVM to produce the likelihood:

Input: $\vec{z}_t$, K, L, U

Output: $p(\vec{z}_t|\vec{x}_t^k)$

for $k = 1$ to K **do**

Extract image patch at frame t according to $\{a_t^k(1), b_t^k(1), a_t^k(2), b_t^k(2)\}$;

Extract L features $\vec{f}_l$, $l = 1, ..., L$ from the image patch;

Create image patch representation $\vec{v} = \{v_1, v_2, \ldots v_U\}$, where

$v_u = \max_l \varrho_u(\vec{f}_l), \ l = 1, \ldots, L;$

Classify each image patch using SVM classifier to produce the likelihood $p(\vec{z}_t|\vec{x}_t^k)$.

end for

Dictionary Construction

- Dictionary construction can be regarded as a density estimation problem using a Gaussian mixture model (GMM) via the optimation of the following likelihood function:

  $$\Lambda(\mathcal{X};\theta) = \prod_{l=1}^{\bar{L}} \sum_{u=1}^{U} \omega_u g(\vec{f}_l;\vec{m}_u,\vec{\sigma}_u),$$

  where

  $$g(\vec{f}_l;\vec{m}_u,\vec{\sigma}_u) = ([(2\pi)^M \cdot |\mathbf{\Sigma}_u|]^{-\frac{1}{2}})exp(-\frac{1}{2}(\vec{f}_l - \vec{m}_u)^T \mathbf{\Sigma}_u^{-1}(\vec{f}_l - \vec{m}_u)),$$

- The parameters of the GMM can be estimated e.g. using an expectation maximisation (EM) algorithm. In our work, the means of the Gaussian mixtures is obtained by the k-means clustering.

Histogram Generation (Coding)

- Hard assignment (HA):

$$v_u = \frac{1}{L}\sum_{l=1}^{L} \begin{cases} 1 & \text{if } \vec{d}_u = \underset{\vec{d}\in\mathbf{D}}{\arg\min}(\mathbb{E}(\vec{d},\vec{f}_l)) \\ 0 & \text{otherwise} \end{cases} .$$

- Soft assignment (SA): Koniusz, et al., 2013

$$v_u = \frac{1}{L}\sum_{l=1}^{L} \varrho_u(\vec{f}_l),$$

where

$$\varrho_u(\vec{f}_l) = \frac{\omega_u g(\vec{f}_l;\vec{m}_u,\vec{\sigma}_u)}{\sum_{u'=1}^{U}\omega_{u'} g(\vec{f}_l;\vec{m}_{u'},\vec{\sigma}_{u'})}.$$

Histogram Generation (Coding)

- Approximate locality constrained SA (LcSA):

$$\varrho_u(\vec{f}_l) = \begin{cases} \frac{g(\vec{f}_l;\vec{m}_u,\vec{\sigma})}{\sum_{\vec{m}_{u'}\in\mathbf{D}_l^c} g(\vec{f}_l;\vec{m}_{u'},\vec{\sigma})} & \text{if } \vec{m}_u \in \mathbf{D}_l^c \\ 0 & \text{otherwise} \end{cases}$$

where

$$\mathbf{D}_l^c = NN_{\mathbf{D}}\left(\vec{f}_l, c\right)$$

- Fast Hierarchical Nearest Neighbour Search (FHNN):

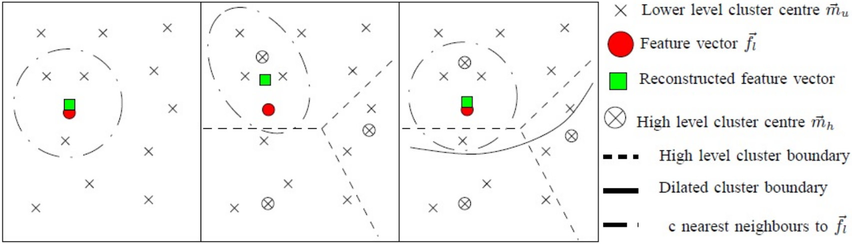
$$\mathbf{D}_l^c = NN_{\mathbf{D}_h}\left(\vec{f}_l, c\right)$$

where

$$\mathbf{D}_h = NN_{\mathbf{D}}\left(\vec{m}_h, \rho_h\right)$$

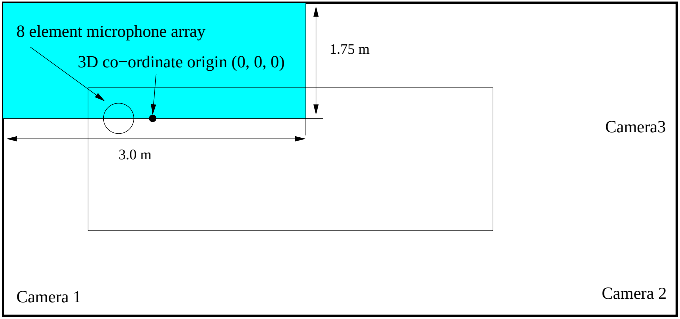
FHNN

Comparison among SA, LcSA and FHNN:



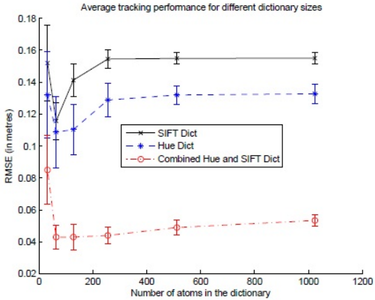
Experiments on AV16.3 dataset

Room layout (camera and microphone array set-up):

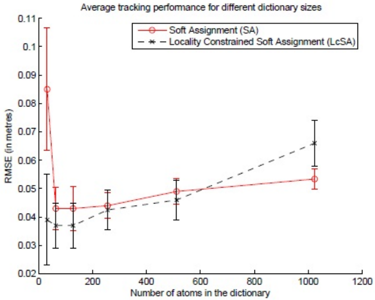


Tracking Errors v.s. Dictionary Size

Average results of 50 independent random tests measured on sequences 11, 12, and 15:



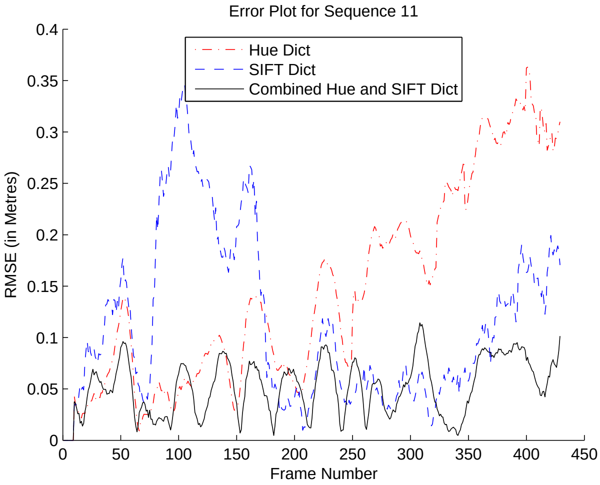
(a) Hue, SIFT and combined Hue and SIFT dictionaries using HA



(b) SA and LcSA dictionaries

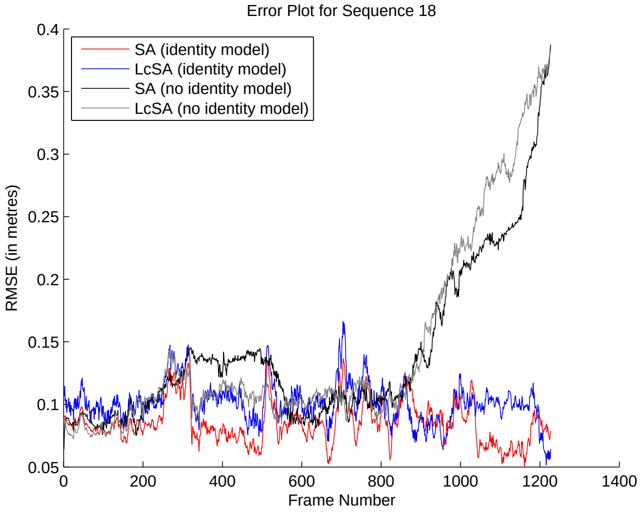
Single-Speaker Tracking Errors

RMSE (in meters) for sequence 11 (single speaker) over frames:



Multip-Speaker Tracking Errors

RMSE (in meters) for sequence 18 (two speakers) over frames:



Overall Tracking Errors for the Tested Sequences

Tracking errors measured over all the frames.

- Single speaker

Sequence	Hue Hist	SIFT Hist	Hue Dict	SIFT Dict	Combined Hue and SIFT Dict
Sequence 15	0.11	0.12	0.9	0.10	0.03
Sequence 11	0.13	0.15	0.10	0.10	0.05
Sequence 12	0.22	0.13	0.15	0.10	0.06

- Two speakers

Sequence	SA	LcSA	SA (with identity)	LcSA (with identity)
Sequence 18	0.19	0.17	0.13	0.10
Sequence 24	0.11	0.10	0.09	0.09

Video Demonstrations

- Single-speaker tracking

- Two-speaker tracking

A Summary

- Underdetermined blind speech separation
- Image separation and denoising
- Audio-visual source separation
- Multi-speaker tracking

Key References

- T. Xu, W. Wang, and W. Dai, "Sparse coding with adaptive dictionary learning for underdetermined blind speech separation", Speech Communication, vol. 55, no. 3, pp. 432-450, 2013.
  - http://personal.ee.surrey.ac.uk/Personal/W.Wang/papers/XuWD_SPCOM_2013.pdf
- W. Dai, T. Xu, and W. Wang. "Simultaneous codeword optimization (SimCO) for dictionary update and learning. " IEEE Transactions on Signal Processing, vol. 60, no. 12 (2012): 6340-6353.
  - http://personal.ee.surrey.ac.uk/Personal/W.Wang/papers/DaiXW_TSP_2012.pdf
- Q. Liu, W. Wang, P. Jackson, M. Barnard, J. Kittler, and J.A. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", IEEE Transactions on Signal Processing, 2013. (under review)
  - http://personal.ee.surrey.ac.uk/Personal/W.Wang/papers/LiuWBJKC_TSP_2013_ms.pdf
- M. Barnard, P.K. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J.A. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modelling", IEEE Transactions on Multimedia, 2013. (under review)
  - http://personal.ee.surrey.ac.uk/Personal/W.Wang/papers/Barnard_PWKNC_TMM_2013_ms.pdf

Key References

- M. Aharon, M. Elad, and A. Bruckstein. "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation." IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, 2006.
- M. Elad, and M. Aharon. "Image denoising via sparse and redundant representations over learned dictionaries." IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736–3745, 2006.
- G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval, "Learning multi-modal dictionaries," IEEE Trans. Image Process., vol. 16, no. 9, pp. 2272–2283, September 2007.
- V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Blind separation of image sources via adaptive dictionary learning," IEEE Trans. on Image Processing, vol. 21, no. 6, pp. 2921–2930 , 2012.
- M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," IEEE J. Selected Topics in Signal Process., vol. 5, no. 5, pp. 1025–1031, 2011.
- G. Monaci, P. Vandergheynst, and F. T. Sommer, "Learning bimodal structure in audio-visual data," IEEE Trans. Neural Netw., vol. 20, no. 12, pp. 1898–1910, December 2009.

Key References

- M. I. Mandel, R. J. Weiss, and D. Ellis, “Model-based expectation-maximization source separation and localization,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 382–394, February 2010.
- A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, “Blind audiovisual source separation based on sparse redundant representations,” IEEE Trans. Multimed., vol. 12, no. 5, pp. 358–371, August 2010.
- M. Barnard, W. Wang, J. Kittler, S.M.R. Naqvi, and J.A. Chambers, "A dictionary learning approach to tracking," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), Kyoto, Japan, March 25-30, 2012.
- Q. Liu, W. Wang, P. Jackson and M. Barnard, "Reverberant speech separation based on audio-visual dictionary learning and binaural cues", in Proc. IEEE Statistical Signal Processing Workshop (SSP 2012), pp. 664-667, Ann Arbor, USA, 5-8 August, 2012.
- W. Dai, T. Xu, and W. Wang, "Dictionary learning and update based on simultaneous codeword optimisation (SIMCO)," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012), Kyoto, Japan, March 25-30, 2012.
- Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve permutation problem in convolutive BSS," Signal Processing, vol. 92, vol. 8, pp. 1916-1927, 2012.

Key References

- W. Dai, T. Xu, and W. Wang, "Simultaneous codeword optimization (SimCO) for dictionary learning," in Proc. 49th Annual Allerton Conference on Communication, Control, and Computing (ALLERTON 2011), Monticello, Illinois, USA, Sept 28-30, 2011.
- B.V. Gowreesunker, A.H. Tewfik, "Blind source separation using monochannel overcomplete dictionaries," In: Proc. IEEE Internat. Conf. Acoustics, Speech and Signal Processing, pp. 33–36, 2008.
- B.V. Gowreesunker, A.H. Tewfik, "A novel subspace clustering method for dictionary design," In: Proc. Internat. Conf. on Independent Component Analysis and Signal Separation, pp. 34–41, 2009.
- X. Zhao, T. Xu, G. Zhou, W. Dai, and W. Wang, "Joint image separation and dictionary learning," in Proc. 18th Internat. Conf. on Digital Signal Processing, Greece, 2013 (to appear).
- P. Koniusz, F. Yan, and K. Mikolajczyk, "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection," Computer Vision and Image Understanding, vol. 117, no. 5, pp. 479 – 492, 2013.

Key References

- T. Xu and W. Wang, "Methods for learning adaptive dictionary for underdetermined speech separation," in Proc. IEEE 21st International Workshop on Machine Learning for Signal Processing (MLSP 2011), Beijing, China, Sept 18-21, 2011.
- T. Xu and W. Wang, "A block-based compressed sensing method for underdetermined blind speech separation incorporating binary mask," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), Dallas, Texas, USA, March 14-19, 2010.
- T. Xu and W. Wang, "A compressed sensing approach for underdetermined blind audio source separation with sparse representations," in Proc. IEEE International Workshop on Statistical Signal Processing (SSP 2009), Cardiff, UK, August 31-Sept 3, 2009.
- V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2046–2057, September 2011.
- E. Vincent, R. Gribonval, and C. Fe´votte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech Language Process., vol. 14, no. 4, pp. 1462–1469, 2006.