Dictionary Learning

Karin Schnass

Department of Mathematics
University of Innsbruck

karin.schnass@uibk.ac.at

Innsbruck
May 19, 2020

about me - education

April '04	Master in mathematics, University of Vienna, AT. Thesis: Gabor Multipliers - A Self-Contained Survey, Supervisor: Hans G. Feichtinger.
March '09	PhD in computer, communication and information sciences, Swiss Federal Institute of Technology Lausanne (EPFL), CH. Thesis: Sparsity & Dictionaries - Algorithms & Design, Advisor: Pierre Vandergheynst.

- '04 - '05 Leonardo da Vinci Industrial Internship at Philips Research, Eindhoven, NL.
- '05 - '09 Research Assistant, Signal Processing Laboratory 2, EPFL, CH.
- '09 - '10 Maternity leave.
- '10 - '11 Postdoc (part-time), RICAM, Linz, AT.
- '11 - '12 Maternity leave.
- '12 - '14 Erwin Schrödinger Research Fellow, Computer Vision Laboratory, University of Sassari, IT.
- June '14 FWF - START Prize: *Optimisation Principles, Models and Algorithms for Dictionary Learning*.
- since '15 University Assistant (since '19 assoc. Prof.), Department of Mathematics, University of Innsbruck.

Given N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \dots, y_N) \in \mathbb{R}^{d \times N}$
N large,

Given N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N) \in \mathbb{R}^{d \times N}$
N large,

find a decomposition into $\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^{d \times K}$, the dictionary, and sparse coefficients $X = (x_1, \ldots, x_N) \in \mathbb{R}^{K \times N}$,

$$Y \approx \Phi X \quad \text{where} \quad \|x_n\|_0 \leq S \ll d.$$

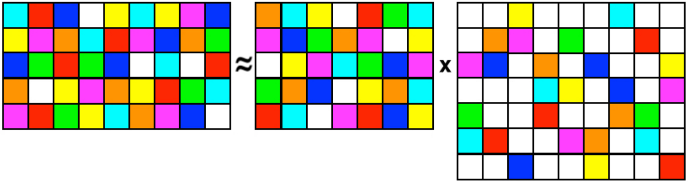
The columns ϕ_k are called atoms and normalised, $\|\phi_k\|_2 = 1$.

Given N vectors $y_n \in \mathbb{R}^d$
$Y = (y_1, \ldots, y_N) \in \mathbb{R}^{d \times N}$
N large,

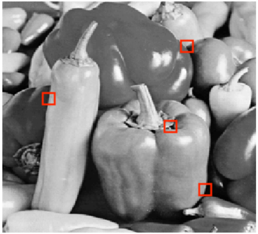
find a decomposition into $\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^{d \times K}$, the dictionary, and sparse coefficients $X = (x_1, \ldots, x_N) \in \mathbb{R}^{K \times N}$,

$$Y \approx \Phi X \quad \text{where} \quad \|x_n\|_0 \leq S \ll d.$$

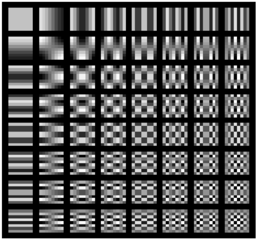
The columns ϕ_k are called atoms and normalised, $\|\phi_k\|_2 = 1$.



$K \ll N$ $\square \equiv 0$

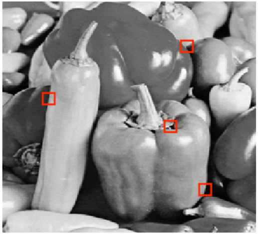


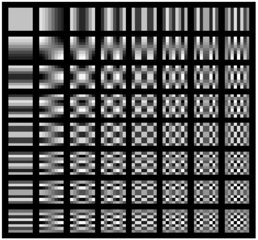
DCT-basis (jpg)

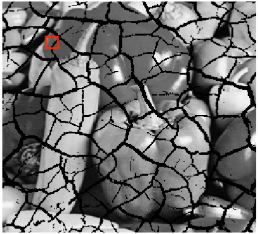


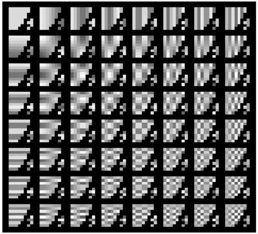
inpainting

DCT-basis (jpg)



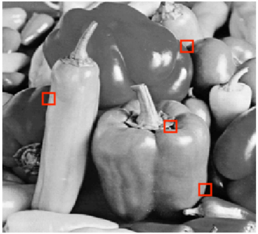


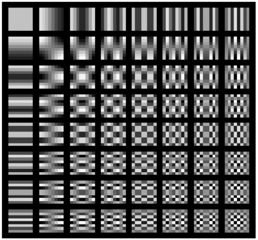


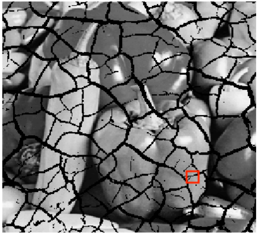


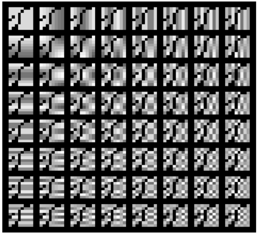
inpainting

DCT-basis (jpg)









Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

back to dictionary learning - alternating optimisation

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,. . .

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,...

back to dictionary learning - alternating optimisation

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

but it becomes nicer if we keep certain variables fixed,

back to dictionary learning - alternating optimisation

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

but it becomes nicer if we keep certain variables fixed, e.g.

$$\text{fix } \Psi : \quad \min_{X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2 = \sum_n \min_{\|x_n\|_0 \leq S} \|y_n - \Psi x_n\|_2^2.$$

$\Rightarrow$ N sparse approximation problems.

back to dictionary learning - alternating optimisation

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

but it becomes nicer if we keep certain variables fixed, e.g.

$$\text{fix } \Psi : \quad \min_{X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2 = \sum_n \min_{\|x_n\|_0 \leq S} \|y_n - \Psi x_n\|_2^2.$$

$\Rightarrow$ N sparse approximation problems.

$$\text{fix } X : \quad \arg\min_{\Psi \in \mathcal{D}_K} \|Y - \Psi X\|_F^2$$

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

but it becomes nicer if we keep certain variables fixed, e.g.

$$\text{fix } \Psi : \quad \min_{X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2 = \sum_n \min_{\|x_n\|_0 \leq S} \|y_n - \Psi x_n\|_2^2.$$

$\Rightarrow$ N sparse approximation problems.

$$\text{fix } X : \quad \arg \min_{\color{red}{\Psi \in \mathbb{R}^{d \times K}}} \|Y - \Psi X\|_F^2$$

back to dictionary learning - alternating optimisation

Given a sparsity level S and a dictionary size K, we try to find

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

non-linear,... non-convex,... non-trivial,...

but it becomes nicer if we keep certain variables fixed, e.g.

$$\text{fix } \Psi : \quad \min_{X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2 = \sum_n \min_{\|x_n\|_0 \le S} \|y_n - \Psi x_n\|_2^2.$$

$\Rightarrow$ N sparse approximation problems.

$$\text{fix } X : \quad \arg \min_{\Psi \in \mathbb{R}^{d \times K}} \|Y - \Psi X\|_F^2 = YX^T(XX^T)^{-1}$$

$\Rightarrow$ a least square problem & renormalisation.

sparse approximation

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2$$

sparse approximation

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Calculate* $x = \Phi^T y$.
- *Find the locations of the largest* S *entries of* x *in magnitude*

$$I = \operatorname{argmax}_{|J|=S} \|x_J\|_2^2$$

- *Set* $a = \Phi_I \Phi_I^T y$

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \le S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \le S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Calculate* $x = \Phi^T y$.
- *Find the locations of the largest* S *entries of* x *in magnitude*

$$I = \text{argmax}_{|J|=S} \|x_J\|_2^2$$

- *Set* $a = \Phi_I \Phi_I^\dagger y =: P(\Phi_I) y$

sparse approximation

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Calculate* $x = \Phi^T y$.
- *Find the locations of the largest* S *entries of* x *in magnitude*

$$I = \operatorname{argmax}_{|J|=S} \|x_J\|_2^2$$

- *Set* $a = \Phi_I \Phi_I^\dagger y =: P(\Phi_I) y$

If Φ is only a dictionary, this is called thresholding.

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Initialise* $I = \emptyset$, $r = y$, $a = 0$.
- *Repeat until* $|I| = S$
  - *Find* $i = \text{argmax}_j |\langle \phi_i, r \rangle|$
  - *Update* $I \leftarrow I \cup \{i\}$ *and* $r = y - \Phi_I \Phi_I^T y$.
- *Set* $a = \Phi_I \Phi_I^T y$

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Initialise* $I = \emptyset$, $r = y$, $a = 0$.
- *Repeat until* $|I| = S$
  - *Find* $i = \text{argmax}_j |\langle \phi_i, r \rangle|$
  - *Update* $I \leftarrow I \cup \{i\}$ *and* $r = y - \Phi_I \Phi_I^\dagger y = y - P(\Phi_I) y$.
- *Set* $a = \Phi_I \Phi_I^\dagger y = P(\Phi_I) y$

Given a dictionary Φ and a signal y we want to minimize

$$\min_{\|x\|_0 \leq S} \|y - \Phi x\|_2^2 \quad \Leftrightarrow \quad \min_{|I| \leq S} \|y - \Phi_I \Phi_I^\dagger y\|_2^2.$$

If Φ is an orthonormal basis, this is easy.

Algorithm

- *Initialise* $I = \emptyset$, $r = y$, $a = 0$.
- *Repeat until* $|I| = S$
  - *Find* $i = \operatorname{argmax}_j |\langle \phi_i, r\rangle|$
  - *Update* $I \leftarrow I \cup \{i\}$ *and* $r = y - \Phi_I \Phi_I^\dagger y = y - P(\Phi_I) y$.
- *Set* $a = \Phi_I \Phi_I^\dagger y = P(\Phi_I) y$

If Φ is only a dictionary, this is called Orthogonal Matching Pursuit.

MOD, K-SVD, ITKrM

Choose (be given) a sparsity level S a dictionary size K and

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

Choose (be given) a sparsity level S a dictionary size K and

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

Algorithm (MOD - Method of Optimal Directions)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Psi_{I_n}) y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *Calculate*

$$\bar{\Psi} = Y X^T (X X^T)^{-1}$$

- *Update:* $\psi_k \leftarrow \bar{\psi}_k / \|\bar{\psi}_k\|_2$.

MOD, K-SVD, ITKrM

Choose (be given) a sparsity level S a dictionary size K and

$$\min_{\Psi\in\mathcal{D}_K, X\in\mathcal{X}_S} \|Y-\Psi X\|_F^2$$

Algorithm (K-SVD)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Psi_{I_n})y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *For all k calculate*

$$R_k = \sum_{n: k\in I_n} [y_n - \Psi x_n + \psi_k x_n(k)][y_n - \Psi x_n + \psi_k x_n(k)]^T.$$

- *Update:* $\psi_k \leftarrow \arg\max_{\|v\|_2=1} \|R_k v\|_2$, *(via K SVDs).*

MOD, K-SVD, ITKrM

Choose (be given) a sparsity level S a dictionary size K and

$$\min_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \|Y - \Psi X\|_F^2$$

Algorithm (Iterative Thresholding and K residual means - ITKrM)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use thresholding to sparsely approximate y_n*

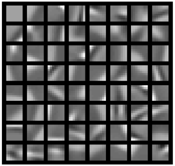
$$a_n = P(\Psi_{I_n}) y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^\dagger y, \quad x_n|_{I_n^c} = 0.$$

- *For all k calculate*

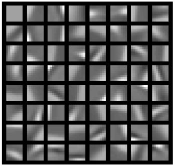
$$\bar{\psi}_k = \sum_{n: k \in I_n} \left[y_n - \Psi x_n + \psi_k \langle \psi_k, y_n \rangle \right] \cdot \mathrm{sign}(\langle \psi_k, y_n \rangle).$$

- *Update:* $\psi_k \leftarrow \bar{\psi}_k / \|\bar{\psi}_k\|_2$.

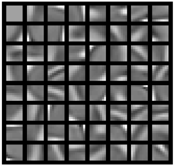
MOD



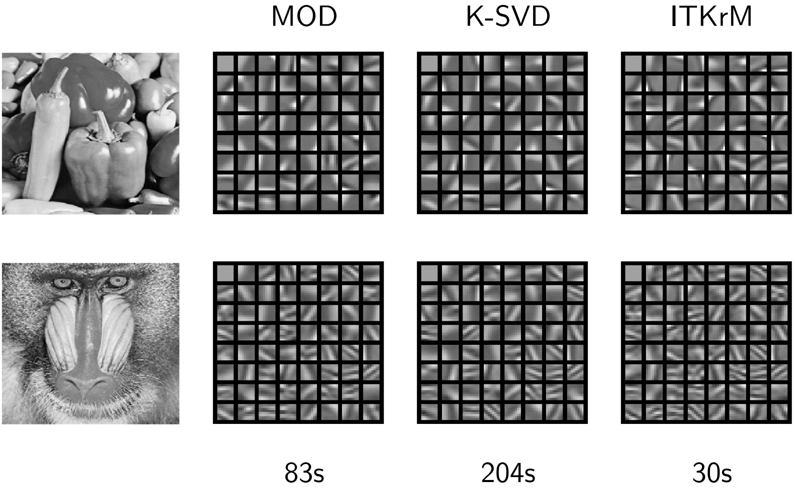
K-SVD

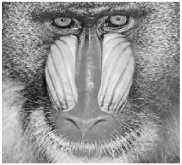


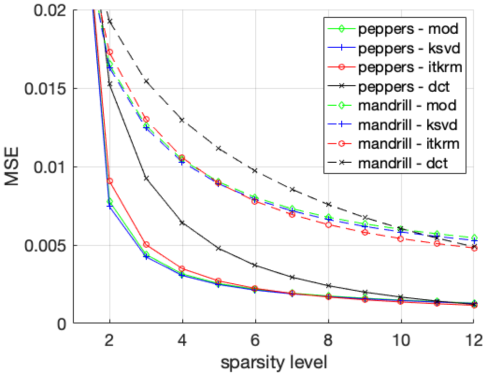
ITKrM



some learned dictionaries






0.02
0.015
0.01
0.005
0
MSE
2
4
6
8
10
12
sparsity level
peppers - mod
peppers - ksvd
peppers - itkrm
peppers - dct
mandrill - mod
mandrill - ksvd
mandrill - itkrm
mandrill - dct

how can we do math with this?

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist

how can we do math with this?

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist unless $K \approx e^d$, ($\equiv$ a packing problem).

how can we do math with this?

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist unless $K \approx e^d$, ($\equiv$ a packing problem).

A simple S-sparse model:
Fix $\Phi \in \mathcal{D}_K$ and coefficients c with $c_1 \geq c_2 \ldots \geq c_S > 0$ and $c_k = 0$ for $k > S$. Choose a permutation p of $\{1 \ldots K\}$ and signs $\sigma \in \{-1, 1\}^K$ uniformly at random and set

$$y = \sum_{i=1}^{S} c_i \sigma_i \phi_{p(i)} =: \Phi_I x_I \quad \text{with} \quad I = \{p(1), \ldots p(S)\} \qquad (1)$$

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist unless $K \approx e^d$, ($\equiv$ a packing problem).

A simple S-sparse model:
Fix $\Phi \in \mathcal{D}_K$ and coefficients c with $c_1 \geq c_2 \ldots \geq c_S > 0$ and $c_k = 0$ for $k > S$. Choose a permutation p of $\{1 \ldots K\}$ and signs $\sigma \in \{-1, 1\}^K$ uniformly at random and set

$$y = \sum_{i=1}^{S} c_i \sigma_i \phi_{p(i)} =: \Phi_I x_I \quad \text{with} \quad I = \{p(1), \ldots p(S)\} \qquad (1)$$

Question: Can an algorithm recover Φ given N samples $Y = (y_1, \ldots y_N)$ and a good/random/any initialisation?

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist unless $K \approx e^d$, ($\equiv$ a packing problem).

A simple S-sparse model:
Fix $\Phi \in \mathcal{D}_K$ and coefficients c with $c_1 \geq c_2 \ldots \geq c_S > 0$ and $c_k = 0$ for $k > S$. Choose a permutation p of $\{1 \ldots K\}$ and signs $\sigma \in \{-1, 1\}^K$ uniformly at random and set

$$y = \sum_{i=1}^{S} c_i \sigma_i \phi_{p(i)} =: \Phi_I x_I \quad \text{with} \quad I = \{p(1), \ldots p(S)\} \qquad (1)$$

Question: Can an algorithm recover Φ given N samples $Y = (y_1, \ldots y_N)$ and a good/random/any initialisation?

Quick Answers: Only up to signs and permutations

$$Y = \Phi X \quad \Rightarrow Y = \Phi DP \cdot PDX.$$

If $y \sim \mathcal{N}(0, \mathbb{I}_d)$ no algorithm will find a good dictionary, because it does not exist unless $K \approx e^d$, ($\equiv$ a packing problem).

A simple S-sparse model:
Fix $\Phi \in \mathcal{D}_K$ and coefficients c with $c_1 \geq c_2 \ldots \geq c_S > 0$ and $c_k = 0$ for $k > S$. Choose a permutation p of $\{1 \ldots K\}$ and signs $\sigma \in \{-1, 1\}^K$ uniformly at random and set

$$y = \sum_{i=1}^{S} c_i \sigma_i \phi_{p(i)} =: \Phi_I x_I \quad \text{with} \quad I = \{p(1), \ldots p(S)\} \qquad (1)$$

Question: Can an algorithm recover Φ given N samples $Y = (y_1, \ldots y_N)$ and a good/random/any initialisation?

Quick Answers: Not if

$$\mu(\Phi) := \max_{i \neq j} |\langle \phi_i, \phi_j \rangle| = 1.$$

ideally the generating dictionary is a fixed point

We know

$$Y = \Phi X \quad \text{with} \quad \|x_n\|_0 = S.$$

ideally the generating dictionary is a fixed point

We know

$$Y = \Phi X \quad \text{with} \quad \|x_n\|_0 = S.$$

Algorithm (MOD)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Psi_{I_n})y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *Calculate*

$$\bar{\Psi} = YX^T(XX^T)^{-1}$$

- *Update:* $\psi_k \leftarrow \bar{\psi}_k / \|\bar{\psi}_k\|_2$.

ideally the generating dictionary is a fixed point

We know

$$Y = \Phi X \quad \text{with} \quad \|x_n\|_0 = S.$$

Algorithm (MOD)

Given an input dictionary $\Psi = \Phi$ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Phi_{I_n})y_n = \Phi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Phi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *Calculate*

$$\bar{\Psi} = YX^T(XX^T)^{-1} = \Phi XX^T(XX^T)^{-1} = \Phi$$

- *Update:* $\psi_k \leftarrow \bar{\psi}_k / \|\bar{\psi}_k\|_2 = \phi_k$.

ideally the generating dictionary is a fixed point

We know

$$Y = \Phi X \quad \text{with} \quad \|x_n\|_0 = S.$$

Algorithm (K-SVD)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Psi_{I_n})y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *For all k calculate*

$$R_k = \sum_{n:k \in I_n} [y_n - \Psi x_n + \psi_k x_n(k)][y_n - \Psi x_n + \psi_k x_n(k)]^T.$$

- *Update:* $\psi_k \leftarrow \arg\max_{\|v\|_2=1} \|R_k v\|_2.$

ideally the generating dictionary is a fixed point

We know

$$Y = \Phi X \quad \text{with} \quad \|x_n\|_0 = S.$$

Algorithm (K-SVD)

Given an input dictionary $\Psi = \Phi$ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Phi_{I_n})y_n = \Phi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Phi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *For all k calculate*

$$R_k = \sum_{n:k \in I_n} [\phi_k x_n(k)][\phi_k x_n(k)]^T = \sum_{n:k \in I_n} x_n(k)^2 \phi_k \phi_k^T$$

- *Update:* $\psi_k \leftarrow \arg\max_{\|v\|_2=1} \|R_k v\|_2 = \phi_k$.

ITKrM is quite well understood theoretically

Theorem (M.C. Pali & K. S.)

Assume that the signals y_n follow model (1) for coefficients with gap $c(S+1)/c(S) \leq \gamma_{gap}$, dynamic sparse range $c(1)/c(S) \leq \gamma_{dyn}$, noise to coefficient ratio $\rho/c(S) \leq \gamma_{rho}$ and relative approximation error $\|c(\mathbb{S}^c)\|_2/c(1) \leq \gamma_{app} \leq \frac{12}{7}\log K$. Further, assume that the coherence and operator norm of the current dictionary estimate Ψ satisfy,

$$\mu(\Psi) \leq \frac{1}{20\log K} \quad \textit{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1.$$

If $d(\Psi,\Phi) \geq \frac{1}{32\sqrt{S}}$ but the cross Gram matrix $\Phi^\star\Psi$ is diagonally dominant in the sense that

$$\min_k |\langle \psi_k, \phi_k\rangle| \geq \max\Big\{ 8\,\gamma_{gap}\cdot \max_k |\langle \psi_k, \phi_k\rangle|,\; 40\,\gamma_{rho}\cdot\sqrt{\log K},\; 48\,\gamma_{dyn}\cdot\log K\cdot\mu(\Phi,\Psi),\; 14\,\gamma_{dyn}\cdot\sqrt{\|\Phi\|_{2,2}^2 S\log K/(K-S)}\Big\},$$

then one iteration of ITKrM using N training signals will reduce the distance by at least a factor $\kappa \leq 0.94$, meaning $d(\bar\Psi,\Phi) \leq 0.94\cdot d(\Psi,\Phi)$, except with probability

$$2K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon}{768K\max\{S,\|\Phi\|_{2,2}^2+1\}^{\frac{3}{2}}}\right) + 2K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon^2}{512K\max\{S,\|\Phi\|_{2,2}^2+1\}\left(1+d\rho^2\right)}\right).$$

MOD & K-SVD not so much

because we first need to understand OMP.

MOD & K-SVD not so much

because we first need to understand OMP.

Theorem (J. Tropp '04)

OMP will succeed for $y = \Phi_I x_I$, that is, recover any support I with $|I| = S$, if

$$2\mu S \leq 1.$$

Remember $\mu = \max_{i \neq j} |\langle \phi_j, \phi_i \rangle|$.

MOD & K-SVD not so much

because we first need to understand OMP.

Theorem (J. Tropp '04)

OMP will succeed for $y = \Phi_I x_I$, that is, recover any support I with $|I| = S$, if

$$2\mu S \leq 1.$$

Remember $\mu = \max_{i \neq j} |\langle \phi_j, \phi_i \rangle|$.

Proof idea:
OMP will succeed if for any $J \subset I$ with $J^c = I \backslash J$ the residual

$$r_J = y - P(\Phi_J)y = \Phi_{J^c} x_{J^c} - P(\Phi_J)\Phi_{J^c} x_{J^c}$$

satisfies $\max_{i \in I} |\langle \phi_i, r_J \rangle| > \max_{j \notin I} |\langle \phi_j, r_J \rangle|$.

MOD & K-SVD not so much

because we first need to understand OMP.

Theorem (J. Tropp '04)

OMP will succeed for $y = \Phi_I x_I$, that is, recover any support I with $|I| = S$, if

$$2\mu S \leq 1.$$

Remember $\mu = \max_{i \neq j} |\langle \phi_j, \phi_i \rangle|$.

Proof idea:
OMP will succeed if for any $J \subset I$ with $J^c = I \backslash J$ the residual

$$r_J = y - P(\Phi_J)y = \Phi_{J^c} x_{J^c} - P(\Phi_J)\Phi_{J^c} x_{J^c}$$

satisfies $\max_{i \in I} |\langle \phi_i, r_J \rangle| > \max_{j \notin I} |\langle \phi_j, r_J \rangle|$.

$$|\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx |x_i| \pm |\sum_{k \in J^c} x_k \langle \phi_i, \phi_k \rangle| \approx |x_i| \pm \|x_{J^c}\|_1 \cdot \mu$$

MOD & K-SVD not so much

because we first need to understand OMP.

Theorem (J. Tropp '04)

OMP will succeed for $y = \Phi_I x_I$, *that is, recover any support* I *with* $|I| = S$, *if*

$$2\mu S \leq 1.$$

Remember $\mu = \max_{i \neq j} |\langle \phi_j, \phi_i \rangle|$.

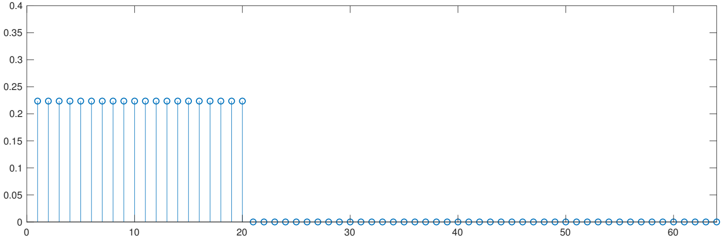
Proof idea:
OMP will succeed if for any $J \subset I$ with $J^c = I \backslash J$ the residual

$$r_J = y - P(\Phi_J)y = \Phi_{J^c} x_{J^c} - P(\Phi_J)\Phi_{J^c} x_{J^c}$$

satisfies $\max_{i \in I} |\langle \phi_i, r_J \rangle| > \max_{j \notin I} |\langle \phi_j, r_J \rangle|$.

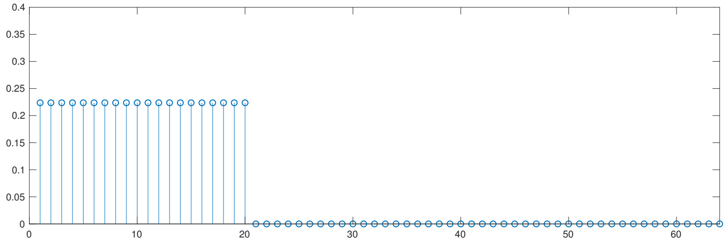
$$|\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx |x_i| \pm |\sum_{k \in J^c} x_k \langle \phi_i, \phi_k \rangle| \approx |x_i| \pm \|x_{J^c}\|_1 \cdot \mu$$

pessimistic but nothing we can do...

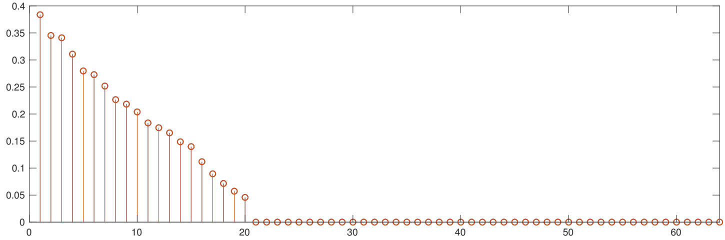


Sorted absolute coefficients of a sparse signal.

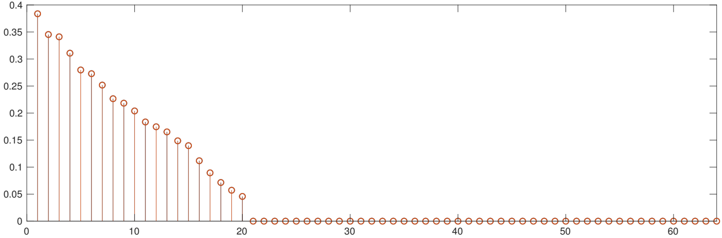
pessimistic but nothing we can do...



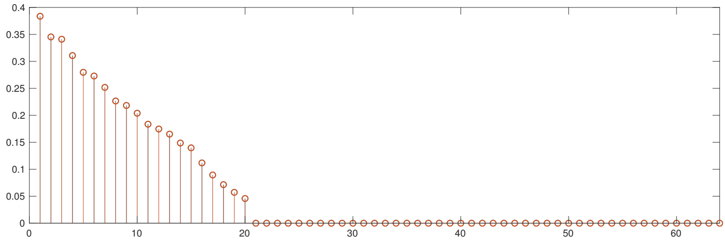
Sorted absolute coefficients of a sparse signal.



unless we look at decaying coefficients...

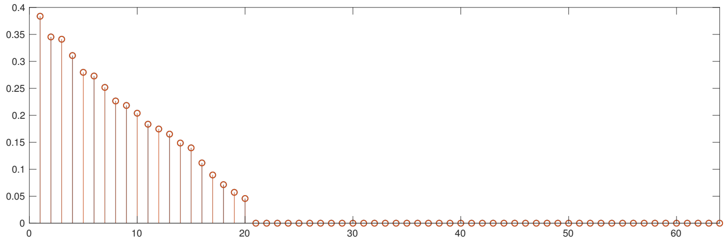


unless we look at decaying coefficients...



and add randomness, $x_i = c_i \sigma_i$.

Idea: $|\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx c_i \pm | \sum_{k \in J^c} c_k \sigma_k \langle \phi_i, \phi_k \rangle|$

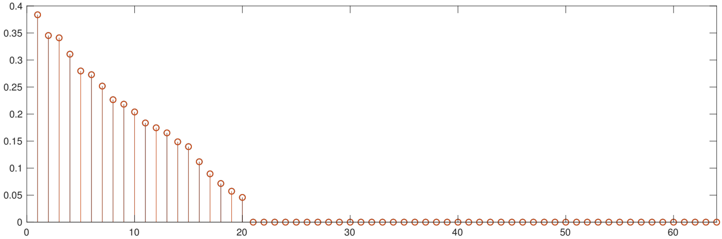


and add randomness, $x_i = c_i \sigma_i$.

Idea: $|\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx c_i \pm |\sum_{k \in J^c} c_k \sigma_k \langle \phi_i, \phi_k \rangle|$

$$\rightsquigarrow |x_i| \pm \|x_{J^c}\|_2 \cdot \mu$$

unless we look at decaying coefficients...

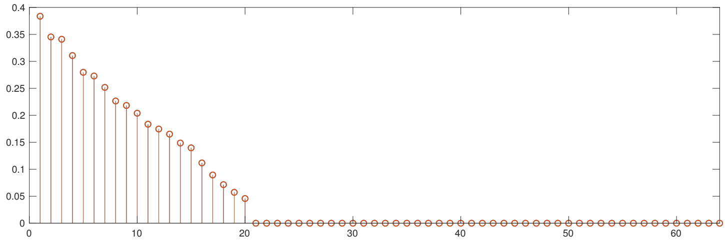


and add randomness, $x_i = c_i \sigma_i$.

Idea: $|\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx c_i \pm |\sum_{k \in J^c} c_k \sigma_k \langle \phi_i, \phi_k \rangle|$

$$\rightsquigarrow |x_i| \pm \|x_{J^c}\|_2 \cdot \mu$$

- Decay reduces the destructive energy of not recovered atoms



and add randomness, $x_i = c_i \sigma_i$.

$$\text{Idea:} \quad |\langle \phi_i, r_J \rangle| \approx |\langle \phi_i, \Phi_{J^c} x_{J^c} \rangle| \approx c_i \pm \Big| \sum_{k \in J^c} c_k \sigma_k \langle \phi_i, \phi_k \rangle \Big|$$

$$\rightsquigarrow |x_i| \pm \|x_{J^c}\|_2 \cdot \mu$$

- Decay reduces the destructive energy of not recovered atoms
- and the number of likely intermediate supports J.

I will not bore you with technicalities...

let's just say that you need

- to be a little creative to further reduce the number of intermediate subsets for which you need concentration
- and to remember that for $\lambda \in (0,1)$

$$(1-\lambda)^{1/\lambda} < e^{-1}.$$

Theorem (simplest case)

Assume that the support I satisfies $\delta_I := \|\Phi_I^T \Phi_I - I_S\|_{2,2} \leq \frac{1}{2}$ and additionally that the sorted coefficients c_i form a subgeometric sequence with parameter $\alpha < 1$ meaning $c_{i+1} \leq \alpha c_i$. Then OMP will recover the correct support except with probability $2SK^{1-m}$ as long as

$$S\mu^2 \lesssim 1-\alpha \qquad \text{and} \qquad S\mu^2\sqrt{m \log K} \lesssim \sqrt{1-\alpha}$$

I will not bore you with technicalities...

let's just say that you need

- to be a little creative to further reduce the number of intermediate subsets for which you need concentration
- and to remember that for $\lambda \in (0, 1)$

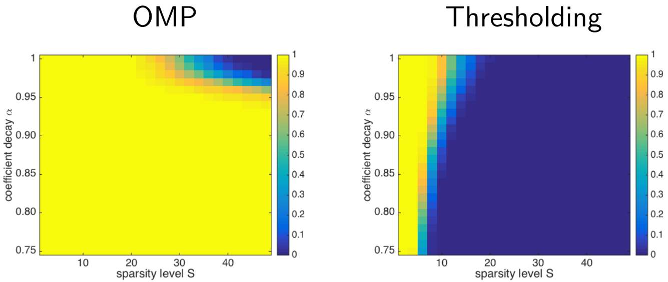
$$(1-\lambda)^{1/\lambda} < e^{-1}.$$

Theorem (simplest case)

Assume that the support I satisfies $\delta_I := \|\Phi_I^T \Phi_I - I_S\|_{2,2} \leq \frac{1}{2}$ and additionally that the sorted coefficients c_i form a subgeometric sequence with parameter $\alpha < 1$ meaning $c_{i+1} \leq \alpha c_i$. Then OMP will recover the correct support except with probability $2SK^{1-m}$ as long as

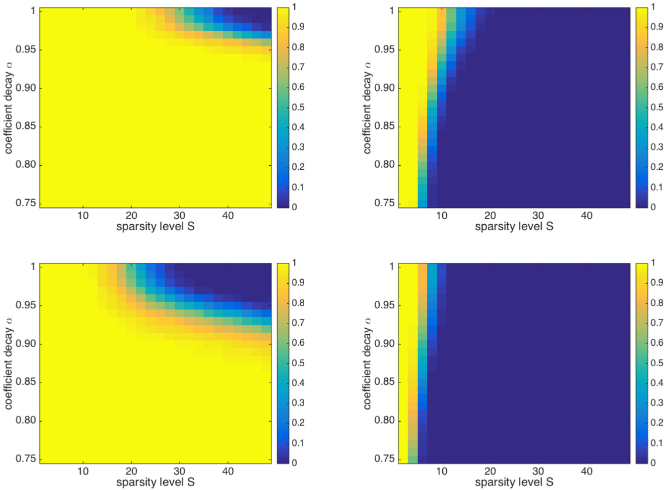
$$S\mu^2 \lesssim 1-\alpha \qquad \text{and} \qquad S\mu^2 \sqrt{m \log K} \lesssim \sqrt{1-\alpha}$$

instead some pretty pictures



instead some pretty pictures

OMP Thresholding



Percentage of correctly recovered supports for noiseless signals with various sparsity levels and coefficient decay parameters in the Dirac-DCT dictionary (top) and the Dirac-DCT-random dictionary (bottom).

average success with perturbations

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

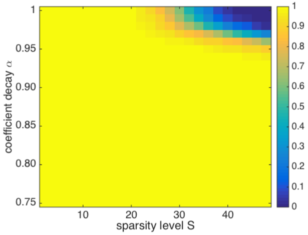
($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

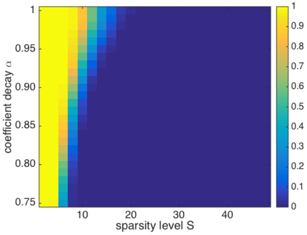
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding



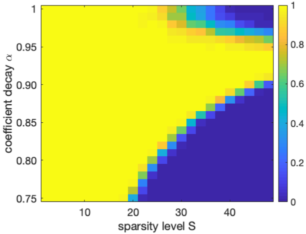
$$\omega_k = 0$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

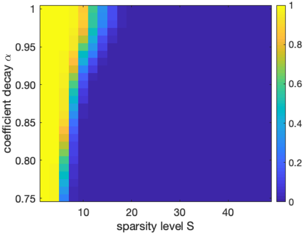
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding



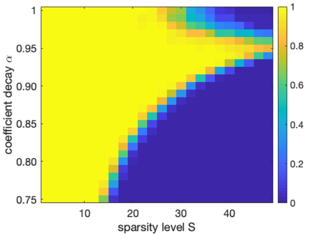
$$\gamma_k : \omega_k = 100 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

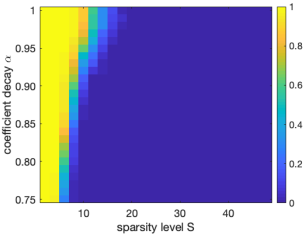
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding

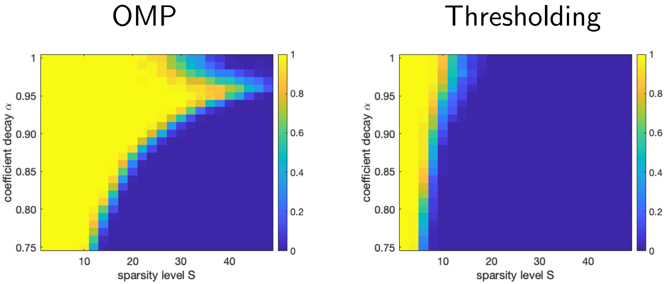


$$\gamma_k : \omega_k = 20 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)



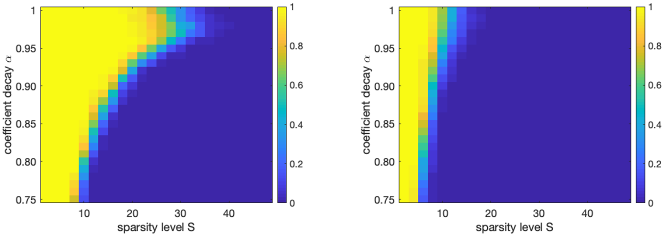
$$\gamma_k : \omega_k = 10 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP | Thresholding



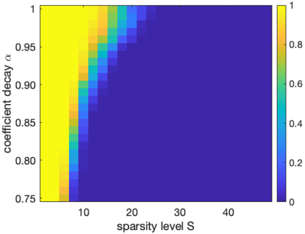
$$\gamma_k : \omega_k = 4 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

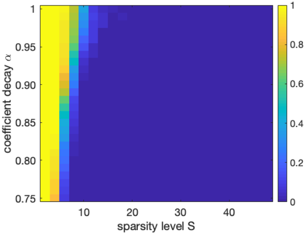
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding



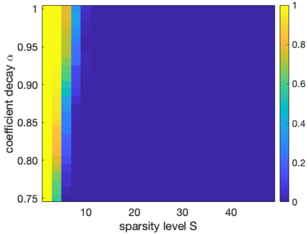
$$\gamma_k : \omega_k = 2 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

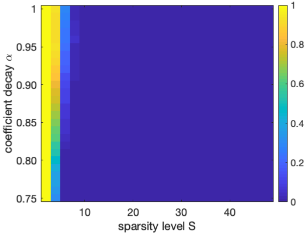
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding



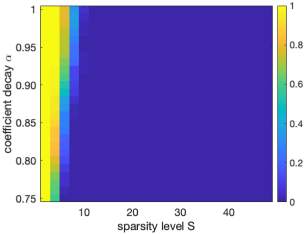
$$\gamma_k : \omega_k = 1 : 1$$

In dictionary learning we have: $y = \Phi_I x_I$
and need to recover I using Ψ rather than Φ.
Let's see what happens if

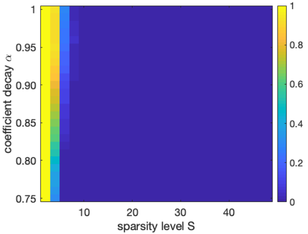
$$\psi_k = \gamma_k \phi_k + \omega_k z_k$$

($\gamma_k^2 + \omega_k^2 = 1$ and z_k chosen uniformly from the sphere $S^{d-1} \perp \phi_k$.)

OMP



Thresholding



$$\gamma_k : \omega_k = 1 : 1$$

So maybe thresholding is not cheap but sensible.

back to dictionary learning...

Algorithm (K-SVD)

Given an input dictionary Ψ *and* N *training signals* y_n *do:*

- *For all* n *use OMP to sparsely approximate* y_n

$$a_n = P(\Psi_{I_n})y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^{\dagger} y, \quad x_n|_{I_n^c} = 0.$$

- *For all* k *calculate*

$$R_k = \sum_{n:k \in I_n} [y_n - \Psi x_n + \psi_k x_n(k)][y_n - \Psi x_n + \psi_k x_n(k)]^T.$$

- *Update:* $\psi_k \leftarrow \arg\max_{\|v\|_2=1} \|R_k v\|_2$, *(via K SVDs).*

Algorithm (K-SVD)

Given an input dictionary Ψ and N training signals y_n do:

- *For all n use OMP to sparsely approximate y_n*

$$a_n = P(\Psi_{I_n})y_n = \Psi x_n \quad \Leftrightarrow \quad x_n|_{I_n} = \Psi_{I_n}^\dagger y, \quad x_n|_{I_n^c} = 0.$$

- *For all k calculate*

$$R_k = \sum_{n:k\in I_n} [y_n - \Psi x_n + \psi_k x_n(k)][y_n - \Psi x_n + \psi_k x_n(k)]^T.$$

- *Update: $\psi_k \leftarrow \arg\max_{\|v\|_2=1} \|R_k v\|_2$, (via K SVDs).*

and the smallprint in OMP:
Assume that the support I satisfies $\delta_I := \|\Phi_I^T \Phi_I - I_S\|_{2,2} \leq \frac{1}{2}$, ...

conditioning of random supports

Theorem (S. Chretien & S. Darses)

Let Φ *be a dictionary with coherence* μ *and operator norm* $B = \|\Phi\|_{2,2}$*. If* I *is chosen uniformly at random from all subsets* $J \subset \{1 \dots K\}$ *with* $|J| = S$ *then for* $\delta \in (0,1)$

$$\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2 K}{4e^2 S B^2}\right\}\right).$$

Theorem (S. Chretien & S. Darses)

Let Φ be a dictionary with coherence μ and operator norm $B = \|\Phi\|_{2,2}$. If I is chosen uniformly at random from all subsets $J \subset \{1 \ldots K\}$ with $|J| = S$ then for $\delta \in (0,1)$

$$\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2 K}{4e^2 S B^2}\right\}\right).$$

But actually we need $\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta | k \in I\right)$.

conditioning of random supports

Theorem (S. Chretien & S. Darses)

Let Φ be a dictionary with coherence μ and operator norm $B = \|\Phi\|_{2,2}$. If I is chosen uniformly at random from all subsets $J \subset \{1 \dots K\}$ with $|J| = S$ then for $\delta \in (0,1)$

$$\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2 K}{4e^2 S B^2}\right\}\right).$$

But actually we need $\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta | k \in I, j \in I\right)$.

conditioning of random supports

Theorem (S. Chretien & S. Darses)

Let Φ be a dictionary with coherence μ and operator norm $B = \|\Phi\|_{2,2}$. If I is chosen uniformly at random from all subsets $J \subset \{1 \dots K\}$ with $|J| = S$ then for $\delta \in (0,1)$

$$\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2 K}{4e^2 S B^2}\right\}\right).$$

But actually we need $\mathbb{P}\left(\|\Phi_I^T \Phi_I - \mathbb{I}_S\| \geq \delta | k \in I, j \notin I\right)$.

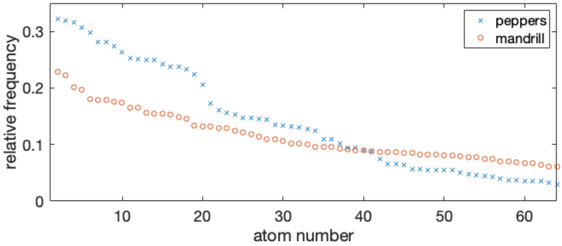
Theorem (S. Chretien & S. Darses)

Let Φ be a dictionary with coherence μ and operator norm $B = \|\Phi\|_{2,2}$. If I is chosen uniformly at random from all subsets $J \subset \{1 \dots K\}$ with $|J| = S$ then for $\delta \in (0,1)$

$$\mathbb{P}\left(\|\Phi_I^T\Phi_I - \mathbb{I}_S\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2 K}{4e^2 SB^2}\right\}\right).$$

But actually we need $\mathbb{P}\left(\|\Phi_I^T\Phi_I - \mathbb{I}_S\| \geq \delta | k \in I, j \notin I\right)$.

Also dictionaries for real data are not used uniformly:



conditioning of random supports (non uniform)

We can model this using weights $p_k \geq 0$ for $k \in K$ with $\sum p_k = S$, and choosing a support I according to

$$\mathbb{P}(I) = \begin{cases} \frac{1}{c} \prod_{i \in I} p_i \prod_{j \notin I} (1 - p_j) & \text{if} \quad |I| = S \\ 0 & \text{else} \end{cases}$$

Theorem (S. Ruetz & K.S.)

Let $\delta \in (0,1)$. Define the diagonal matrix W with $W_{kk} = \sqrt{p_k}$ and set $B = \max\{\|W\Phi^T\|_{2,2}, \|W\Psi^T\|_{2,2}\}$. Then we have for I being chosen according to the model above

$$\mathbb{P}\left(\|\Phi_I^T \Psi_I - D_I\| \geq \delta\right) \leq 216K \exp\left(-\min\left\{\frac{\delta}{2\mu}, \frac{\delta^2}{4e^2 B^2}\right\}\right).$$

I could go on forever...

...but I'm out of time

Questions

Comments

Thanks for your attention!!