On the Reconstruction Risk of Convolutional Sparse Dictionary Learning

Shashank Singh, Barnabás Póczos, Jian Ma

Carnegie Mellon University

4 Oct. 2017
Allerton Conference

Motivation

- Sparsity is key to high-dimensional problems.
- Many data have **unknown** sparse representations.
- **Sparse dictionary learning** models data using **sparse linear combinations**.
- Many data have different sparse structure.
  - Naturalistic data are often **convolutionally** sparse.
    - Consistent local patterns in different positions.
    - Images, speech, genomic data, etc.
  - Several benefits of incorporating this structure into the model:
    - Faster computation
    - Greater interpretability
    - Reduced error

Motivation

- Sparsity is key to high-dimensional problems.
- Many data have **unknown** sparse representations.
- **Sparse dictionary learning** models data using **sparse linear combinations**.
- Many data have different sparse structure.
  - Naturalistic data are often **convolutionally** sparse.
    - Consistent local patterns in different positions.
    - Images, speech, genomic data, etc.
  - Several benefits of incorporating this structure into the model:
    - Faster computation
    - Greater interpretability
    - Reduced error

Motivation

- Sparsity is key to high-dimensional problems.
- Many data have **unknown** sparse representations.
- **Sparse dictionary learning** models data using **sparse linear combinations**.
- Many data have different sparse structure.
  - Naturalistic data are often **convolutionally** sparse.
    - Consistent local patterns in different positions.
    - Images, speech, genomic data, etc.
  - Several benefits of incorporating this structure into the model:
    - Faster computation
    - Greater interpretability
    - Reduced error

Motivation

- Sparsity is key to high-dimensional problems.
- Many data have **unknown** sparse representations.
- **Sparse dictionary learning** models data using **sparse linear combinations**.
- Many data have different sparse structure.
  - Naturalistic data are often **convolutionally** sparse.
    - Consistent local patterns in different positions.
    - Images, speech, genomic data, etc.
  - Several benefits of incorporating this structure into the model :
    - Faster computation
    - Greater interpretability
    - Reduced error

Background : IID Sparse Dictionary Learning (SDL)

Decompose data matrix $X \in \mathbb{R}^{d \times N}$ into $X \approx DR$, where

(a) Dictionary $D \in \mathbb{R}^{d \times K}$

(b) Encoding $R \in \mathbb{R}^{K \times N}$ is sparse

Example with $N = 1$:

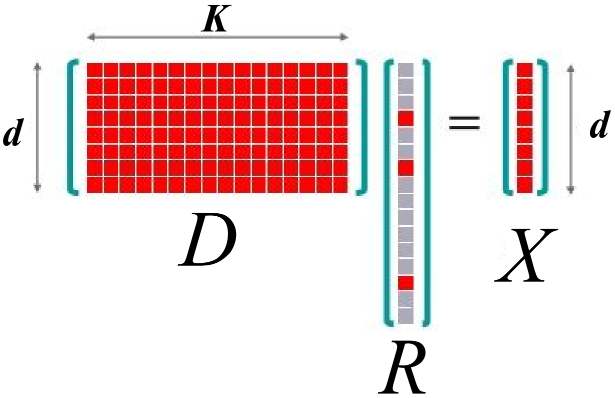


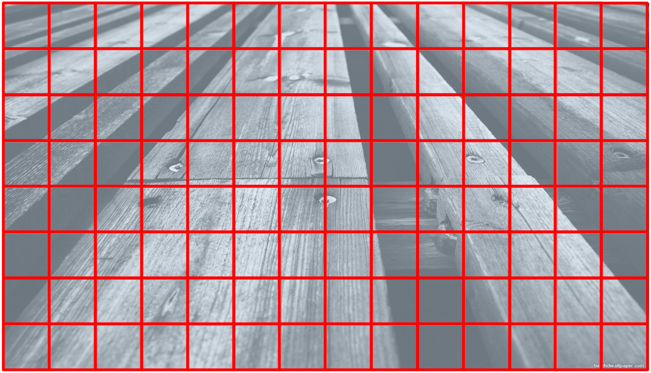
Image Credit : Manchor Ko

Example: IID SDL with Images

Example : IID SDL with Images

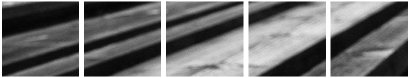
Images are **locally sparse, but not globally sparse**.

Example : IID SDL with Images



Images are **locally sparse, but not globally sparse**.

Example: IID SDL with Images



Dictionary Learned by IID SDL

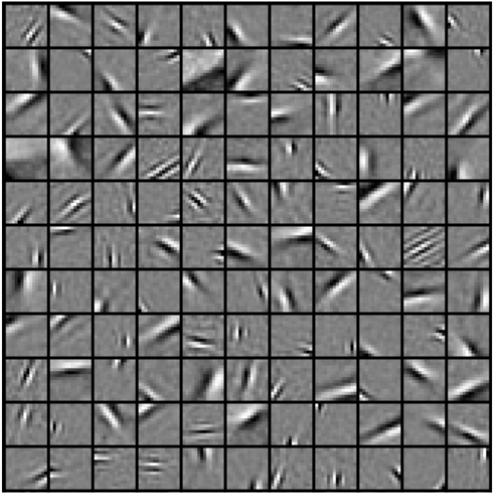


Image Credit : Olshausen & Field (1996)

Dictionary Learned by IID SDL

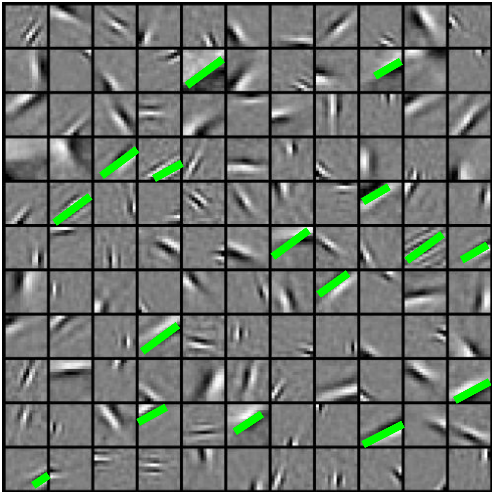


Image Credit : Olshausen & Field (1996)

- Highly redundant dictionary

Dictionary Learned by IID SDL

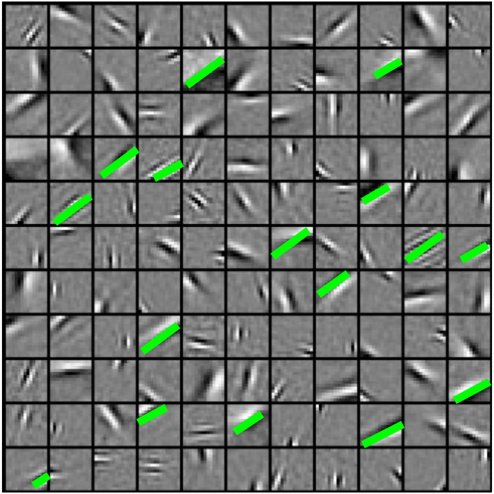


Image Credit : Olshausen & Field (1996)

- Highly redundant dictionary
- $\Rightarrow$ Computationally and statistically inefficient

Dictionary Learned by IID SDL

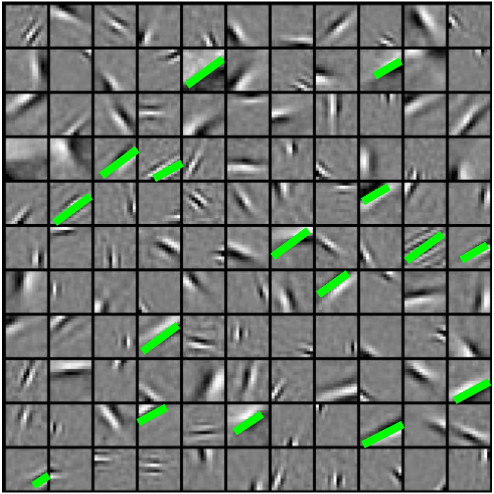


Image Credit : Olshausen & Field (1996)

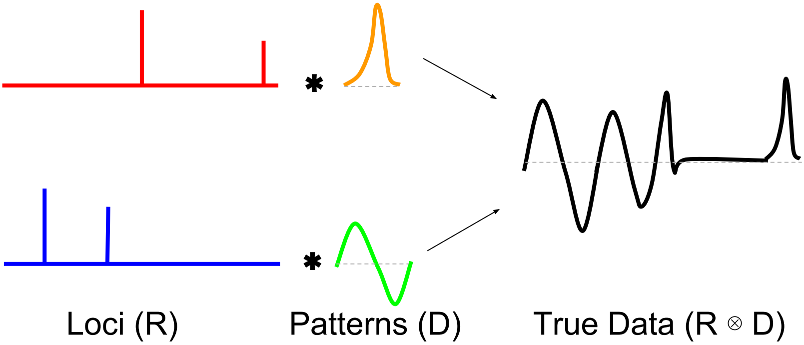
- Highly redundant dictionary
- $\Rightarrow$ Computationally and statistically inefficient
- Linear combinations ($X \approx RD$) lack translation invariance.

Not the right sparsity model!

Multi-convolution

For $R \in \mathbb{R}^{(N-n+1)\times K}$ and $D \in \mathbb{R}^{n \times K}$:

$$X = R \otimes D = \sum_{k=1}^{K} R_k * D_k \in \mathbb{R}^N.$$



CSDL Model and Goal

- Suppose we observe $Y = X + \epsilon \in \mathbb{R}^N$, where $\epsilon \in \mathbb{R}^N$ is noise and

$$X = R \otimes D \in \mathbb{R}^N,$$

  for some fixed **sparse** $R \in \mathbb{R}^{(N-n+1)\times K}$ and $D \in \mathbb{R}^{n\times K}$.

- Potential goals:
  - recover dictionary D
  - recover encoding R
  - recover true sequence $X = R \otimes D$

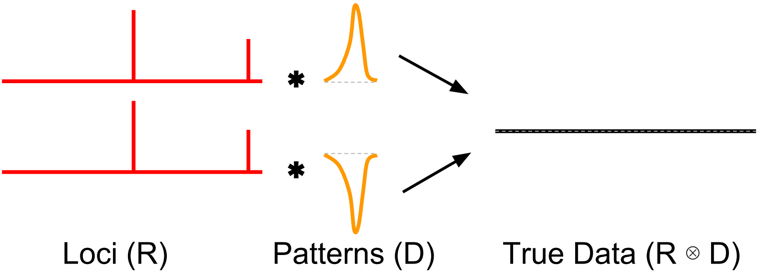
- We focus on recovering X (*reconstruction error*).

Why Reconstruction Error?

- Applications to denoising and compression.

Why Reconstruction Error ?

- Applications to denoising and compression.
- Recovering R and D requires potentially strong assumptions on D.



- Our bounds for recovering X require **almost no assumptions**.

Some Notation

1. To ensure sparsity, assume $\|R\|_{1,1} \leq \lambda$.
2. To fix scale, assume columns of D have at most unit $\mathcal{L}_2$ norm.

Problem Domain :

$$\mathcal{S}_\lambda = \left\{ (R, D) \in \mathbb{R}^{(N-n+1)\times K} \times \mathbb{R}^{n\times K} : \|R\|_{1,1} \leq \lambda, \|D\|_{2,\infty} \leq 1 \right\}.$$

Upper Bound

Optimization Formulation

$\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda$, where

$$\left(\widehat{R}_\lambda, \widehat{D}_\lambda\right) := \underset{(R,D)\in\mathcal{S}_\lambda}{\operatorname{argmin}} \|Y - R \otimes D\|_2 .$$

Upper Bound

Optimization Formulation

$\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda$, where

$$\left(\widehat{R}_\lambda, \widehat{D}_\lambda\right) := \underset{(R,D)\in\mathcal{S}_\lambda}{\operatorname{argmin}} \|Y - R \otimes D\|_2 .$$

Theorem (Upper Bound)

Suppose

1. $\lambda \geq \|R\|_{1,1}$.
2. *the coordinates of* $\epsilon \in \mathbb{R}^N$ *are sub-Gaussian with constant* σ.

Then,

$$\frac{1}{N} \mathop{\mathbb{E}}_{\epsilon}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}.$$

Lower Bound

Theorem (Minimax Lower Bound)

There exists an independent noise pattern ϵ, which is sub-Gaussian with parameter σ, such that

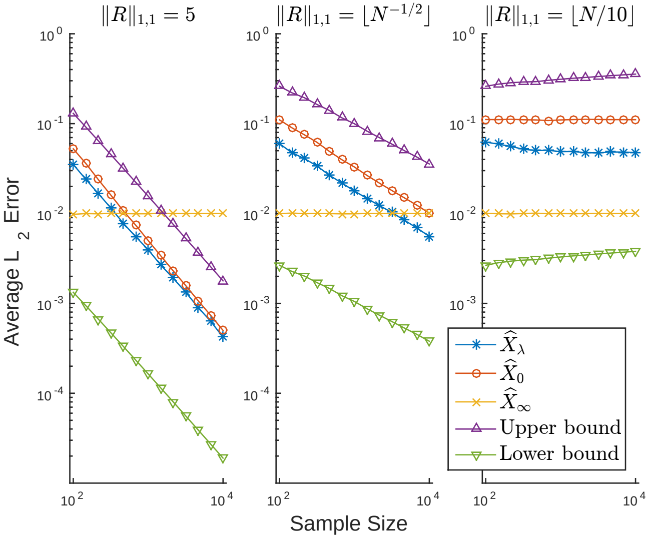
$$\inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N} \mathop{\mathbb{E}}_{\epsilon}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right] \geq \frac{\lambda}{8N} \min\left\{\lambda, \sigma\sqrt{\log(N-n+1)}\right\}.$$

- In the extremely sparse/noisy setting where

$$\lambda \leq \sigma\sqrt{\log(N-n+1)},$$

the trivial estimator $\widehat{X} = 0$ becomes optimal (with risk $\leq \lambda^2/N$).

Simulation : Convergence Rates and Sparsity



Comparing upper and lower bounds

For

$$M(\lambda, \sigma, N, n) := \inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N}\mathbb{E}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right],$$

we have (for $\lambda \geq \sigma\sqrt{\log(N-n+1)}$):

$$\text{Dependent Upper Bound:} \quad M(\lambda, \sigma, N, n) \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}$$

$$\text{Lower Bound:} \quad M(\lambda, \sigma, N, n) \geq \frac{\lambda\sigma\sqrt{\log(N-n+1)}}{8N}$$

Comparing upper and lower bounds

For

$$M(\lambda, \sigma, N, n) := \inf_{\widehat{X}} \sup_{(R,D) \in \mathcal{S}_\lambda} \frac{1}{N} \mathbb{E}\left[\left\| X - \widehat{X}_\lambda \right\|_2^2 \right],$$

we have (for $\lambda \geq \sigma\sqrt{\log(N-n+1)}$):

$$\text{Dependent Upper Bound:} \quad M(\lambda, \sigma, N, n) \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}$$

$$\text{Lower Bound:} \quad M(\lambda, \sigma, N, n) \geq \frac{\lambda\sigma\sqrt{\log(N-n+1)}}{8N}$$

Comparing upper and lower bounds

For

$$M(\lambda, \sigma, N, n) := \inf_{\widehat{X}} \sup_{(R,D)\in\mathcal{S}_\lambda} \frac{1}{N}\mathbb{E}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right],$$

we have (for $\lambda \geq \sigma\sqrt{\log(N-n+1)}$):

$$\text{Dependent Upper Bound:} \quad M(\lambda, \sigma, N, n) \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}$$

$$\text{Lower Bound:} \quad M(\lambda, \sigma, N, n) \geq \frac{\lambda\sigma\sqrt{\log(N-n+1)}}{8N}$$

$$\text{Independent Upper Bound:} \quad M(\lambda, \sigma, N, n) \leq \frac{4\lambda\sigma\sqrt{2\log(2N)}}{N}$$

Independence of Noise

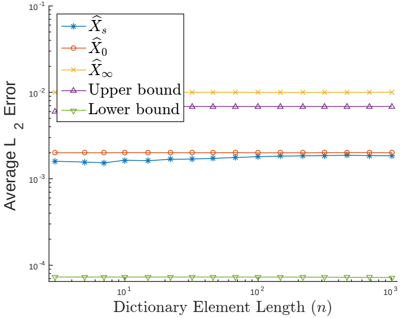


FIGURE – $\mathcal{N}(0, 0.1)$ Noise **independent** across signal

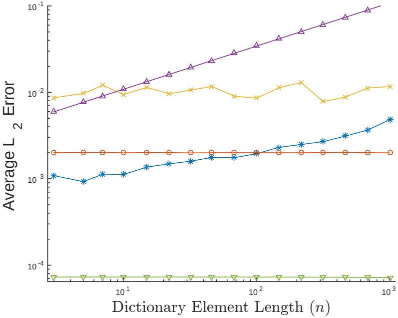


FIGURE – $\mathcal{N}(0, 0.1)$ Noise **identical** across signal

Summary

- Many data exhibit convolutional sparsity
  - For these data, CSDL $>$ SDL
- For fixed n, CSDL is guaranteed consistent (in reconstruction risk) if and only if

$$\frac{\lambda\sigma\sqrt{\log(N)}}{N} \to 0.$$

- Role of dictionary length n depends on dependence pattern of noise

Summary

- Many data exhibit convolutional sparsity
  - For these data, CSDL $>$ SDL
- For fixed n, CSDL is guaranteed consistent (in reconstruction risk) if and only if

$$\frac{\lambda\sigma\sqrt{\log(N)}}{N} \to 0.$$

- Role of dictionary length n depends on dependence pattern of noise

Thank you!

Appendix

Appendix

Upper Bound with Independence

Theorem (Upper Bound with Independence)

Suppose

1. $\lambda \geq \|R\|_{1,1}$.
2. *the coordinates of* $\epsilon \in \mathbb{R}^N$ *are sub-Gaussian with constant* σ.
3. $\widehat{X}_\lambda = \widehat{R}_\lambda \otimes \widehat{D}_\lambda$, *as previously.*
4. *Additionally, coordinates of* ϵ *are* ***independent***

Then,

$$\frac{1}{N}\mathbb{E}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2\log(2N)}}{N}.$$

Compare

$$\frac{1}{N}\mathbb{E}\left[\left\|X - \widehat{X}_\lambda\right\|_2^2\right] \leq \frac{4\lambda\sigma\sqrt{2n\log(2N)}}{N}$$

without independence assumption.