Sparse and spurious: dictionary learning with noise and outliers

Rodolphe Jenatton
Amazon, Berlin

Joint work with **Rémi Gribonval** & **Francis Bach**

Optimization and Big Data, May 2015 @ Edinburgh

Movie quiz

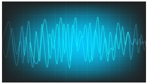
Which cinematographic reference?

Movie quiz

Which cinematographic reference?

Motivation: feature learning

Raw data

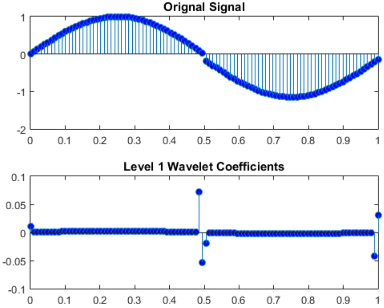


YouTube

amazon instant video

$\Longrightarrow$ **Good features**

Motivation: sparse representation



- Expensive to manipulate high-dimensional signals
  - Storage
  - Latency
  - ...
- Try to find sparse representation

Sparse coding in a nutshell

- **Idea**: Represent signals as combinations of few learned atoms

Sparse coding in a nutshell

- **Idea**: Represent signals as combinations of few learned atoms

- Applied to various settings and classes of signals
  - Neuroscience [OF97]
  - Image processing/Computer vision [EA06, Pey09, Mai10]
  - Audio processing [PABD06, GRKN07]
  - Topic modeling [JMOB11]
  - ...

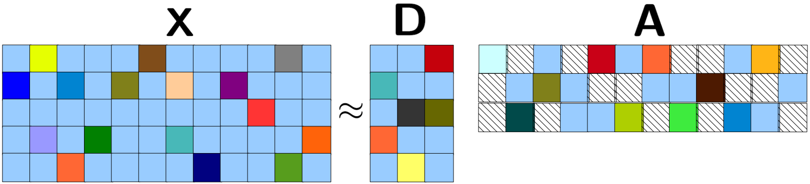
Sparse coding in a nutshell

- **Idea**: Represent signals as combinations of few learned atoms
- Applied to various settings and classes of signals
  - Neuroscience [OF97]
  - Image processing/Computer vision [EA06, Pey09, Mai10]
  - Audio processing [PABD06, GRKN07]
  - Topic modeling [JMOB11]
  - ...
- Several approaches:
  - Convex [BMP08, BB09]
  - Submodular [KC10]
  - Bayesian [ZCP^{+}09]
  - Non-convex matrix-factorization [OF97, LBRN07, MBPS10]

Sparse coding setting

- **Data**: n signals, $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{m \times n}$
- **Dictionary**: p atoms, $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m \times p}$
- **Decomposition**: $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] \in \mathbb{R}^{p \times n}$
- **Goal**:

$$\mathbf{X} \approx \mathbf{DA}, \text{ with sparse } \mathbf{A}$$



Sparse coding objective function

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times n},\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\Big[\overbrace{\frac{1}{2}\|\mathbf{x}^i-\mathbf{D}\boldsymbol{\alpha}^i\|_2^2}^{\text{data-fitting term}}+\lambda\underbrace{\|\boldsymbol{\alpha}^i\|_1}_{\text{sparsity-inducing norm}}\Big]$$

- $\mathcal{D}$: dictionaries with unit ℓ_2-norm atoms

Sparse coding objective function

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times n},\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\Big[\overbrace{\frac{1}{2}\|\mathbf{x}^i-\mathbf{D}\boldsymbol{\alpha}^i\|_2^2}^{\text{data-fitting term}}+\lambda\underbrace{\|\boldsymbol{\alpha}^i\|_1}_{\text{sparsity-inducing norm}}\Big]$$

- $\mathcal{D}$: dictionaries with unit ℓ_2-norm atoms
- Equivalently:

$$\min_{\mathbf{D}\in\mathcal{D}} F_{\mathbf{X}}(\mathbf{D}), \quad \text{with} \quad F_{\mathbf{X}}(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}^i}(\mathbf{D})$$

and

$$f_{\mathbf{x}}(\mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p}\Big[\frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2+\lambda\|\boldsymbol{\alpha}\|_1\Big]$$

Theoretical analysis of sparse coding

Theoretical analysis of sparse coding

- Excess risk analysis
    - [MP10, VMB10, MG12, GJB+13]
    - Non-asymptotic generalization bounds

Theoretical analysis of sparse coding

- Excess risk analysis
  - [MP10, VMB10, MG12, GJB+13]
  - Non-asymptotic generalization bounds
- **Identifiability** analysis
  - Study of local/global minima

Theoretical analysis of sparse coding

- Excess risk analysis
  - [MP10, VMB10, MG12, GJB+13]
  - Non-asymptotic generalization bounds
- **Identifiability** analysis
  - Study of local/global minima
  - Assume sparse model $\mathbf{x} = \mathbf{D}_0\alpha_0$ for a fixed $\mathbf{D}_0$
  - Possibly corrupted by some additive noise

Theoretical analysis of sparse coding

- Excess risk analysis
  - [MP10, VMB10, MG12, GJB+13]
  - Non-asymptotic generalization bounds
- **Identifiability** analysis
  - Study of local/global minima
  - Assume sparse model $\mathbf{x} = \mathbf{D}_0 \alpha_0$ for a fixed $\mathbf{D}_0$
  - Possibly corrupted by some additive noise
  - **Questions**:
    - Local minimum of $F_{\mathbf{X}} : \mathcal{D} \to \mathbb{R}$ around $\mathbf{D}_0$?

Theoretical analysis of sparse coding

- Excess risk analysis
  - [MP10, VMB10, MG12, GJB$^+$13]
  - Non-asymptotic generalization bounds
- **Identifiability** analysis
  - Study of local/global minima
  - Assume sparse model $\mathbf{x} = \mathbf{D}_0\alpha_0$ for a fixed $\mathbf{D}_0$
  - Possibly corrupted by some additive noise
  - **Questions**:
    - Local minimum of $F_{\mathbf{X}} : \mathcal{D} \to \mathbb{R}$ around $\mathbf{D}_0$?
    - Algorithmic schemes to reach those minima?

Previous work

Reference	Over.	Noise	Outliers	Global min/algo.	Poly. algo.	Sample comp. (no noise)
[GTC05] *Combinatorial*	YES	NO	NO	YES	NO	$m\binom{p}{m-1}$
[AEB06] *Combinatorial*	YES	NO	NO	YES	NO	$(k+1)\binom{p}{k}$
[GS10] ℓ^1	NO	NO	NO	NO	NO	$\frac{m^2 \log m}{k}$
[GWW11] ℓ^1	YES	NO	NO	NO	NO	kp^3
[SWW13] ℓ^0 *ER-SpUD*	NO	NO	NO	YES YES	NO YES	$m \log m$ $m^2 \log^2 m$
[Sch14b] *K-SVD criterion*	YES	YES	NO	NO	NO	$\frac{mp^3 k}{r^2}$
[AGM13] *Clustering*	YES	YES	NO	YES	YES	$\frac{p^2 \log p}{k^2}$
[AAN13] *Clustering & ℓ^1*	YES	NO	NO	YES	YES	$p \log mp$
[AAJN13] *Clustering & ℓ^1 with alt.*	YES	NO	NO	YES	YES	$p^2 \log p$
[Sch14a] *Resp. max. criterion*	YES	YES	NO	NO	NO	$\frac{mp^3 k}{r^2}$
Our *ℓ^1-regularized*	YES	YES	YES	NO	NO	mp^3

Our goal

- Consider probabilistic model with noise, $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$
  - Fixed reference dictionary $\mathbf{D}_0$
  - Random $(\boldsymbol{\alpha}_0, \boldsymbol{\varepsilon})$ with sparse $\boldsymbol{\alpha}_0$

Our goal

- Consider probabilistic model with noise, $\mathbf{x} = \mathbf{D}_0\boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon}$
  - Fixed reference dictionary $\mathbf{D}_0$
  - Random $(\boldsymbol{\alpha}_0, \boldsymbol{\varepsilon})$ with sparse $\boldsymbol{\alpha}_0$

- **Goal**: non-asymptotic characterization of

$$\mathbb{P}\big(F_{\mathbf{X}} \text{ has a local minimum in a ``neighborhood'' of } \mathbf{D}_0\big) \approx 1$$

with

$$F_{\mathbf{X}}(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathbf{x}^i}(\mathbf{D})$$

and

$$f_{\mathbf{x}}(\mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \left[\frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\right]$$

Some motivations

- Better account for empirical success

Some motivations

- Better account for empirical success
- Better understand roles of parameters, e.g.,
  - Number of atoms
    - Over-complete regimes
    - Model selection
  - Acceptable level of noise
  - Acceptable fraction of outliers
  - Sample complexity

Some motivations

- Better account for empirical success
- Better understand roles of parameters, e.g.,
  - Number of atoms
    - Over-complete regimes
    - Model selection
  - Acceptable level of noise
  - Acceptable fraction of outliers
  - Sample complexity
- Which parameters contribute to the curvature of $F_{\mathbf{X}}$?
  - E.g., design of new optimization strategies

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \varepsilon \quad \textbf{(Inliers)}$$

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \quad \textbf{(Inliers)}$$

- Fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$
  - Cumulative-coherence assumption [Fuc05, DH01]

$$\mu_k(\mathbf{D}_0) \triangleq \sup_{|\mathrm{J}| \leq k} \sup_{j \notin \mathrm{J}} \|[\mathbf{D}_0]_{\mathrm{J}}^{\top} \mathbf{d}_0^j\|_1$$

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \quad \textbf{(Inliers)}$$

- Fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$
  - Cumulative-coherence assumption [Fuc05, DH01]

$$\mu_k(\mathbf{D}_0) \triangleq \sup_{|\mathrm{J}| \leq k} \sup_{j \notin \mathrm{J}} \| [\mathbf{D}_0]_{\mathrm{J}}^{\top} \mathbf{d}_0^j \|_1$$

- Sparse random vector $\boldsymbol{\alpha}_0$
  - Uniformly draw $\mathrm{J} \subseteq \{1, \cdots, p\}, \ |\mathrm{J}| = k$

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \quad \textbf{(Inliers)}$$

- Fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$
  - Cumulative-coherence assumption [Fuc05, DH01]

$$\mu_k(\mathbf{D}_0) \triangleq \sup_{|\mathrm{J}| \leq k} \sup_{j \notin \mathrm{J}} \|[\mathbf{D}_0]_{\mathrm{J}}^{\top} \mathbf{d}_0^j\|_1$$

- Sparse random vector $\boldsymbol{\alpha}_0$
  - Uniformly draw $\mathrm{J} \subseteq \{1, \cdots, p\}$, $|\mathrm{J}| = k$
  - $[\boldsymbol{\alpha}_0]_{\mathrm{J}^c} = \mathbf{0}$

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \quad \textcolor{red}{\textbf{(Inliers)}}$$

- Fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$
  - Cumulative-coherence assumption [Fuc05, DH01]

$$\mu_k(\mathbf{D}_0) \triangleq \sup_{|\mathrm{J}| \leq k} \sup_{j \notin \mathrm{J}} \|[\mathbf{D}_0]_{\mathrm{J}}^\top \mathbf{d}_0^j\|_1$$

- Sparse random vector $\boldsymbol{\alpha}_0$
  - Uniformly draw $\mathrm{J} \subseteq \{1, \cdots, p\}, \ |\mathrm{J}| = k$
  - $[\boldsymbol{\alpha}_0]_{\mathrm{J}^c} = \mathbf{0}$
- Random noise $\boldsymbol{\varepsilon}$

Generative model of signals

$$\mathbf{x} = \mathbf{D}_0 \boldsymbol{\alpha}_0 + \boldsymbol{\varepsilon} \quad \textbf{(Inliers)}$$

- Fixed reference dictionary $\mathbf{D}_0 \in \mathcal{D}$
  - Cumulative-coherence assumption [Fuc05, DH01]

$$\mu_k(\mathbf{D}_0) \triangleq \sup_{|\mathrm{J}| \leq k} \sup_{j \notin \mathrm{J}} \|[\mathbf{D}_0]_{\mathrm{J}}^{\top} \mathbf{d}_0^j\|_1$$

- Sparse random vector $\boldsymbol{\alpha}_0$
  - Uniformly draw $\mathrm{J} \subseteq \{1, \cdots, p\}, \ |\mathrm{J}| = k$
  - $[\boldsymbol{\alpha}_0]_{\mathrm{J}^c} = \mathbf{0}$
- Random noise $\boldsymbol{\varepsilon}$

- **Outliers:** No assumption on $\mathbf{x}_{\text{outlier}}$

Signal assumptions

$$\begin{aligned}
\mathbb{E}\left\{[\boldsymbol{\alpha}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}^\top \mid \mathrm{J}\right\} &= \quad \mathbb{E}\{\alpha^2\}\cdot \mathbf{I} && \textbf{(coefficient whiteness)}\\
\mathbb{E}\left\{\boldsymbol{\varepsilon}[\boldsymbol{\alpha}_0]_{\mathrm{J}}^\top \mid \mathrm{J}\right\} &= \quad \mathbf{0} && \textbf{(decorrelation)}\\
\mathbb{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top | \mathrm{J}\right\} &= \quad \mathbb{E}\{\epsilon^2\}\cdot \mathbf{I} && \textbf{(noise whiteness)}
\end{aligned}$$

(similar assumptions for $\mathrm{sign}(\boldsymbol{\alpha}_0)$)

Signal assumptions

$$
\begin{aligned}
\mathbb{E}\left\{[\boldsymbol{\alpha}_0]_{\mathrm{J}}[\boldsymbol{\alpha}_0]_{\mathrm{J}}^{\top} \mid \mathrm{J}\right\} &= \mathbb{E}\{\alpha^2\}\cdot\mathbf{I} && \textbf{(coefficient whiteness)}\\
\mathbb{E}\left\{\boldsymbol{\varepsilon}[\boldsymbol{\alpha}_0]_{\mathrm{J}}^{\top} \mid \mathrm{J}\right\} &= \mathbf{0} && \textbf{(decorrelation)}\\
\mathbb{E}\left\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top} | \mathrm{J}\right\} &= \mathbb{E}\{\epsilon^2\}\cdot\mathbf{I} && \textbf{(noise whiteness)}
\end{aligned}
$$

(similar assumptions for $\mathrm{sign}(\boldsymbol{\alpha}_0)$)

"Flatness" of the distribution:

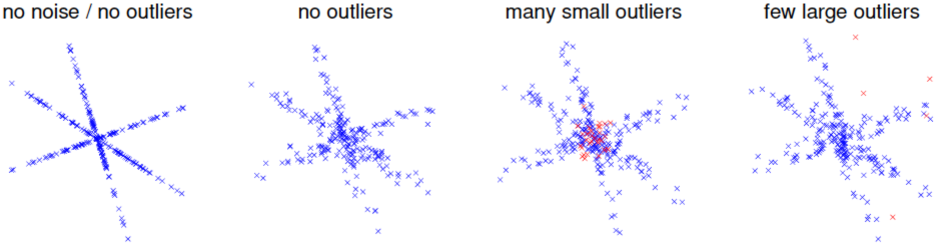
$$\kappa_\alpha \triangleq \frac{\mathbb{E}[|\alpha|]}{\sqrt{\mathbb{E}[\alpha^2]}}$$

Further boundedness assumptions

$$\begin{aligned}
\mathbb{P}(\min_{j\in \mathrm{J}} |[\boldsymbol{\alpha}_0]_j| < \underline{\alpha} \mid \mathrm{J}) = 0, &\quad \text{for some } \underline{\alpha} > 0 \quad &&(\textbf{coefficient threshold}) \\
\mathbb{P}(\|\boldsymbol{\alpha}_0\|_2 > M_{\boldsymbol{\alpha}}) = 0, &\quad \text{for some } M_{\boldsymbol{\alpha}} &&(\textbf{coefficient boundedness}) \\
\mathbb{P}(\|\boldsymbol{\varepsilon}\|_2 > M_{\boldsymbol{\varepsilon}}) = 0, &\quad \text{for some } M_{\boldsymbol{\varepsilon}} &&(\textbf{noise boundedness})
\end{aligned}$$

Useful for almost sure **exact recovery** to simplify expression of $F_{\mathbf{X}}$

Illustration



Asymptotic result ($n \gg 1$)

- **Coherence**: $\mu_k(\mathbf{D}_0) \leq 1/4$
- **Sparsity**: $k \lesssim p/|||\mathbf{D}_0|||_2$
- **Regularisation**: $0 < \lambda \lesssim \underline{\alpha}$ and $\tilde{\lambda} \triangleq \frac{\lambda}{\underline{\alpha}}$
- Consider

$$\begin{cases} C_{\min} \asymp \kappa_\alpha^2 \cdot |||\mathbf{D}_0|||_2 \cdot \frac{k}{p} \cdot \|[\mathbf{D}_0]^\top \mathbf{D}_0 - \mathbf{I}\|_{\mathrm{F}} \\ C_{\max} \asymp \frac{\mathbb{E}\ |\alpha|}{M_{\boldsymbol{\alpha}}} \cdot (1 - 2\mu_k(\mathbf{D}_0)) \end{cases}$$

Asymptotic result ($n \gg 1$)

- **Coherence**: $\mu_k(\mathbf{D}_0) \leq 1/4$
- **Sparsity**: $k \lesssim p / |||\mathbf{D}_0|||_2$
- **Regularisation**: $0 < \lambda \lesssim \underline{\alpha}$ and $\tilde{\lambda} \triangleq \frac{\lambda}{\underline{\alpha}}$
- Consider

$$\begin{cases} C_{\min} \asymp \kappa_\alpha^2 \cdot |||\mathbf{D}_0|||_2 \cdot \frac{k}{p} \cdot \|[\mathbf{D}_0]^\top \mathbf{D}_0 - \mathbf{I}\|_{\mathrm{F}} \\ C_{\max} \asymp \frac{\mathbb{E}\ |\alpha|}{M_{\boldsymbol{\alpha}}} \cdot (1 - 2\mu_k(\mathbf{D}_0)) \end{cases}$$

Proposition: For any r such that

$$C_{\min} \cdot \tilde{\lambda} \lesssim r \lesssim C_{\max} \cdot \tilde{\lambda} - \frac{M_{\boldsymbol{\varepsilon}}}{M_{\boldsymbol{\alpha}}}$$

$\mathbf{D} \in \mathcal{D} \mapsto \mathbb{E}\big[F_{\mathbf{x}}(\mathbf{D})\big]$ has a local minimum $\hat{\mathbf{D}}$ with $\|\hat{\mathbf{D}} - \mathbf{D}_0\|_{\mathrm{F}} < r$

Some comments on the proposition

- **Non-empty resolution interval**:
  - Limited over-completeness $p \lesssim m^2$
  - Still meaningful in practice [MBPS10], $m = 8 \times 8$ and $p = 256$

Some comments on the proposition

- **Non-empty resolution interval**:
  - Limited over-completeness $p \lesssim m^2$
  - Still meaningful in practice [MBPS10], $m = 8 \times 8$ and $p = 256$
- **"Flatness"** of the distribution:
  - The peakier, the better (i.e., the smaller κ_α, the better)

$$\kappa_\alpha = \frac{\mathbb{E}[|\alpha|]}{\sqrt{\mathbb{E}[\alpha^2]}}$$

Some comments on the proposition

- **Non-empty resolution interval**:
  - Limited over-completeness $p \lesssim m^2$
  - Still meaningful in practice [MBPS10], $m = 8 \times 8$ and $p = 256$
- **"Flatness"** of the distribution:
  - The peakier, the better (i.e., the smaller κ_α, the better)

$$\kappa_\alpha = \frac{\mathbb{E}[|\alpha|]}{\sqrt{\mathbb{E}[\alpha^2]}}$$

- **Noiseless regime**: $M_\varepsilon = 0$
  - Regime $\tilde{\lambda} \asymp r$
  - No restriction on the minimum resolution

Some comments on the proposition

- **Non-empty resolution interval**:
  - Limited over-completeness $p \lesssim m^2$
  - Still meaningful in practice [MBPS10], $m = 8 \times 8$ and $p = 256$
- **“Flatness”** of the distribution:
  - The peakier, the better (i.e., the smaller κ_α, the better)

$$\kappa_\alpha = \frac{\mathbb{E}[|\alpha|]}{\sqrt{\mathbb{E}[\alpha^2]}}$$

- **Noiseless regime**: $M_\varepsilon = 0$
  - Regime $\tilde{\lambda} \asymp r$
  - No restriction on the minimum resolution
- **Orthogonal dictionary**:
  - $C_{\min} = 0$
  - No restriction on the minimum resolution (even with noise!)

Some instantiations of the result

- Incoherent pair of orthonormal bases:
  - $p = 2m$ and $|||\mathbf{D}_0|||_2 = \sqrt{2}$

Some instantiations of the result

- Incoherent pair of orthonormal bases:
  - $p = 2m$ and $\|\!|\mathbf{D}_0|\!\|_2 = \sqrt{2}$
  - With i.i.d. bounded signals:
    - Coherence: $\mu_1 \lesssim 1/k^{3/2}$
    - Sparsity: $k \lesssim p^{1/3}$

Some instantiations of the result

- Incoherent pair of orthonormal bases:
  - $p = 2m$ and $|\!|\!|\mathbf{D}_0|\!|\!|_2 = \sqrt{2}$
  - With i.i.d. bounded signals:
    - Coherence: $\mu_1 \lesssim 1/k^{3/2}$
    - Sparsity: $k \lesssim p^{1/3}$
  - Fixed amplitude-profile coefficients [Sch14b]:

$$\boldsymbol{\alpha}_j = \epsilon_j \cdot \mathbf{a}_{\sigma(j)}, \text{ for } \begin{cases} \epsilon \text{ i.i.d.}, \mathbb{P}(\epsilon_j = \pm 1) = 1/2 \\ \sigma \text{ random permutation of J} \end{cases}$$

    - Coherence: $\mu_1 \lesssim 1/k$
    - Sparsity: $k \lesssim p^{1/2}$

Non-asymptotic result

Consider confidence $\delta > 0$ and the same assumptions as earlier.

Proposition: With prob. greater than $1 - 2e^{-\delta}$ and provided that

$$n_{\text{inliers}} \gtrsim p^2 \cdot (mp + \delta) \cdot \left[\frac{r + \tilde{\lambda} + \frac{M_\varepsilon}{M_\alpha}}{r - C_{\min} \cdot \tilde{\lambda}} \right]^2$$

$\mathbf{D} \in \mathcal{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ has a local minimum $\hat{\mathbf{D}}$ with $\|\hat{\mathbf{D}} - \mathbf{D}_0\|_{\mathrm{F}} < r$.

Non-asymptotic result

Consider confidence $\delta > 0$ and the same assumptions as earlier.

Proposition: With prob. greater than $1 - 2e^{-\delta}$ and provided that

$$n_{\text{inliers}} \gtrsim p^2 \cdot (mp + \delta) \cdot \left[\frac{r + \tilde{\lambda} + \frac{M_\varepsilon}{M_\alpha}}{r - C_{\min} \cdot \tilde{\lambda}} \right]^2$$

$\mathbf{D} \in \mathcal{D} \mapsto F_{\mathbf{X}}(\mathbf{D})$ has a local minimum $\hat{\mathbf{D}}$ with $\|\hat{\mathbf{D}} - \mathbf{D}_0\|_{\mathrm{F}} < r$.
The conclusion remains valid if

$$\frac{\|\mathbf{X}_{\text{outliers}}\|_{\mathrm{F}}^2}{n_{\text{inliers}} \cdot \mathbb{E}[\|\boldsymbol{\alpha}\|_2^2]} \lesssim \frac{r^2}{p}$$

Summary

	Orthogonal dictionary		General dictionary	
	Noiseless	**Noise**	**Noiseless**	**Noise**
Sample compl.	independent of r	$1/r^2$	independent of r	$1/r^2$
Resolution r	arbitrary small		$r \asymp C_{\min}\tilde{\lambda}$	$r > C_{\min}\tilde{\lambda}$
Outliers	*	depend on r	*	depend on r

*: More refined argument, if there exists $a_0 > 0$ such that

$$\text{for all } \mathbf{x}, \quad a_0 \cdot \|\mathbf{x}\|_2^2 \leq \|\mathbf{D}_0^\top \mathbf{x}\|_2^2,$$

then robustness to outliers can be shown to scale like $\mathcal{O}(a_0^{3/2} \cdot r/\tilde{\lambda})$

Main ideas behind the proof

1. Local-minimum condition over $\mathcal{D}$
   - Sphere $\mathcal{S}(r)$ and ball $\mathcal{B}(r)$ with radius r
   - $\inf_{\mathbf{D}\in\mathcal{S}(r)\cap\mathcal{D}} F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0) > 0$
   - Compact $\mathcal{B}(r)\cap\mathcal{D}$ with $F_{\mathbf{X}}$ continuous [GJB+13]

Main ideas behind the proof

1. Local-minimum condition over $\mathcal{D}$
   - Sphere $\mathcal{S}(r)$ and ball $\mathcal{B}(r)$ with radius r
   - $\inf_{\mathbf{D}\in\mathcal{S}(r)\cap\mathcal{D}} F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0) > 0$
   - Compact $\mathcal{B}(r)\cap\mathcal{D}$ with $F_{\mathbf{X}}$ continuous [GJB+13]
2. Transfer properties from $\mathbf{D}_0$ to $\mathbf{D}$ in a neighbourhood of $\mathbf{D}_0$
   - Coherence properties

Main ideas behind the proof

1. Local-minimum condition over $\mathcal{D}$
   - Sphere $\mathcal{S}(r)$ and ball $\mathcal{B}(r)$ with radius r
   - $\inf_{\mathbf{D}\in\mathcal{S}(r)\cap\mathcal{D}} F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0) > 0$
   - Compact $\mathcal{B}(r)\cap\mathcal{D}$ with $F_{\mathbf{X}}$ continuous [GJB+13]
2. Transfer properties from $\mathbf{D}_0$ to $\mathbf{D}$ in a neighbourhood of $\mathbf{D}_0$
   - Coherence properties
3. Replace $f_{\mathbf{x}}$ by some tractable surrogate
   - Exploit exact recovery [Fuc05, ZY06, Wai09]
   - Always holds thanks to boundedness assumptions
   - $f_{\mathbf{x}}(\mathbf{D})$ coincides with

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) \triangleq \frac{1}{2}\left[\|\mathbf{x}\|_2^2 - (\mathbf{D}_{\mathrm{J}}^\top\mathbf{x} - \lambda\mathbf{s}_{\mathrm{J}})^\top(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}(\mathbf{D}_{\mathrm{J}}^\top\mathbf{x} - \lambda\mathbf{s}_{\mathrm{J}})\right] \text{ with } \begin{cases}\mathrm{J} = \mathrm{supp}(\boldsymbol{\alpha}_0)\\ \mathbf{s} = \mathrm{sign}(\boldsymbol{\alpha}_0)\end{cases}$$

Main ideas behind the proof

1. Local-minimum condition over $\mathcal{D}$
   - Sphere $\mathcal{S}(r)$ and ball $\mathcal{B}(r)$ with radius r
   - $\inf_{\mathbf{D}\in\mathcal{S}(r)\cap\mathcal{D}} F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0) > 0$
   - Compact $\mathcal{B}(r) \cap \mathcal{D}$ with $F_{\mathbf{X}}$ continuous [GJB+13]
2. Transfer properties from $\mathbf{D}_0$ to $\mathbf{D}$ in a neighbourhood of $\mathbf{D}_0$
   - Coherence properties
3. Replace $f_{\mathbf{x}}$ by some tractable surrogate
   - Exploit exact recovery [Fuc05, ZY06, Wai09]
   - Always holds thanks to boundedness assumptions
   - $f_{\mathbf{x}}(\mathbf{D})$ coincides with

$$\phi_{\mathbf{x}}(\mathbf{D}|\mathbf{s}) \triangleq \frac{1}{2}\left[\|\mathbf{x}\|_2^2 - (\mathbf{D}_{\mathrm{J}}^\top\mathbf{x} - \lambda\mathbf{s}_{\mathrm{J}})^\top(\mathbf{D}_{\mathrm{J}}^\top\mathbf{D}_{\mathrm{J}})^{-1}(\mathbf{D}_{\mathrm{J}}^\top\mathbf{x} - \lambda\mathbf{s}_{\mathrm{J}})\right] \text{ with } \begin{cases} \mathrm{J} = \mathrm{supp}(\boldsymbol{\alpha}_0) \\ \mathbf{s} = \mathrm{sign}(\boldsymbol{\alpha}_0) \end{cases}$$

4. Concentration arguments
   - Rademacher averages

Conclusions and take-home messages

- Non-asymptotic analysis of local minimum of sparse coding
  - Noisy signals
  - Can also include generic outliers
  - But no algorithm analysis
- Towards more general signal models
  - Compressible [Cev08]
  - Spike and slab [IR05]
- Other penalties, beyond ℓ_1 [BJMO11]
- Different assumptions on $\mathbf{D}_0$ (better than coherence)

Sparse and spurious: dictionary learning with noise and outliers
http://arxiv.org/abs/1407.5155 (submitted)

Thank you all for your attention

References I

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli.
Learning sparsely used overcomplete dictionaries via alternating minimization.
Technical report, preprint arXiv:1310.7991, 2013.

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli.
Exact recovery of sparsely used overcomplete dictionaries.
Technical report, preprint arXiv:1309.1952, 2013.

Michal Aharon, Michael Elad, and Alfred M Bruckstein.
On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them.
Linear algebra and its applications, 416(1):48–67, 2006.

References II

Sanjeev Arora, Rong Ge, and Ankur Moitra.
New algorithms for learning incoherent and overcomplete dictionaries.
Technical report, preprint arXiv:1308.6273, 2013.

D. M. Bradley and J. A. Bagnell.
Convex coding.
In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski.
Optimization with sparsity-inducing penalties.
Foundations and Trends in Machine Learning, 4(1):1–106, 2011.

References III

F. Bach, J. Mairal, and J. Ponce.
Convex sparse matrix factorizations.
Technical report, Preprint arXiv:0812.1869, 2008.

V. Cevher.
Learning with compressible priors.
In *Advances in Neural Information Processing Systems*, 2008.

D. L. Donoho and X. Huo.
Uncertainty principles and ideal atomic decomposition.
IEEE Transactions on Information Theory, 47(7):2845–2862, 2001.

M. Elad and M. Aharon.
Image denoising via sparse and redundant representations over learned dictionaries.
IEEE Transactions on Image Processing, 54(12):3736–3745, December 2006.

References IV

J. J. Fuchs.
Recovery of exact sparse representations in the presence of bounded noise.
IEEE Transactions on Information Theory, 51(10):3601–3608, 2005.

Remi Gribonval, Rodolphe Jenatton, Francis Bach, Martin Kleinsteuber, and Matthias Seibert.
Sample complexity of dictionary learning and other matrix factorizations.
Technical report, preprint arXiv:1312.3790, 2013.

R Grosse, R Raina, H Kwong, and AY Ng.
Shift-invariant sparse coding for audio classification.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

References V

R. Gribonval and K. Schnass.
Dictionary identification—sparse matrix-factorization via ℓ_1-minimization.
IEEE Transactions on Information Theory, 56(7):3523–3539, 2010.

Pando Georgiev, Fabian Theis, and Andrzej Cichocki.
Sparse component analysis and blind source separation of underdetermined mixtures.
IEEE transactions on neural networks, 16(4):992–996, 2005.

Q. Geng, H. Wang, and J. Wright.
On the Local Correctness of L1 Minimization for Dictionary Learning.
Technical report, Preprint arXiv:1101.5672, 2011.

References VI

H. Ishwaran and J. S. Rao.
Spike and slab variable selection: frequentist and Bayesian strategies.
Annals of Statistics, 33(2):730–773, 2005.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
Journal of Machine Learning Research, 12:2297–2334, 2011.

A. Krause and V. Cevher.
Submodular dictionary selection for sparse representation.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

H. Lee, A. Battle, R. Raina, and A. Y. Ng.
Efficient sparse coding algorithms.
In *Advances in Neural Information Processing Systems*, 2007.

References VII

J. Mairal.
Sparse coding for machine learning, image processing and computer vision.
PhD thesis, École normale supérieure de Cachan - ENS Cachan, 2010.
Available at
http://tel.archives-ouvertes.fr/tel-00595312/fr/.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro.
Online learning for matrix factorization and sparse coding.
Journal of Machine Learning Research, 11(1):19–60, 2010.

N. A. Mehta and A. G. Gray.
On the sample complexity of predictive sparse coding.
Technical report, preprint arXiv:1202.4050, 2012.

References VIII

A. Maurer and M. Pontil.
k-dimensional coding schemes in hilbert spaces.
IEEE Transactions on Information Theory, 56(11):5839–5846, 2010.

B. A. Olshausen and D. J. Field.
Sparse coding with an overcomplete basis set: A strategy employed by V1?
Vision Research, 37:3311–3325, 1997.

M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies.
Sparse representations of polyphonic music.
Signal Processing, 86(3):417–431, 2006.

References IX

G. Peyré.
Sparse modeling of textures.
Journal of Mathematical Imaging and Vision, 34(1):17–31, 2009.

Karin Schnass.
Local identification of overcomplete dictionaries.
Technical report, preprint arXiv:1401.6354, 2014.

Karin Schnass.
On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd.
Applied and Computational Harmonic Analysis, 37(3):464–491, 2014.

Daniel A Spielman, Huan Wang, and John Wright.
Exact recovery of sparsely-used dictionaries.
In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

D. Vainsencher, S. Mannor, and A. M. Bruckstein.
The sample complexity of dictionary learning.
Technical report, Preprint arXiv:1011.5395, 2010.

M. J. Wainwright.
Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1- constrained quadratic programming.
IEEE Transactions on Information Theory, 55:2183–2202, 2009.

References XI

M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin.
Non-parametric Bayesian dictionary learning for sparse image representations.
In *Advances in Neural Information Processing Systems*, 2009.

P. Zhao and B. Yu.
On model selection consistency of Lasso.
Journal of Machine Learning Research, 7:2541–2563, 2006.