Sparse dictionary learning in the presence of noise and outliers

Rémi Gribonval - PANAMA Project-team
INRIA Rennes - Bretagne Atlantique, France
remi.gribonval@inria.fr

erc PLEASE
projection, learning and sparsity for efficient data processing

Overview

- Context: inverse problems & sparsity
- Data-driven dictionaries
- Learning as a nonconvex optimization problem
- Fast dictionaries
- Statistical guarantees
- Conclusion

Main Credits

- Theory for Dictionary Learning
  - **K. Schnass, R. Jenatton, F. Bach, M. Kleinsteuber, M. Seibert**

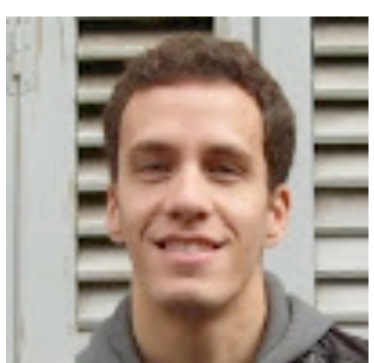




small-project.eu

SMALL

- Learning Fast Dictionaries
  - L. Le Magoarou



erc PLEASE

projection, learning and sparsity for efficient data processing

Additional Thanks To

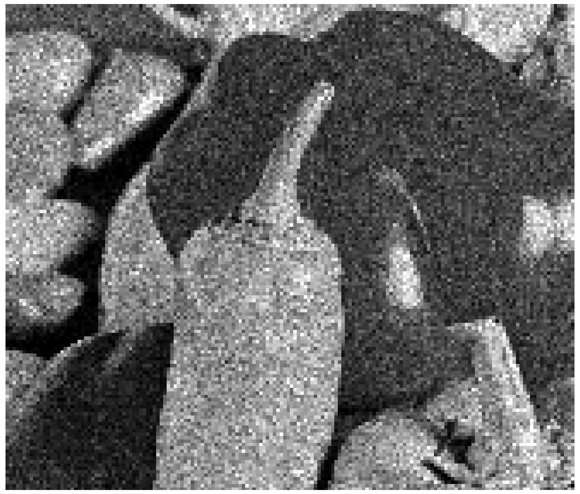
- Audio inpainting
  - A. Adler, N. Bertin, V. Emiya, M. Elad, C.Guichaoua, M. Jafari, M. Plumbley, S. Kitic
- Source localization
  - S. Nam, S. Kitic, N. Bertin, L. Albera
- Acoustic Imaging
  - G. Chardon, L. Daudet, A. Peillot, F. Ollivier, N. Bertin

small-project.eu

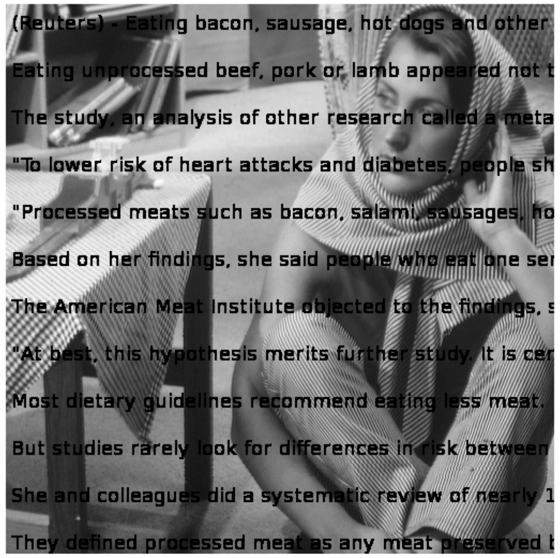
echange.inria.fr

Inverse problems

Inverse Problems in Image Processing



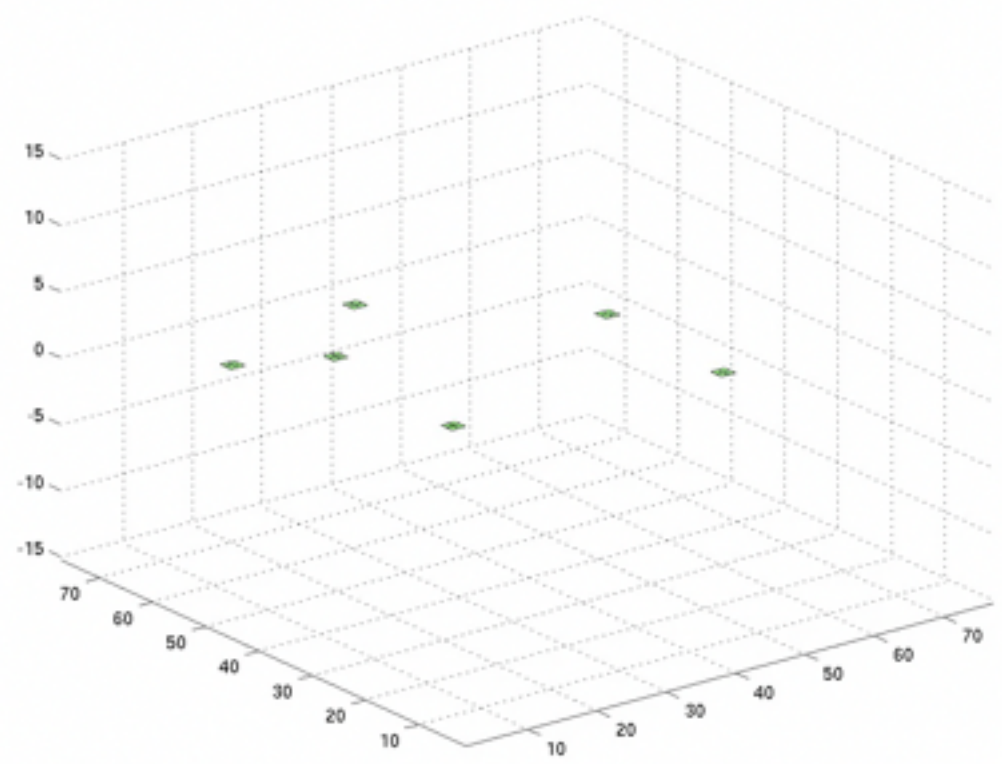
denoising



inpainting

+ Compression,
Source Localization, Separation,
Compressed Sensing ...

Inverse Problems in Acoustics



- **Possible goals**
  - ✓ **localize** emitting sources
  - ✓ **reconstruct** emitted signals
  - ✓ **extrapolate** acoustic field
- **Linear inverse problem**

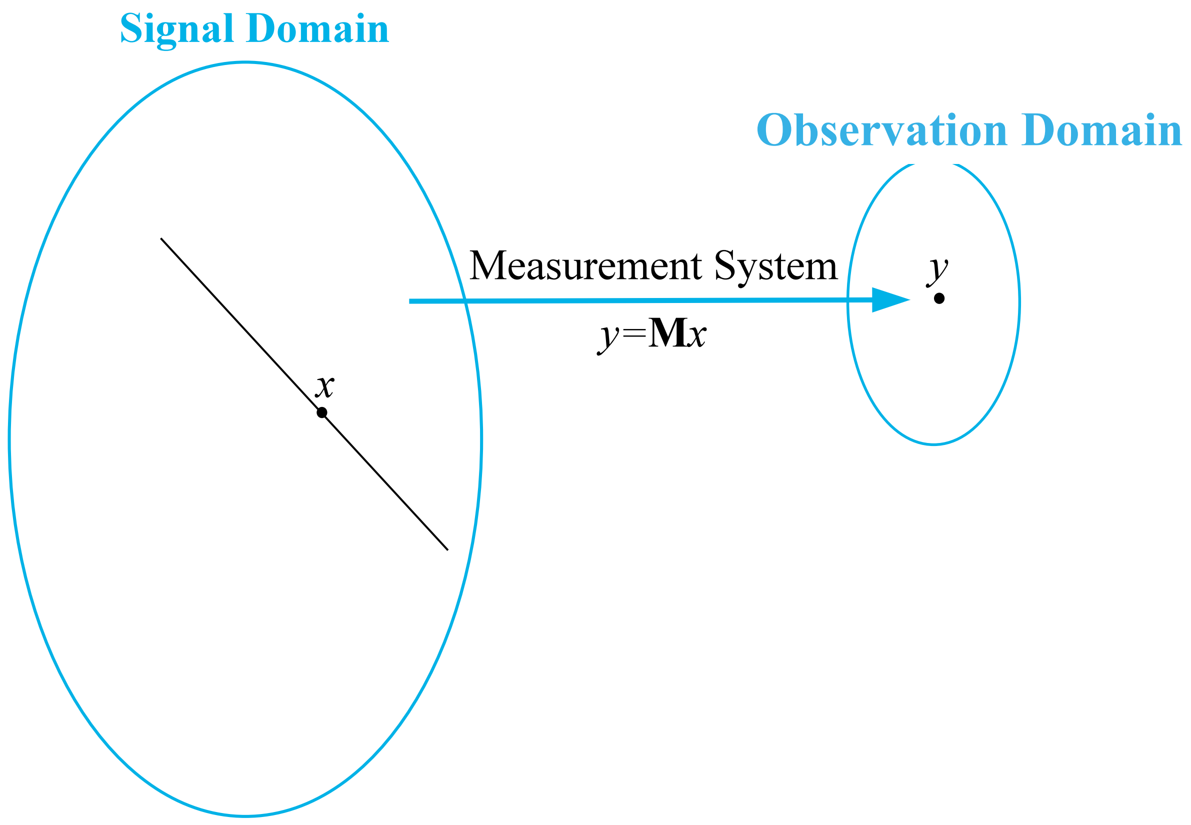
$$\mathbf{y} = \mathbf{M}\mathbf{x}$$

time-series recorded at sensors $\in \mathbb{R}^m$

(discretized) spatio-temporal acoustic field $\in \mathbb{R}^N$

- **Need a model**

Inverse Problems & Signal Models

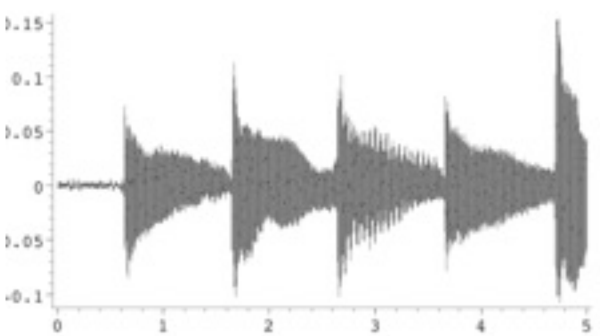


Need for a model = prior knowledge

Sparse signal models

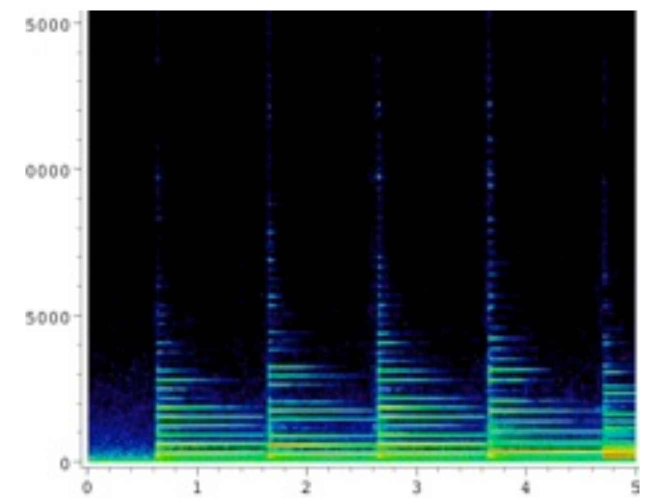
Typical Sparse Models

- Audio : time-frequency representations (MP3)



ANALYSIS

SYNTHESIS

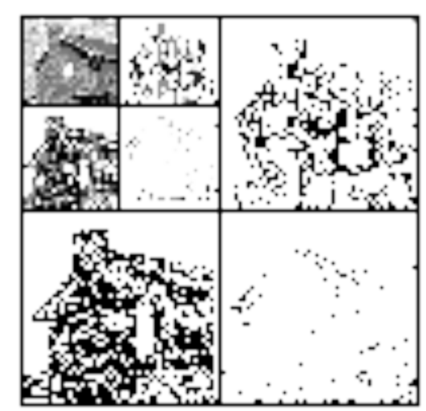


Black
= zero

- Images : wavelet transform (JPEG2000)

ANALYSIS

SYNTHESIS



White
= zero

Mathematical Expression

- Signal / image = high dimensional vector

$$\mathbf{x} \in \mathbb{R}^d$$

- **Model** = linear combination of basis vectors (ex: *time-frequency atoms, wavelets*)

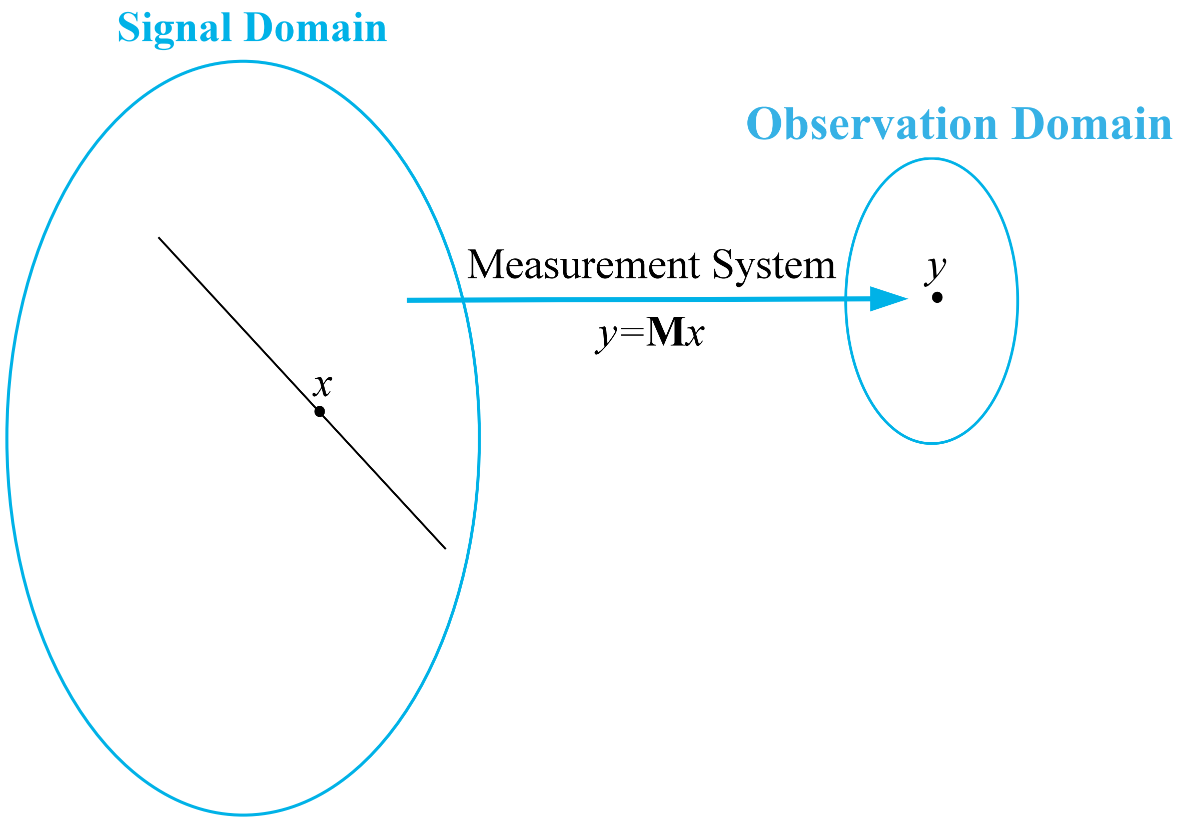
$$\mathbf{x} \approx \sum_k z_k \mathbf{d}_k = \mathbf{D}\mathbf{z}$$

Dictionary of atoms (Mallat & Zhang 93)

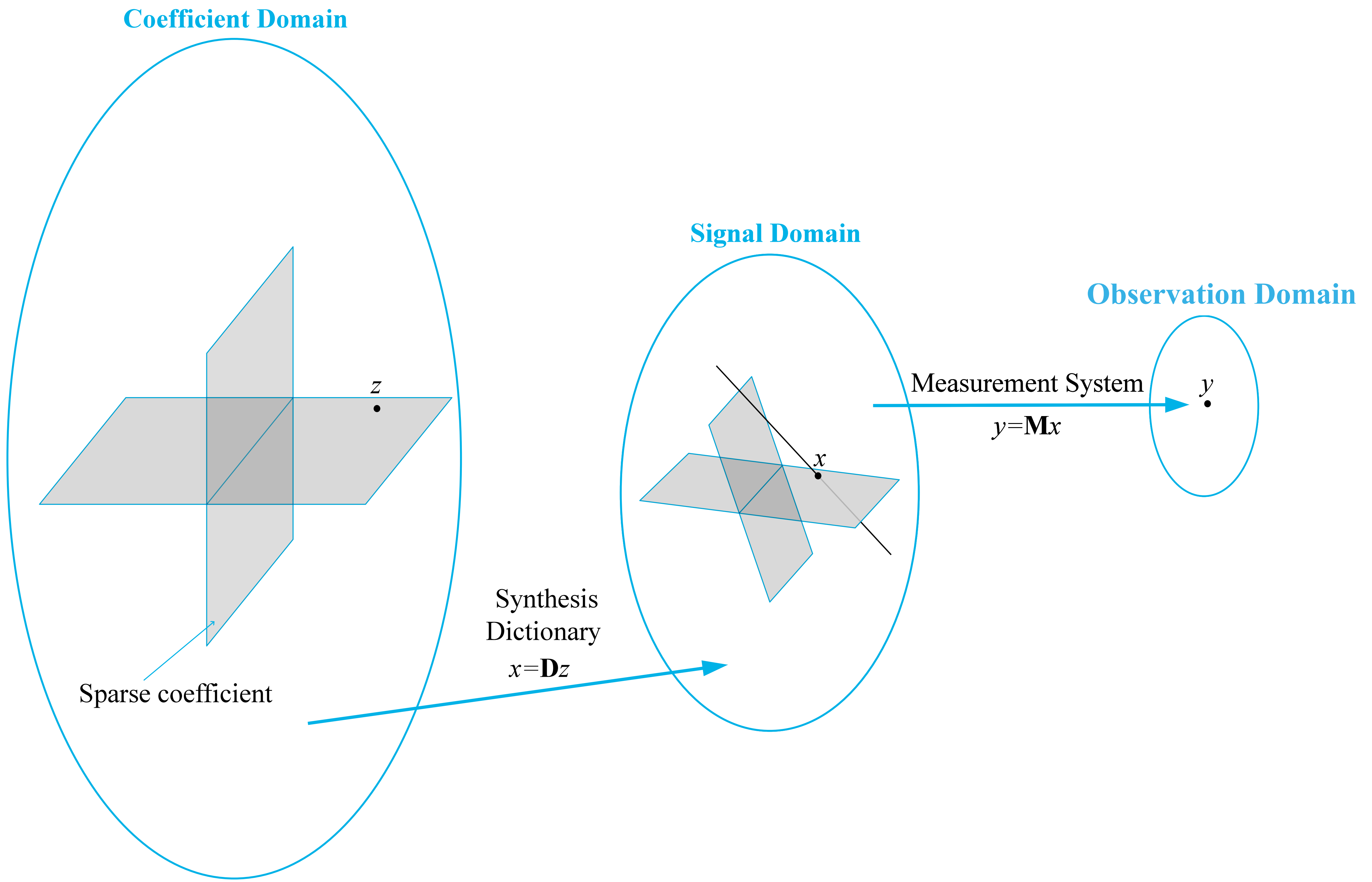
- **Sparsity** = small L0 (quasi)-norm

$$\|\mathbf{z}\|_0 = \sum_k |z_k|^0 = \mathrm{card}\{k, z_k \neq 0\}$$

Sparse Models and Inverse Problems



Sparse Models and Inverse Problems



Algorithmic Principles

- **Sparse regularization = penalized regression**

$$\hat{\mathbf{x}} = \mathbf{D}\hat{z} \qquad \text{with} \qquad \hat{z} = \arg\min_z \tfrac{1}{2}\|\mathbf{y} - \mathbf{MD}z\|_2^2 + \lambda\|z\|_p^p$$

Algorithmic Principles

- **Sparse regularization = penalized regression**

$\hat{\mathbf{x}} = \mathbf{D}\hat{z}$ **with** $\hat{z} = \arg\min_{z} \frac{1}{2}\|\mathbf{y} - \mathbf{M}\mathbf{D}z\|_2^2 + \lambda\|z\|_p^p$

- **In practice: iterative thresholding**
  - ✓ gradient descent to improve data fidelity

$$\hat{z}^{i+1/2} \leftarrow \hat{z}^{i} + \mu\mathbf{D}^{T}\mathbf{M}^{T}(\mathbf{y} - \mathbf{M}\mathbf{D}\hat{z}^{i})$$

  - ✓ thresholding to promote (structured) sparsity

$$\hat{z}^{i+1} \leftarrow \mathrm{Threshold}_p(\hat{z}^{i+1/2}, \lambda)$$

Algorithmic Principles

- **Sparse regularization = penalized regression**

$$\hat{\mathbf{x}} = \mathbf{D}\hat{z} \qquad \text{with} \qquad \hat{z} = \arg\min_{z} \tfrac{1}{2}\|\mathbf{y} - \mathbf{MD}z\|_2^2 + \lambda\|z\|_p^p$$

- **In practice: iterative thresholding**
  - ✓ gradient descent to improve data fidelity

$$\hat{z}^{i+1/2} \leftarrow \hat{z}^{i} + \mu\mathbf{D}^T\mathbf{M}^T(\mathbf{y} - \mathbf{MD}\hat{z}^{i})$$

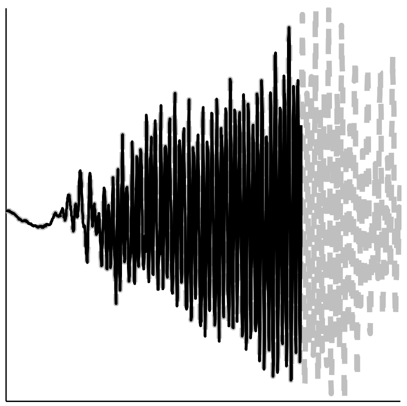
  - ✓ thresholding to promote (structured) sparsity

$$\hat{z}^{i+1} \leftarrow \mathrm{Threshold}_p(\hat{z}^{i+1/2}, \lambda)$$

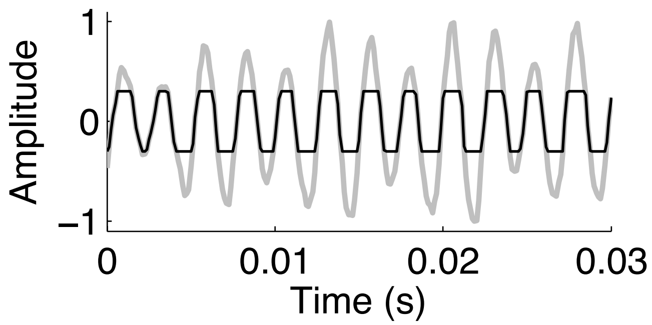
- **See also**: greedy algorithms (Matching Pursuit)

Example: «Audio Inpainting»

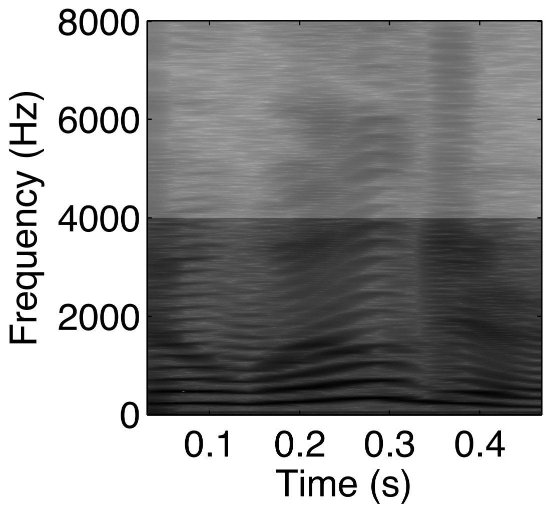
Holes (Packet Loss)



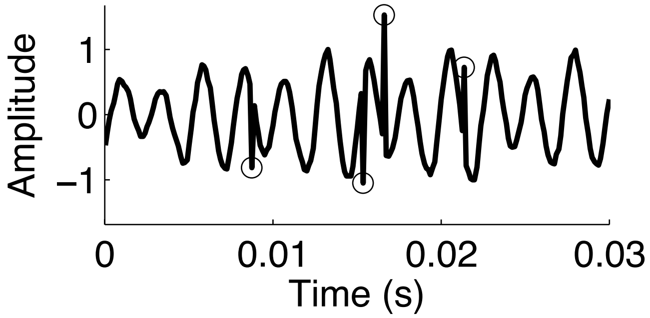
Clipping



Limited bandwidth



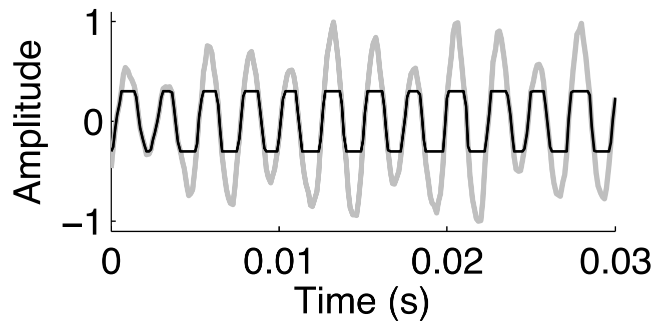
Clicks



A. Adler, V. Emiya, M. Jafari, M. Elad, R.Gribonval and M. Plumbley, <u>Audio Inpainting</u>, IEEE Trans ASLP, 2012

Example: «Audio Inpainting»

Clipping



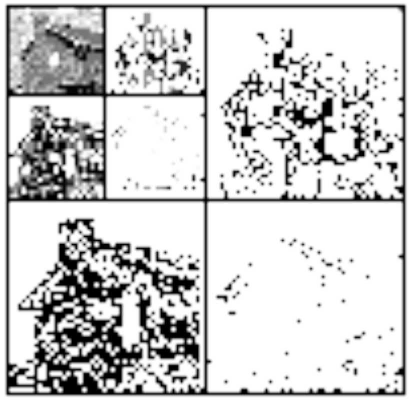
http://people.rennes.inria.fr/Srdan.Kitic/?page_id=40

A. Adler, V. Emiya, M. Jafari, M. Elad, R.Gribonval and M. Plumbley, Audio Inpainting, IEEE Trans ASLP, 2012

Dictionary learning for sparse modeling

Sparse Signal Model

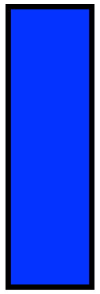
$$\mathbf{x} \approx \mathbf{D}\mathbf{z}$$

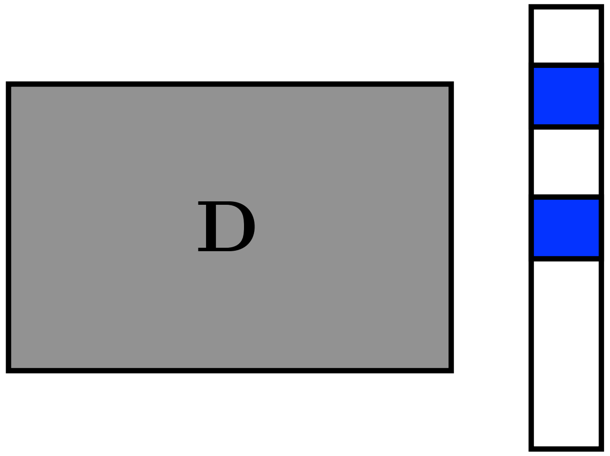


Signal
Image

(Overcomplete)
dictionary of atoms
(wavelets ...)

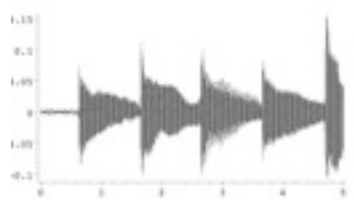
Sparse
Representation
Coefficients





From Analytic to Learned Dictionaries

Analytic dictionaries (Fourier, wavelets ...)

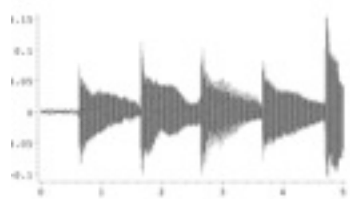


Signals

Images

From Analytic to Learned Dictionaries

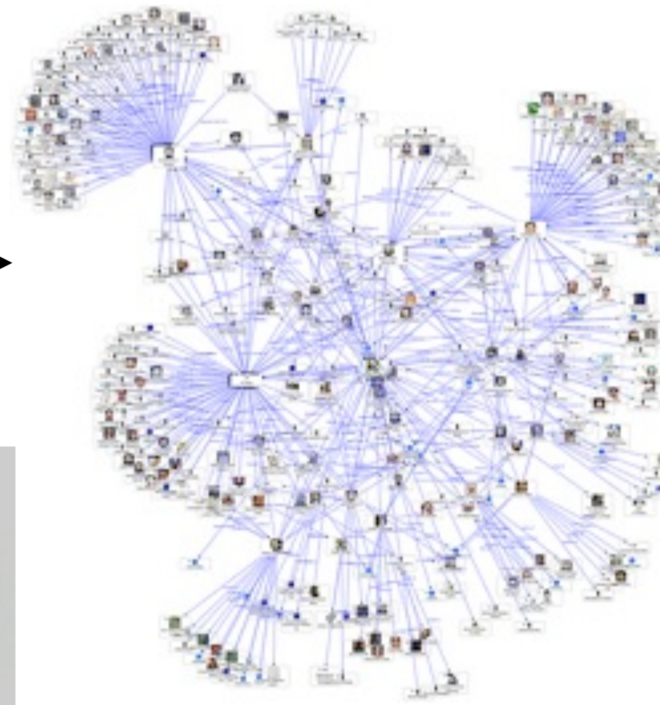
Analytic dictionaries (Fourier, wavelets ...)

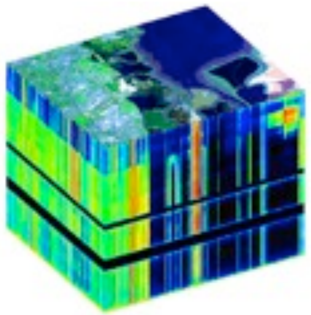


Signals

Images

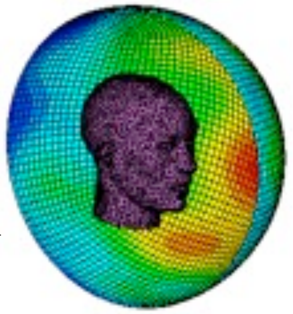
Graph data
Social networks
Brain connectivity

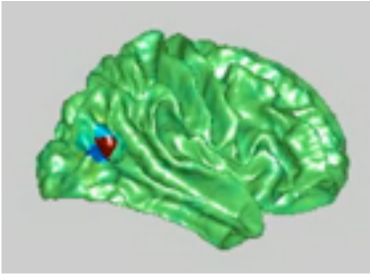




Hyperspectral
Satellite imaging

Spherical geometry
Cosmology, HRTF (3D audio)

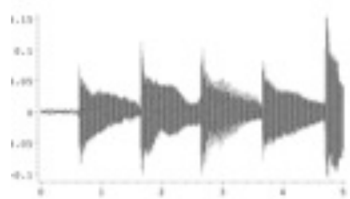




Vector valued
Diffusion tensor

From Analytic to Learned Dictionaries

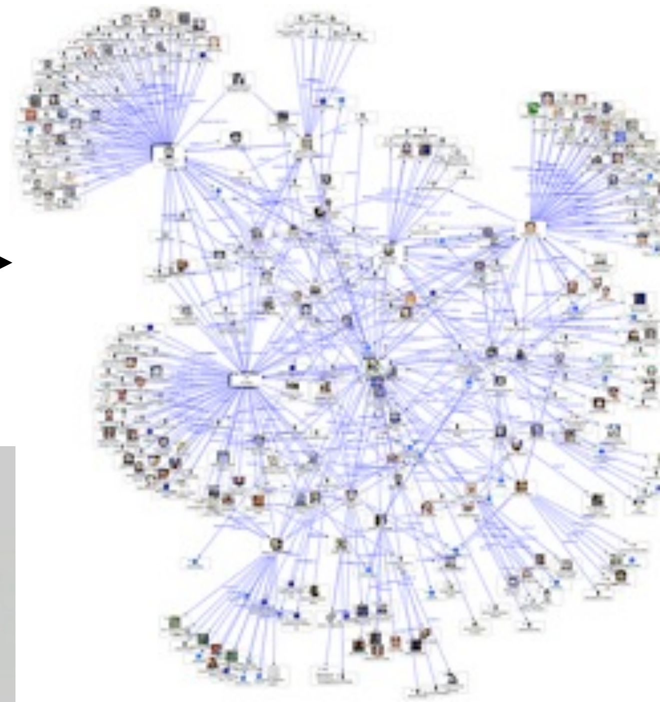
Analytic dictionaries (Fourier, wavelets ...)

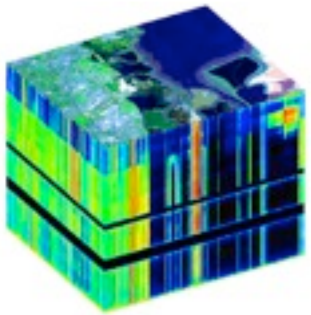


Signals

Images

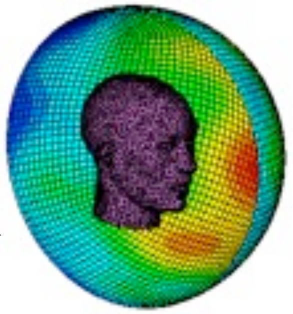
Graph data
Social networks
Brain connectivity

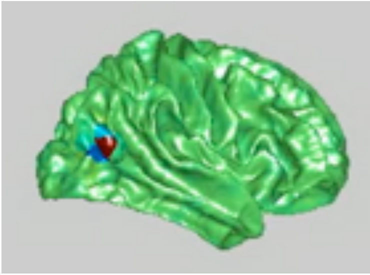




Hyperspectral
Satellite imaging

Spherical geometry
Cosmology, HRTF (3D audio)





Vector valued
Diffusion tensor

Data-driven (learned) dictionaries

A Quest for the Perfect Sparse Model

Training database

A Quest for the Perfect Sparse Model

Training database

Training patches

$\mathbf{x}_n$

$$1 \leq n \leq N$$

A Quest for the Perfect Sparse Model

Training database

patch extraction

Training patches

$$\boxed{\mathbf{x}_n} = \mathbf{D} z_n, \quad 1 \leq n \leq N$$

Unknown dictionary

Unknown sparse coefficients

A Quest for the Perfect Sparse Model

Training database

Training patches

$$\mathbf{x}_n = \mathbf{D} z_n, \quad 1 \leq n \leq N$$

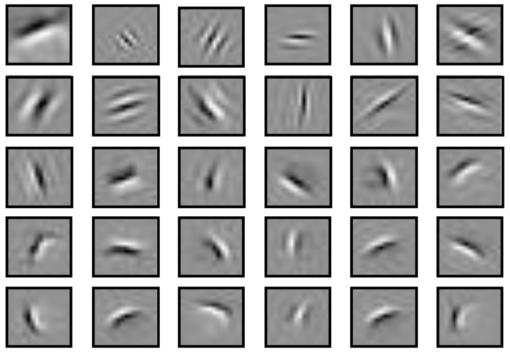
Unknown dictionary	Unknown sparse coefficients

$\hat{\mathbf{D}}$ = edge-like atoms

[Olshausen & Field 96, Aharon et al 06, Mairal et al 09, ...]

= shifts of edge-like motifs

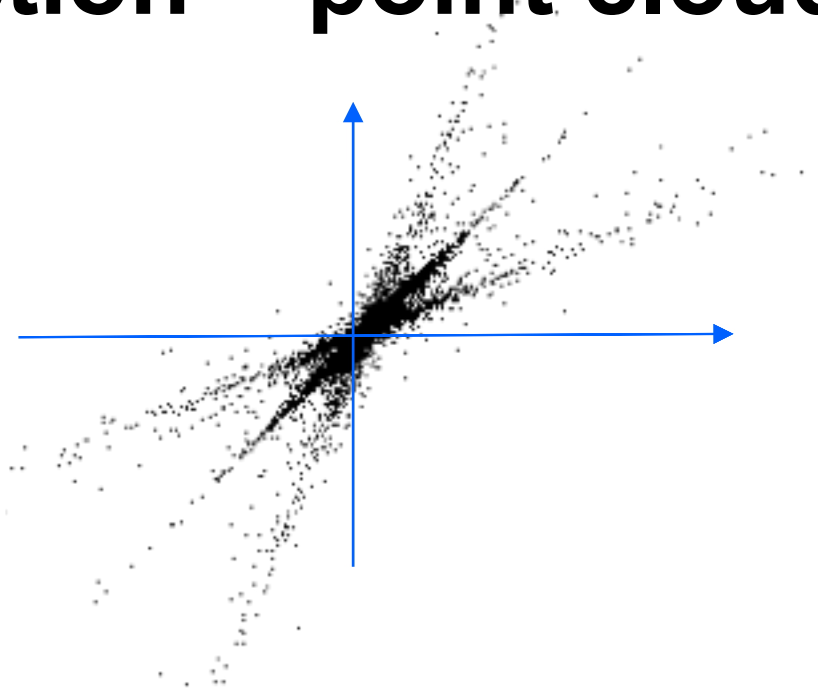
[Blumensath 05, Jost et al 05, ...]



Dictionary learning as sparse matrix factorization

Dictionary Learning
= Sparse Matrix Factorization

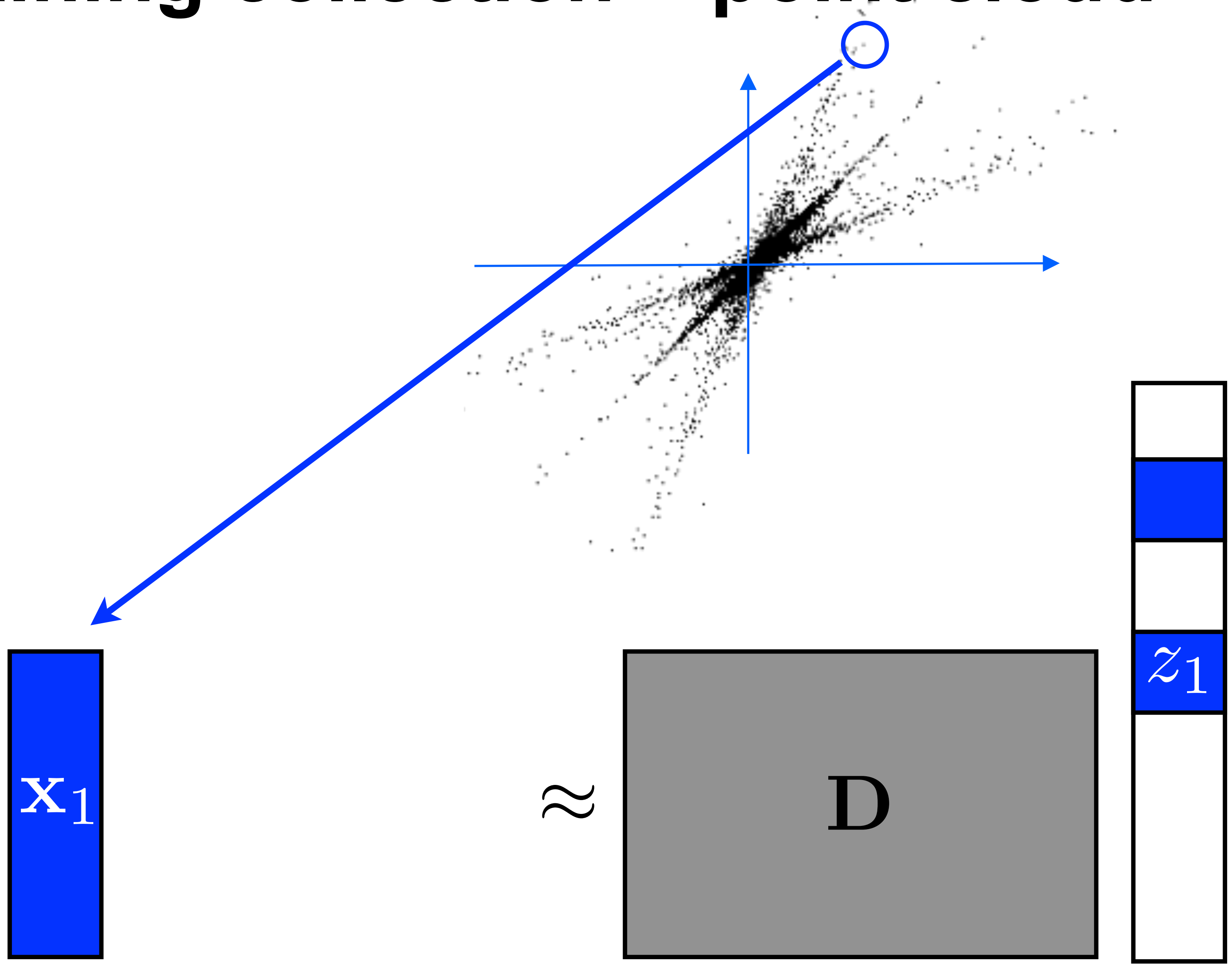
- **Training collection = point cloud**



Dictionary Learning
= Sparse Matrix Factorization

- **Training collection = point cloud**

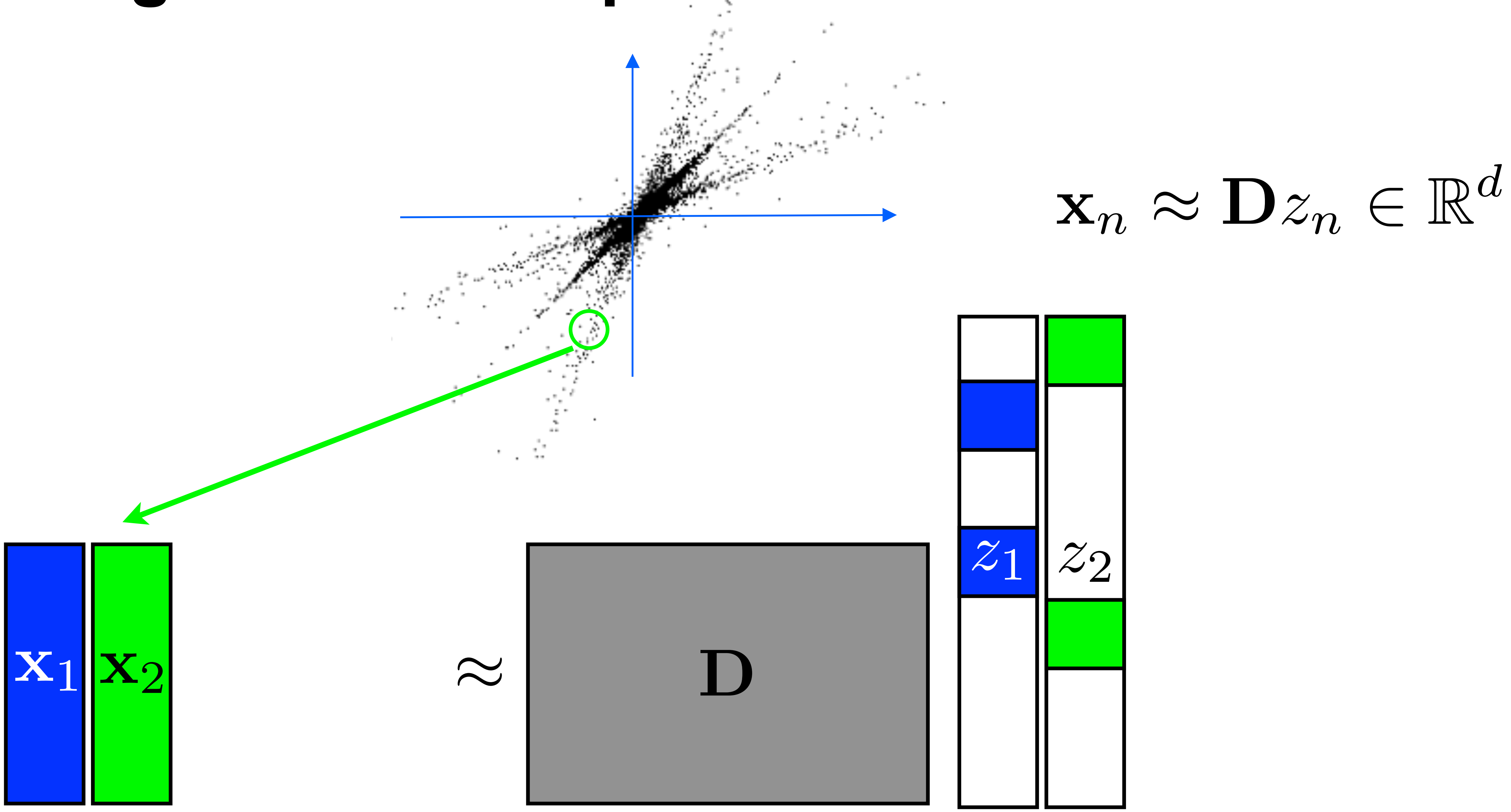
$$\mathbf{x}_n \approx \mathbf{D} z_n \in \mathbb{R}^d$$



***s*-sparse**
= at most *s*
nonzero entries

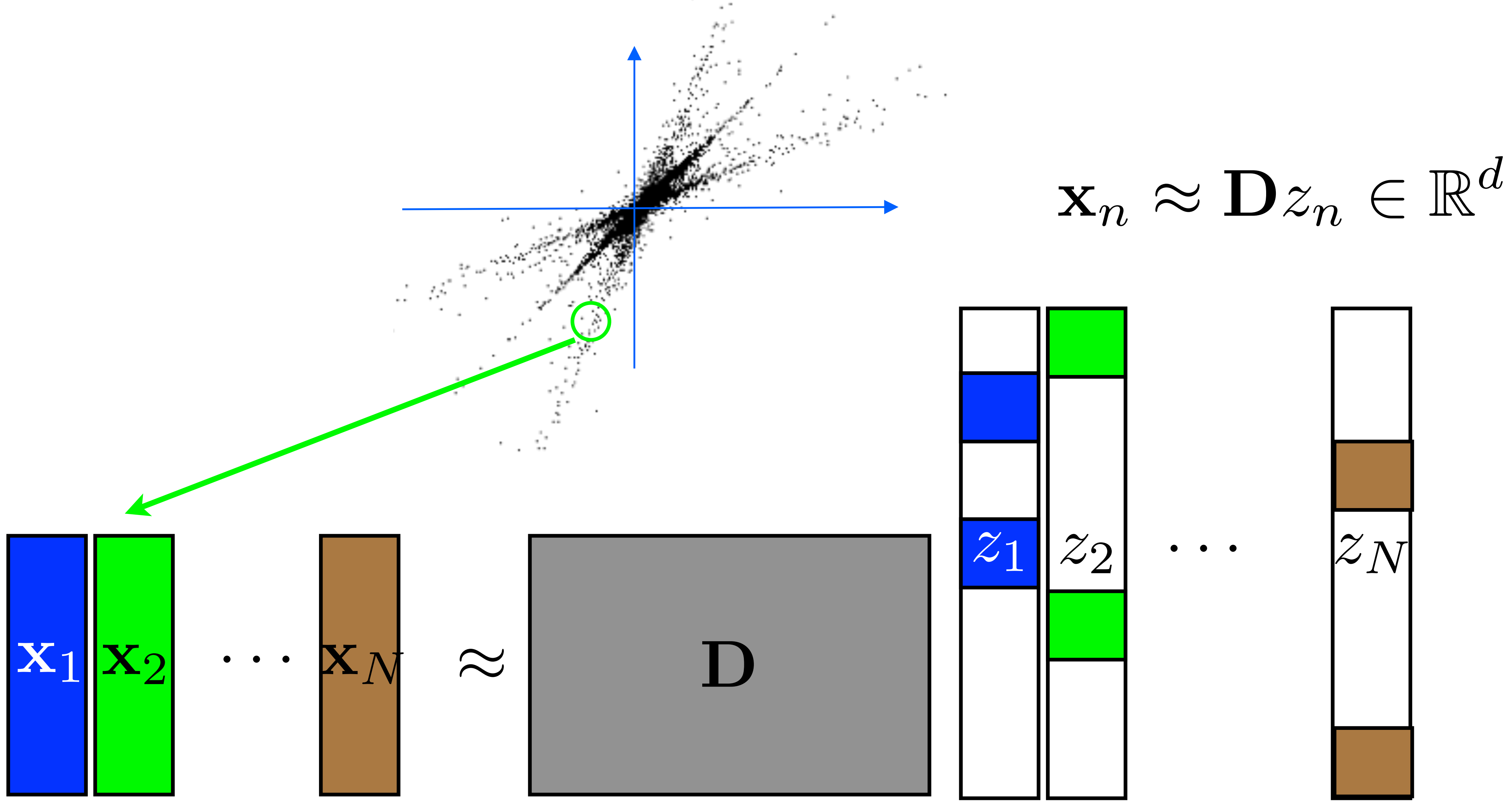
Dictionary Learning
= Sparse Matrix Factorization

- **Training collection = point cloud**



Dictionary Learning
= Sparse Matrix Factorization

- **Training collection = point cloud**



Dictionary Learning
= Sparse Matrix Factorization

$$\underset{d \times N}{\mathbf{X}} \approx \underset{d \times K}{\mathbf{D}} \underset{K \times N}{Z}$$

with *s*-sparse columns

Dictionary Learning
= Sparse Matrix Factorization

$$\mathbf{X} \approx \mathbf{D} Z$$

$d \times N$ $\quad$ $d \times K$ $\quad$ $K \times N$

with *s*-sparse columns

sounds familiar? similar to ICA! **X**=**AS**

Many Approaches

- Independent component analysis
  - [see e.g. book by Comon & Jutten 2011]
- Convex
  - [Bach et al., 2008; Bradley and Bagnell, 2009]
- Submodular
  - [Krause and Cevher, 2010]
- Bayesian
  - [Zhou et al., 2009]
- ***Non-convex optimization***
  - *[Olshausen and Field, 1997; Pearlmutter & Zibulevsky 2001, Aharon et al. 2006; Lee et al., 2007; Mairal et al., 2010 (... and* many *other authors)]*

Nonconvex optimization for dictionary learning

Sparse Coding Objective Function

- **Given one training sample, known D:**
  - ✓ sparse regression

$$f_{\mathbf{x}_n}(\mathbf{D}) = \min_{z_n} \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}z_n\|_2^2 + \phi(z_n)$$

- **Examples:**
  - ✦ **LASSO/Basis Pursuit:**

$$\phi(z) = \lambda\|z\|_1$$

  - ✦ **Ideal s-sparse approximation:**

$$\phi(z) = \chi_s(z) = \begin{cases} 0, & \|z\|_0 \leq s; \\ +\infty, & \text{otherwise} \end{cases}$$

Sparse Coding Objective Function

- **Given one training sample, known D:**
  - ✓ sparse regression

$$f_{\mathbf{x}_n}(\mathbf{D}) = \min_{z_n} \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}z_n\|_2^2 + \phi(z_n)$$

- **Given N training samples, unknown D:**

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

$$\propto \min_{Z} \frac{1}{2}\|\mathbf{X} - \mathbf{D}Z\|_F^2 + \Phi(Z)$$

Learning = *Constrained* Minimization

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D} \in \mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$

$$\propto \min_{Z} \frac{1}{2} \|\mathbf{X} - \mathbf{D}Z\|_F^2 + \Phi(Z)$$

- **Without constraint set** $\mathcal{D}$: degenerate solution

$$\mathbf{D} \to \infty, Z \to 0$$

- **Typical constraint** = unit-norm columns

$$\mathcal{D} = \{\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K], \ \forall k \ \|\mathbf{d}_k\|_2 = 1\}$$

Algorithms for dictionary learning

Principle: Alternate Optimization

- **Global objective** $$\min_{\mathbf{D},Z} \frac{1}{2}\|\mathbf{X}-\mathbf{D}Z\|_F^2 + \Phi(Z)$$
- **Alternate two steps**
  - ✓ ***Update coefficients*** given current dictionary **D**
    $$\min_{z_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}z_i\|_F^2 + \phi(z_i)$$
  - ✓ ***Update dictionary*** given current coefficients *Z*
    $$\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{X}-\mathbf{D}Z\|_F^2$$

Coefficient Update = Sparse Coding

- **Objective** $$\min_{z_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}z_i\|_F^2 + \phi(z_i)$$
- **Two strategies**
  - ✓ **Batch:** for *all* training samples *i* at each iteration
  - ✓ **Online**: for *one* (randomly selected) training sample *i*
- **Implementation: sparse coding algorithm**
  - ✓ L1 minimization , (Orthonormal) Matching Pursuit, ...

Dictionary Update

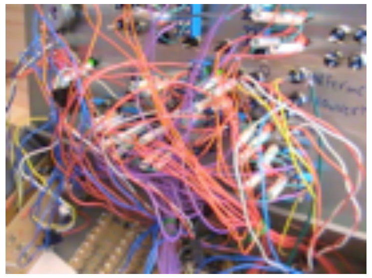
- **Objective** $$\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}Z\|_F^2$$
- **Main approaches**
  - ✓ Method of Optimal Directions (MOD) *[Engan et al., 1999]*

$$\hat{\mathbf{D}} = \mathbf{X} \cdot \mathrm{pinv}(Z) \; = \arg\min_{\mathbf{D}} \|\mathbf{X} - \mathbf{D}Z\|_F^2$$

  - ✓ K-SVD: with PCA *[Aharon et al. 2006]*
    - ✦ coefficients are jointly updated
  - ✓ Online L1: stoch. gradient *[Engan & al 2007, Mairal et al., 2010]*

... but also

- **Related «learning» matrix factorizations**
  - ✓ Non-negativity (NMF):
    - ✦ **Multiplicative update** *[Lee & Seung 1999]*
  - ✓ Known rows *up to gains* (blind calibration) $\mathbf{D} = \mathrm{diag}(\mathbf{g})\mathbf{D}_0$
    - ✦ **Convex formulation** *[G & al 2012, Bilen & al 2013]*
  - ✓ Know-rows *up to permutation* (cable chaos) $\mathbf{D} = \mathbf{\Pi}\mathbf{D}_0$
    - ✦ **Branch & bound** *[Emiya & al, 2014]*



- **(Approximate) Message Passing** *[Krzakala & al, 2013]*

Statistical guarantees

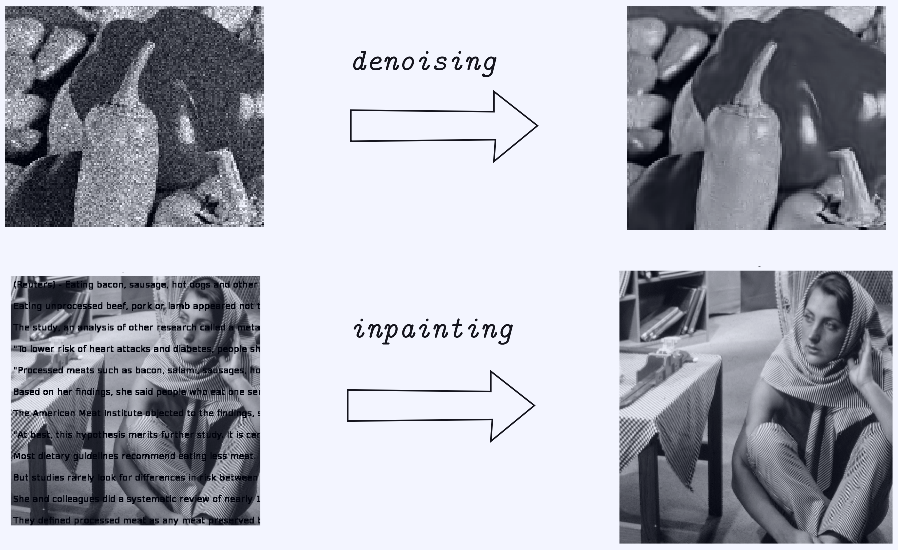
Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathrm{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

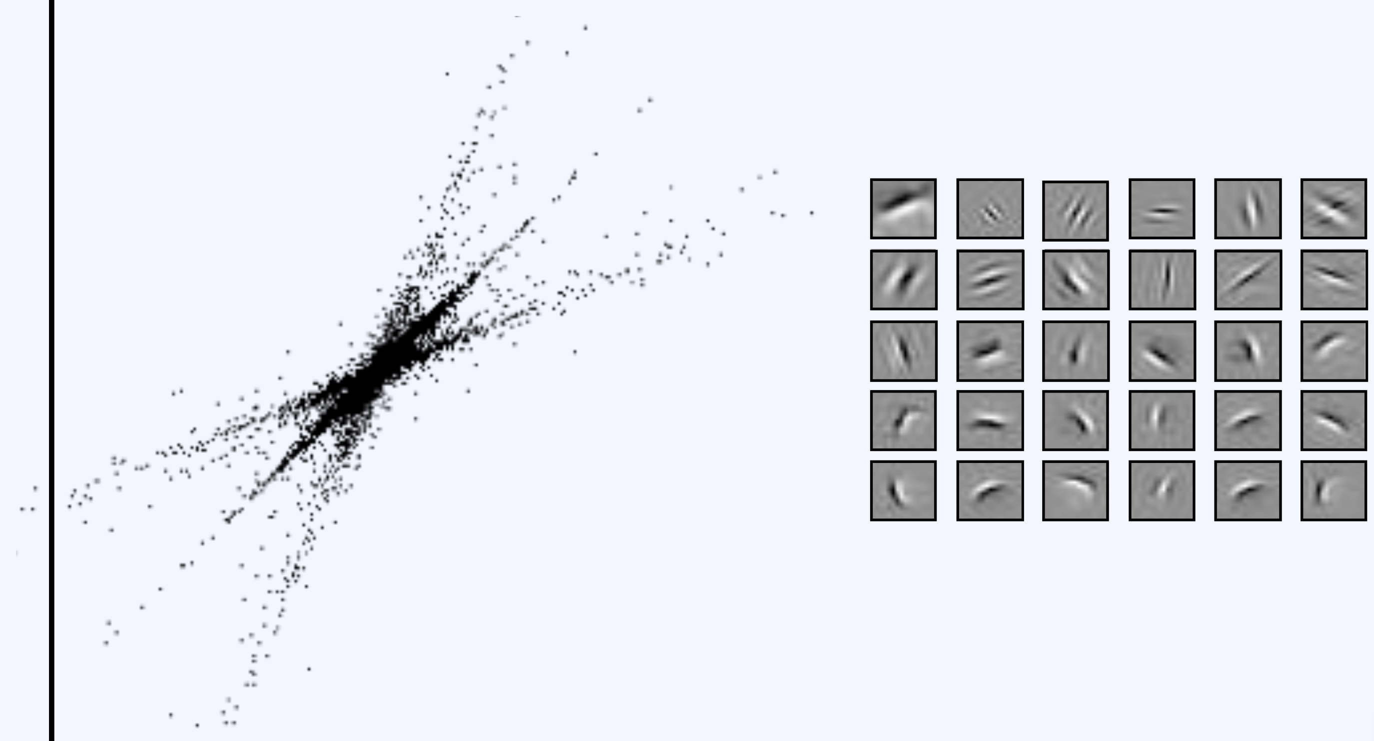
Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...



Source localization, neural coding ...



Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...

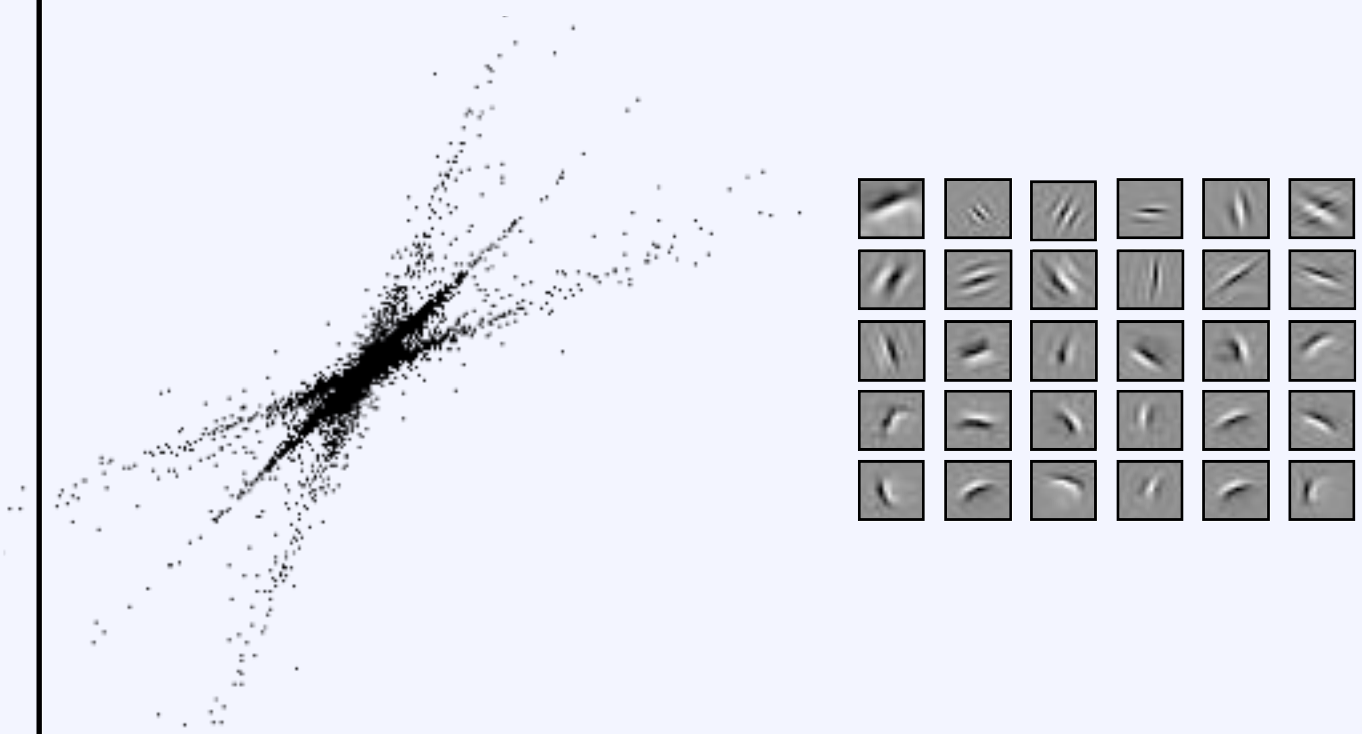
✓ No «ground truth dictionary»

✓ Goal = performance generalization

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***

Source localization, neural coding ...



Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...

✓ No «ground truth dictionary»

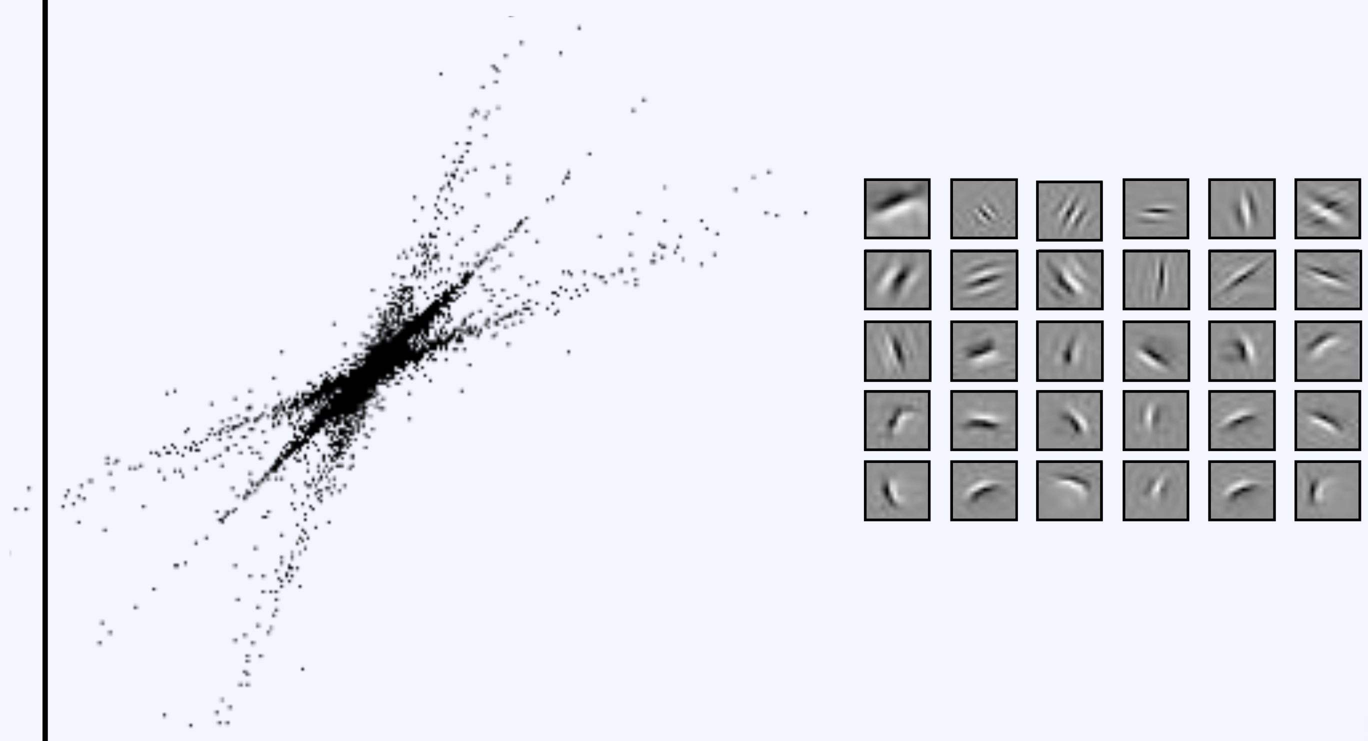
✓ Goal = performance generalization

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***

- **Excess risk analysis** (~Machine Learning)
  - *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*

Source localization, neural coding ...



Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...

✓ No «ground truth dictionary»

✓ Goal = performance generalization

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***
- **Excess risk analysis** (~Machine Learning)
  - *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*

Source localization, neural coding ...

✓ Ground truth $\mathbf{x} = \mathbf{D}_0 z + \varepsilon$

✓ Goal = dictionary estimation

$$\|\hat{\mathbf{D}}_N - \mathbf{D}_0\|_F$$

- ***What recovery conditions ?***

Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...

✓ No «ground truth dictionary»

✓ Goal = performance generalization

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***
- **Excess risk analysis** (~Machine Learning)
  - *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*

Source localization, neural coding ...

✓ Ground truth $\mathbf{x} = \mathbf{D}_0 z + \varepsilon$

✓ Goal = dictionary estimation

$$\|\hat{\mathbf{D}}_N - \mathbf{D}_0\|_F$$

- ***What recovery conditions ?***
- ***Identifiability analysis*** (~Signal Processing)
  - *[Independent Component Analysis, e.g. book Comon & Jutten 2011]*

Theorem: Excess Risk Control

- **Assume:**
  - ✓ **X** obtained from *N* draws, i.i.d., bounded $\mathbb{P}(\|\mathbf{x}\|_2 \leq 1) = 1$
  - ✓ Penalty function $\phi(z)$
    - ✦ **non-negative** and minimum at zero
    - ✦ **lower semi-continuous**
    - ✦ **coercive**
  - ✓ Constraint set $\mathcal{D}$: (upper box-counting) dimension *h*
    - ✦ *typically: h = dK   d = signal dimension, K = number of atoms*

[G. & al, Sample Complexity of Dictionary Learning and Other Matrix Factorizations, 2013, arXiv/HAL]

Theorem: Excess Risk Control

- **Assume:**
  - ✓ **X** obtained from *N* draws, i.i.d., bounded $\mathbb{P}(\|\mathbf{x}\|_2 \leq 1) = 1$
  - ✓ Penalty function $\phi(z)$
    - ✦ **non-negative** and minimum at zero
    - ✦ **lower semi-continuous**
    - ✦ **coercive**
  - ✓ Constraint set $\mathcal{D}$: (upper box-counting) dimension *h*
    - ✦ *typically: h = dK   d = signal dimension, K = number of atoms*
- **Then**: with probability at least $1 - 2e^{-x}$ on **X**

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N \qquad \eta_N \leq C\sqrt{\frac{(h+x)\log N}{N}}$$

[G. & al, Sample Complexity of Dictionary Learning and Other Matrix Factorizations, 2013, arXiv/HAL]

Sample Complexity of Matrix Factorizations

- **General penalty functions**
  - l1 norm / mixed norms / lp quasi-norms
  - ... *but also* non-coercive penalties (with additional RIP on constraint set):
    - s-sparse constraint, non-negativity
- **General constraint sets**
  - unit norm / sparse / shift-invariant / tensor product / tight frame ...
  - «complexity» captured by box-counting dimension
- **«Distribution free»**
  - bounded samples $\mathbb{P}(\|\mathbf{x}\|_2 \leq 1) = 1$
  - ... *but also* sub-Gaussian $\mathbb{P}(\|\mathbf{x}\|_2 \geq At) \leq \exp(-t), \quad t \geq 1$
- **Selected covered examples:**
  - **PCA / NMF / K-Means / sparse PCA**

Analytic vs Learned Dictionaries
Learning Fast Transforms

Ph.D. of Luc Le Magoarou

Analytic *vs* Learned Dictionaries

Dictionary	Adaptation to Training Data
Analytic (Fourier, wavelets, ...)	**No**
Learned	**Yes**

Analytic *vs* Learned Dictionaries

Dictionary	Adaptation to Training Data	Computational Complexity
Analytic (Fourier, wavelets, ...)	**No**	**Low**
Learned	**Yes**	**High**

Analytic *vs* Learned Dictionaries

Dictionary	Adaptation to Training Data	Computational Complexity
Analytic (Fourier, wavelets, ...)	**No**	**Low**
Learned	**Yes**	**High**

Best of both worlds ?

Sparse-KSVD

- **Principle: constrained dictionary learning**
  - ✓ choose reference (fast) dictionary $\mathbf{D}_0$
  - ✓ learn with the constraint: $\mathbf{D} = \mathbf{D}_0\mathbf{S}$ where $\mathbf{S}$ is sparse
- **Resulting *double-sparse* factorization problem**

$$\mathbf{X} \approx \mathbf{D}_0\mathbf{S}Z$$

- [R. Rubinstein, M. Zibulevsky & M. Elad, "*Double Sparsity: Learning Sparse Dictionaries for Sparse Signal Approximation*," IEEE TSP, vol. 58, no. 3, pp. 1553–1564.

Sparse-KSVD

- **Principle: constrained dictionary learning**
  - ✓ choose reference (fast) dictionary $\mathbf{D}_0$ **strong prior!**
  - ✓ learn with the constraint: $\mathbf{D} = \mathbf{D}_0\mathbf{S}$ where $\mathbf{S}$ is sparse

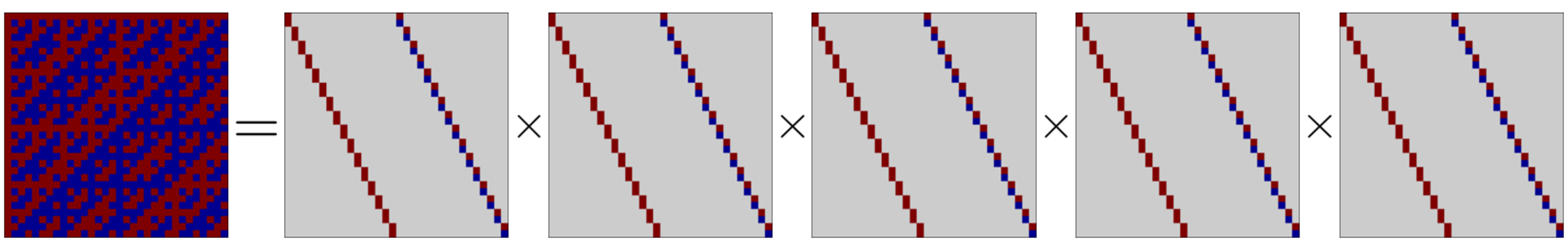
- **Resulting *double-sparse* factorization problem**

$$\mathbf{X} \approx \mathbf{D}_0\mathbf{S}Z$$

- [R. Rubinstein, M. Zibulevsky & M. Elad, "*Double Sparsity: Learning Sparse Dictionaries for Sparse Signal Approximation*," IEEE TSP, vol. 58, no. 3, pp. 1553–1564.

Speed = Factorizable Structure

- **Fourier**: FFT with butterfly algorithm
- **Wavelets**: FWT tree of filter banks
- **Hadamard**: Fast Hadamard Transform



$$\underbrace{\mathbf{D}}_{\mathcal{O}(1024)} = \underbrace{\mathbf{S}_1 \times \mathbf{S}_2 \times \mathbf{S}_3 \times \mathbf{S}_4 \times \mathbf{S}_5}_{\mathcal{O}(320)}$$

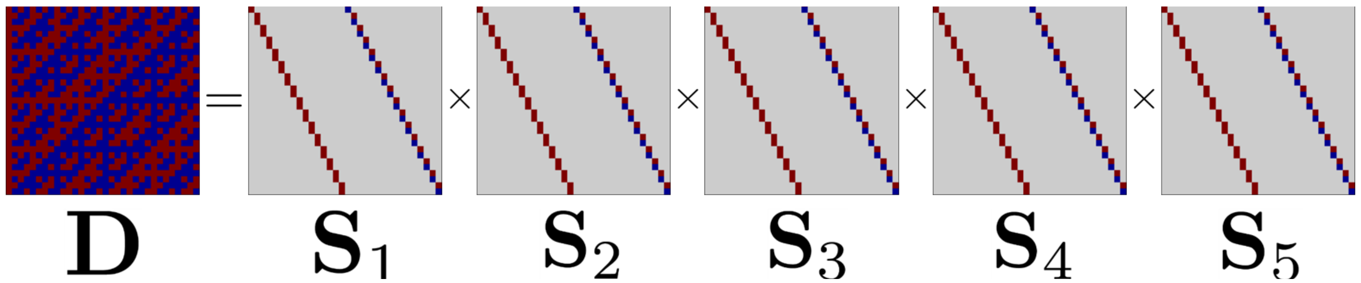
Learning Fast Transforms

- **Class $\mathcal{D}$ of dictionaries of the form** $$\mathbf{D} = \prod_{j=1}^{M} \mathbf{S}_j$$
  - ✓ covers standard fast transforms
  - ✓ more flexible, better adaptation to training data
  - ✓ reduced costs
    - ✦ storage cost: *compression*
    - ✦ sample complexity: *denoising*
    - ✦ computational complexity: *inverse problems and more*
- **Learning***:*
  - ✓ **Nonconvex optimization algorithm**: PALM
    - ✦ guaranteed convergence to stationary point
  - ✓ **Hierarchical strategy**

Example 1: *Reverse-Engineering* the Fast Hadamard Transform

- **Hadamard Dictionary: Reference Factorization**

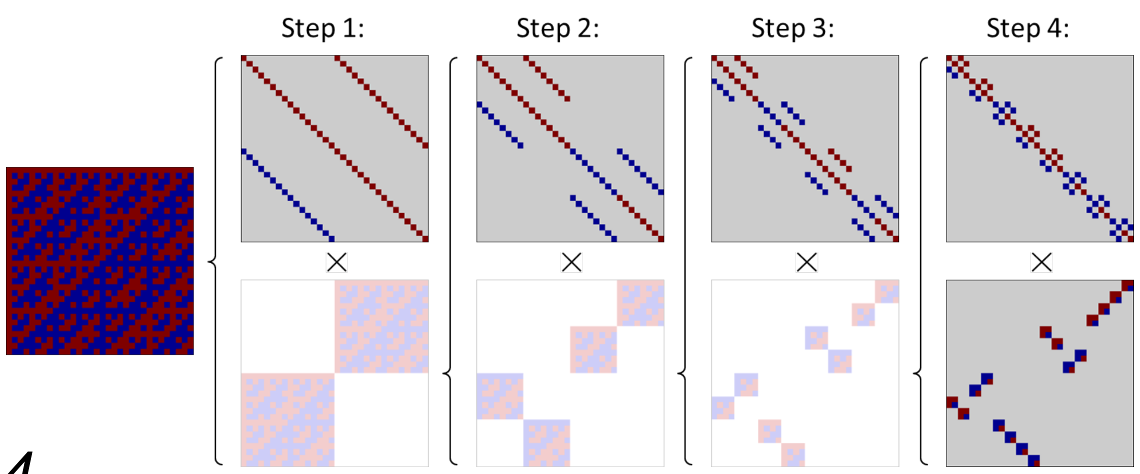
n^2



$2n \log_2 n$

- **Learned Factorization: different, *but as sparse***

n^2

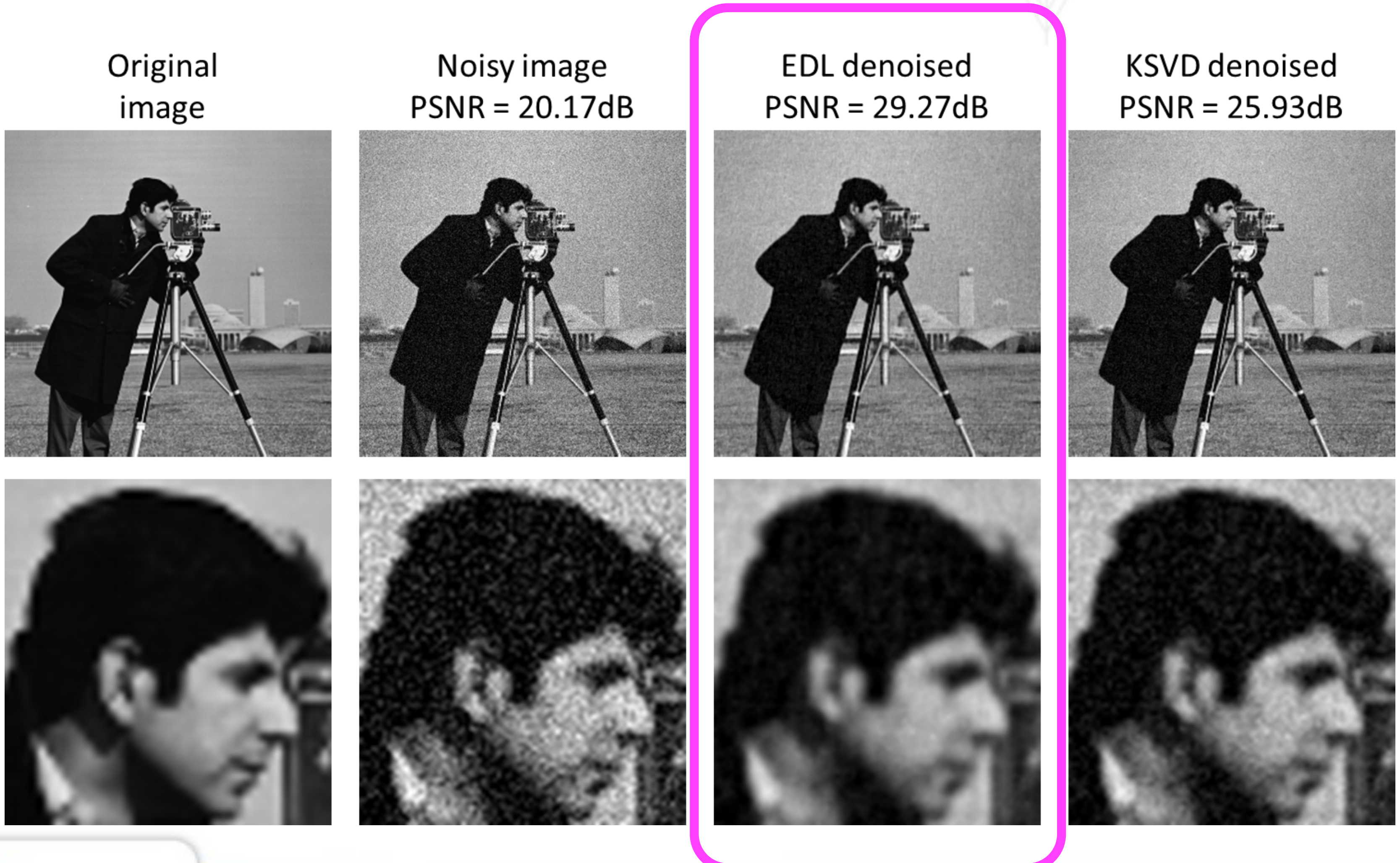


$2n \log_2 n$

tested up to n=1024

Example 2: Image Denoising with Learned Fast Transform

- **Patch-based dictionary learning (*n* = *8x8* pixels)**
- **Comparison using** SMALL box small-project.eu

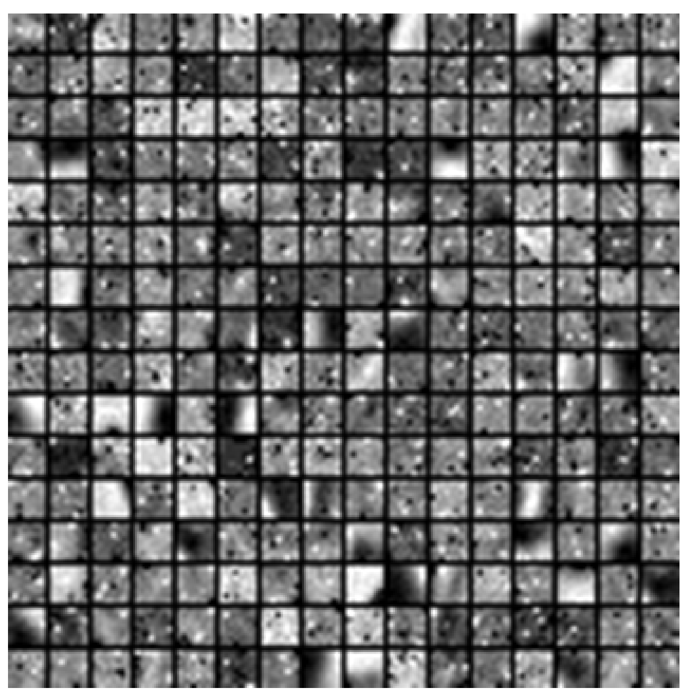


Example 2: Image Denoising with Learned Fast Transform

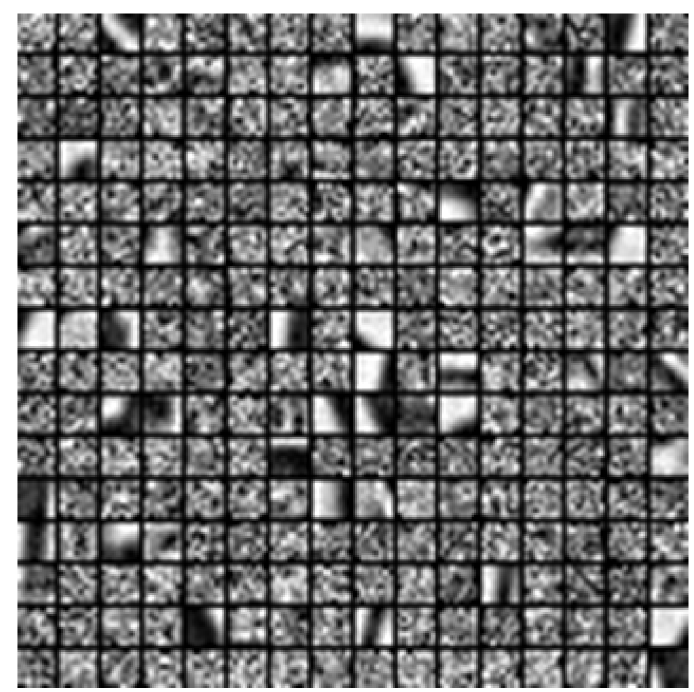
- **Patch-based dictionary learning (*n* = *8x8* pixels)**
- **Comparison using** SMALLbox small-project.eu
- **Learned dictionaries**

$\mathcal{O}(n \log_2 n)$

EDL Dictionary

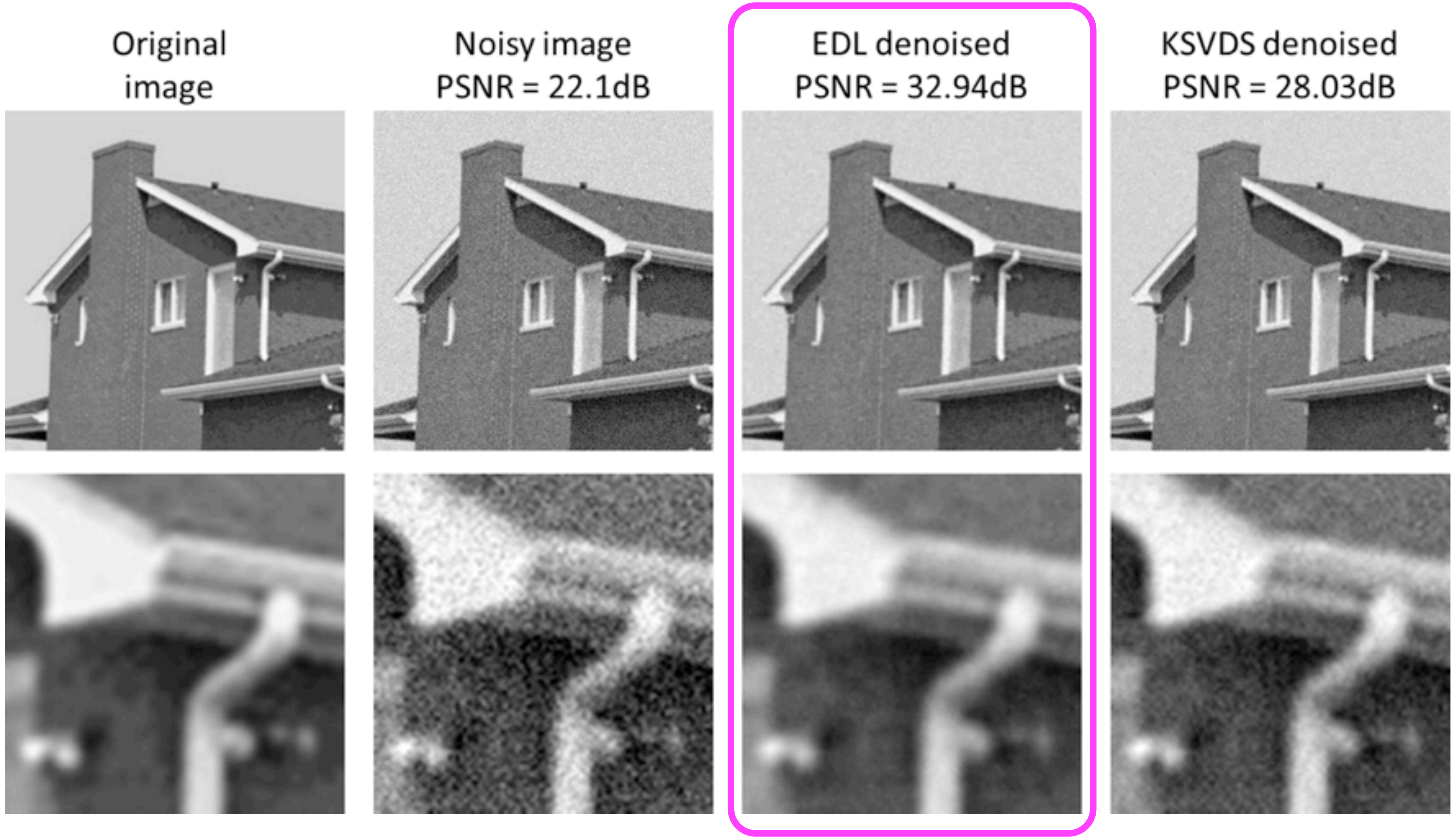


KSVD Dictionary

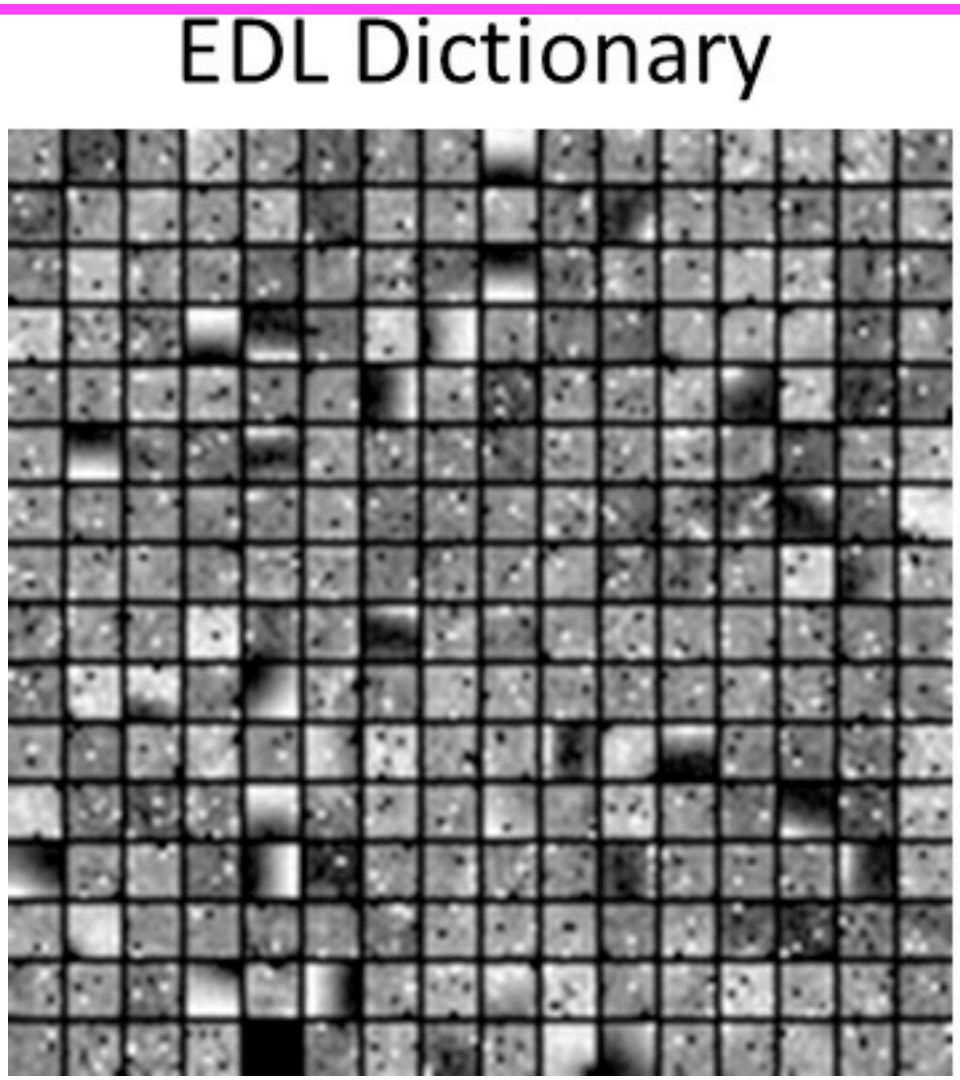


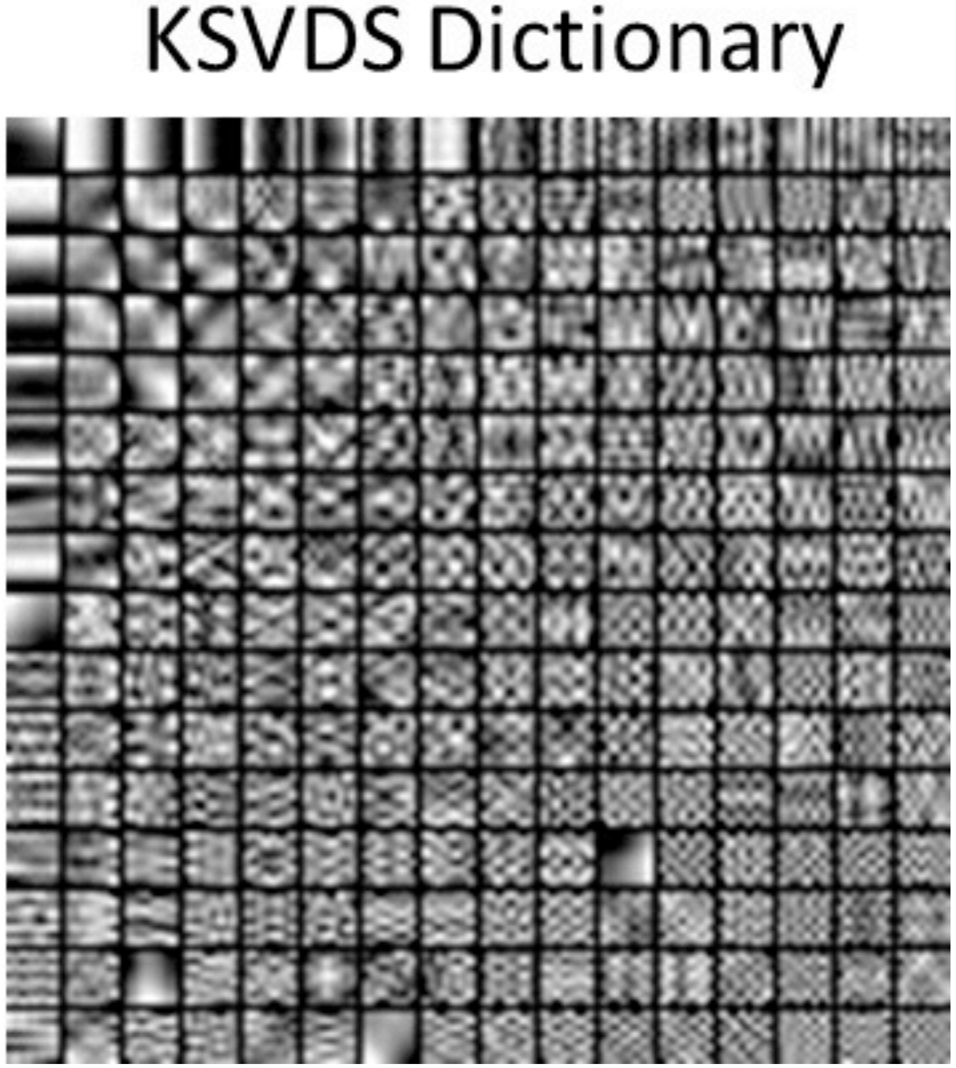
$\mathcal{O}(n^2)$

Comparison with Sparse KSVD (KSVDS)



Comparison with Sparse KSVD (KSVDS)





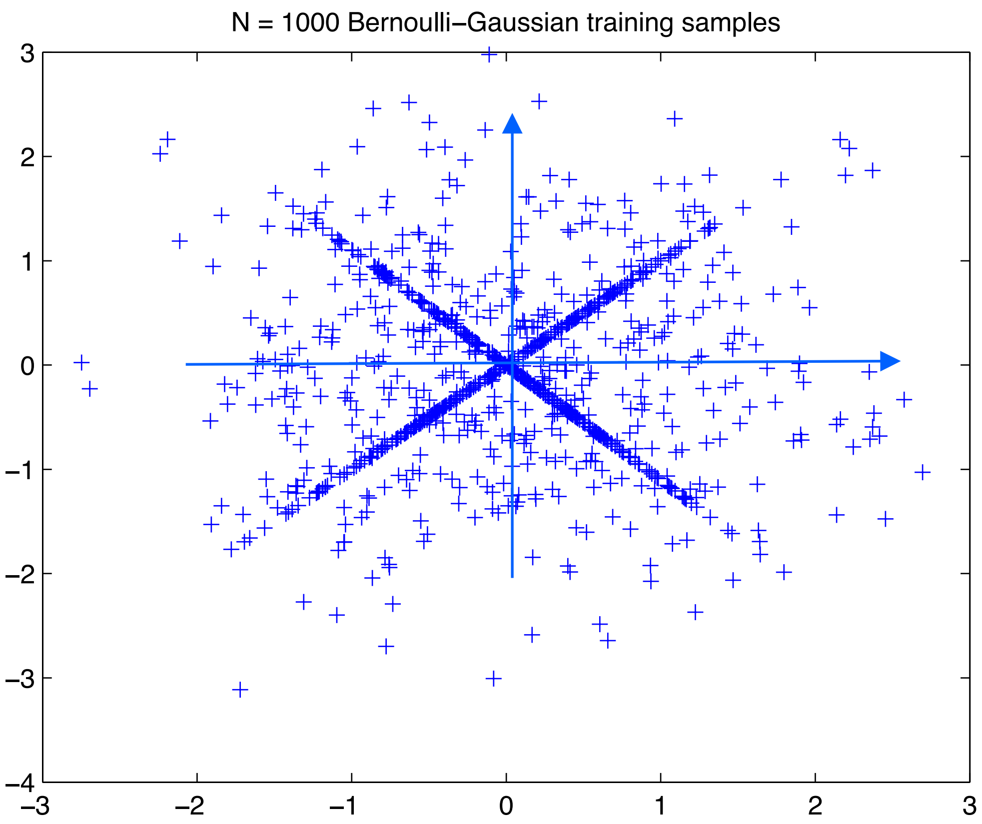
$\mathbf{D} = \mathbf{D}_0\mathbf{S}$

very close to D_0 = DCT

Identifiability analysis ?
Empirical findings

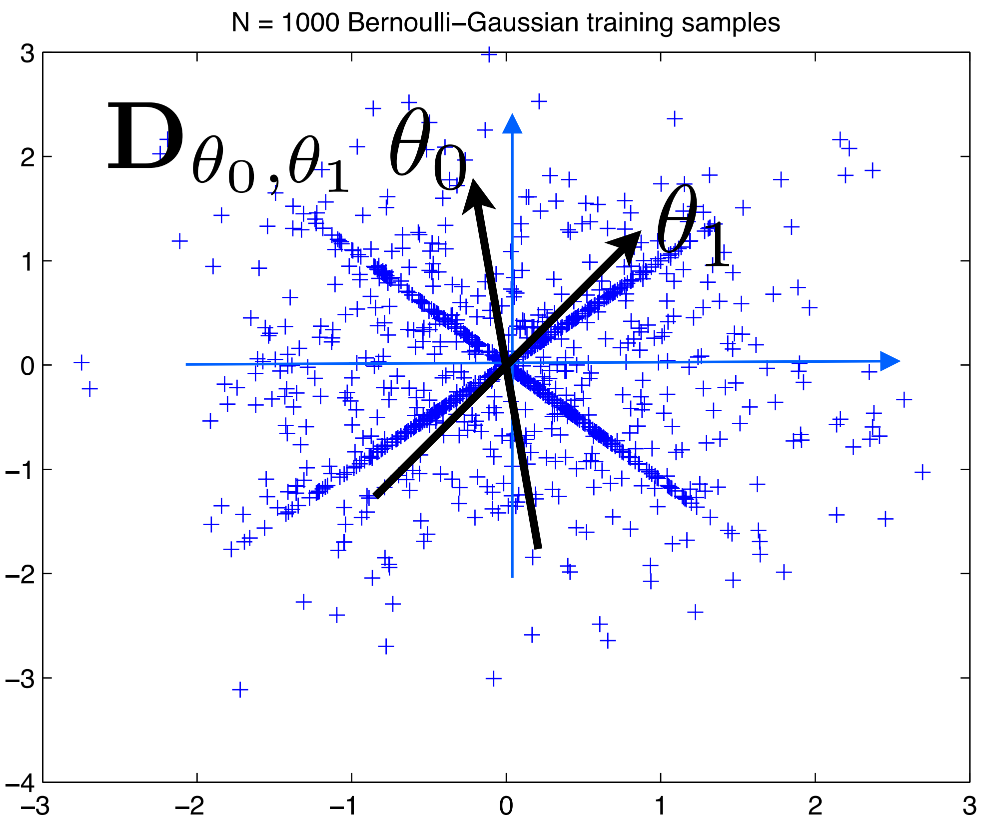
Numerical Example (2D)

$$\mathbf{X} = \mathbf{D}_0 Z_0$$



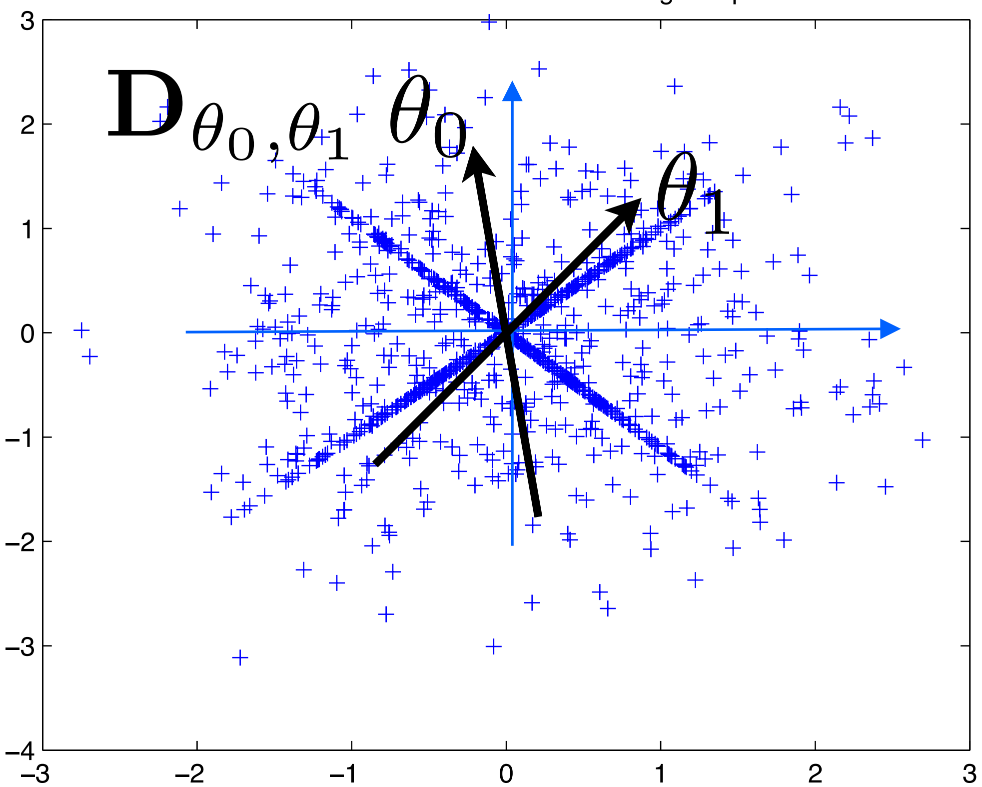
Numerical Example (2D)

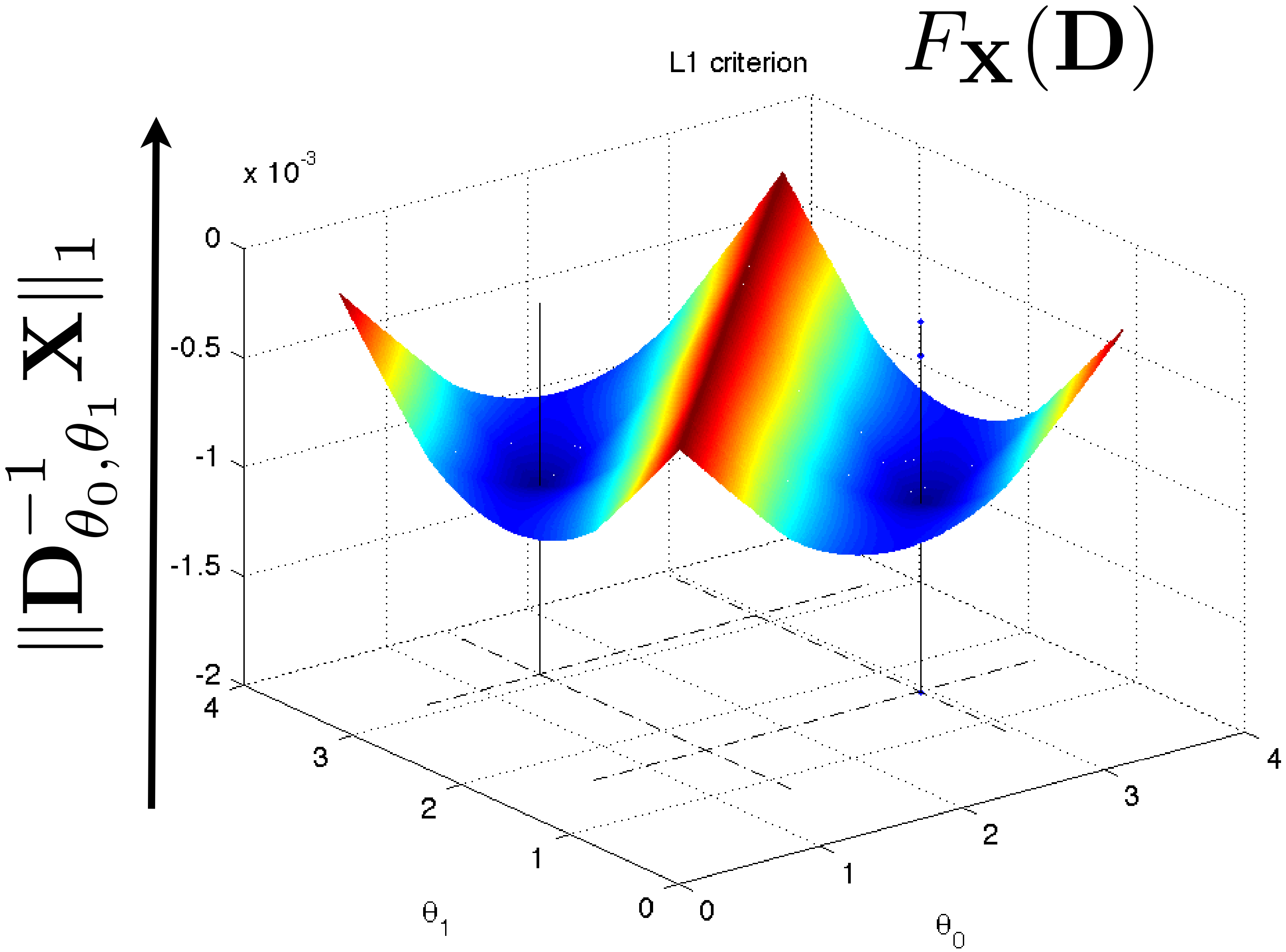
$$\mathbf{X} = \mathbf{D}_0 Z_0$$



Numerical Example (2D)

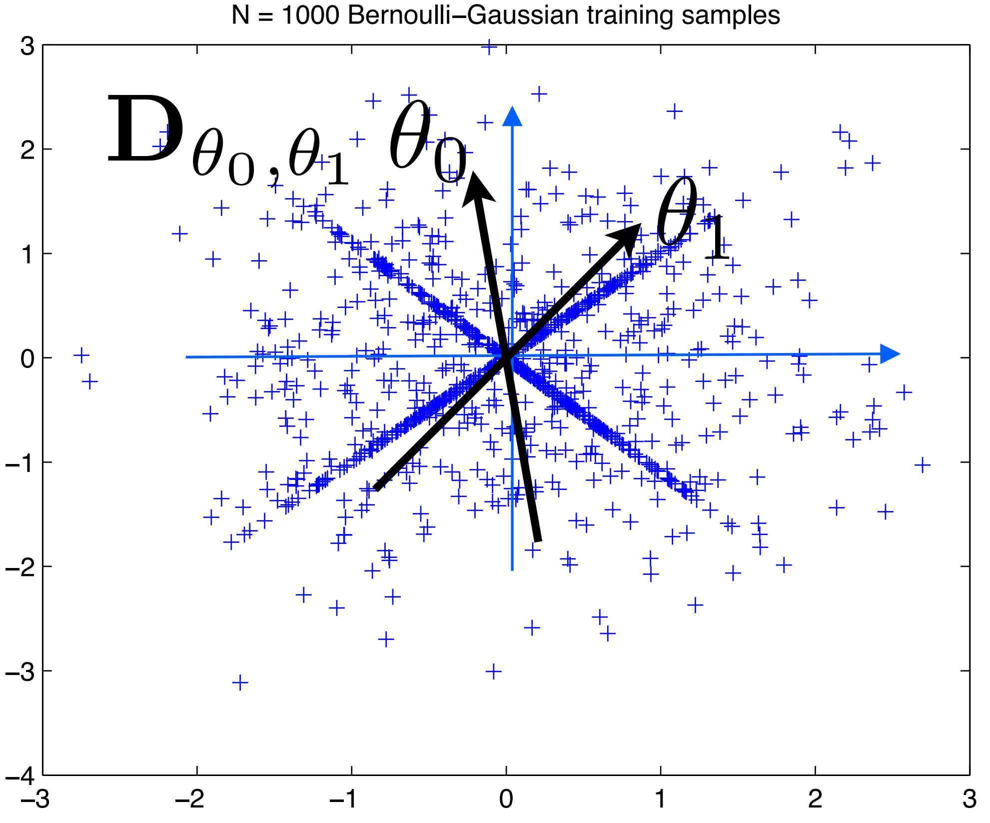
$$\mathbf{X} = \mathbf{D}_0 Z_0$$



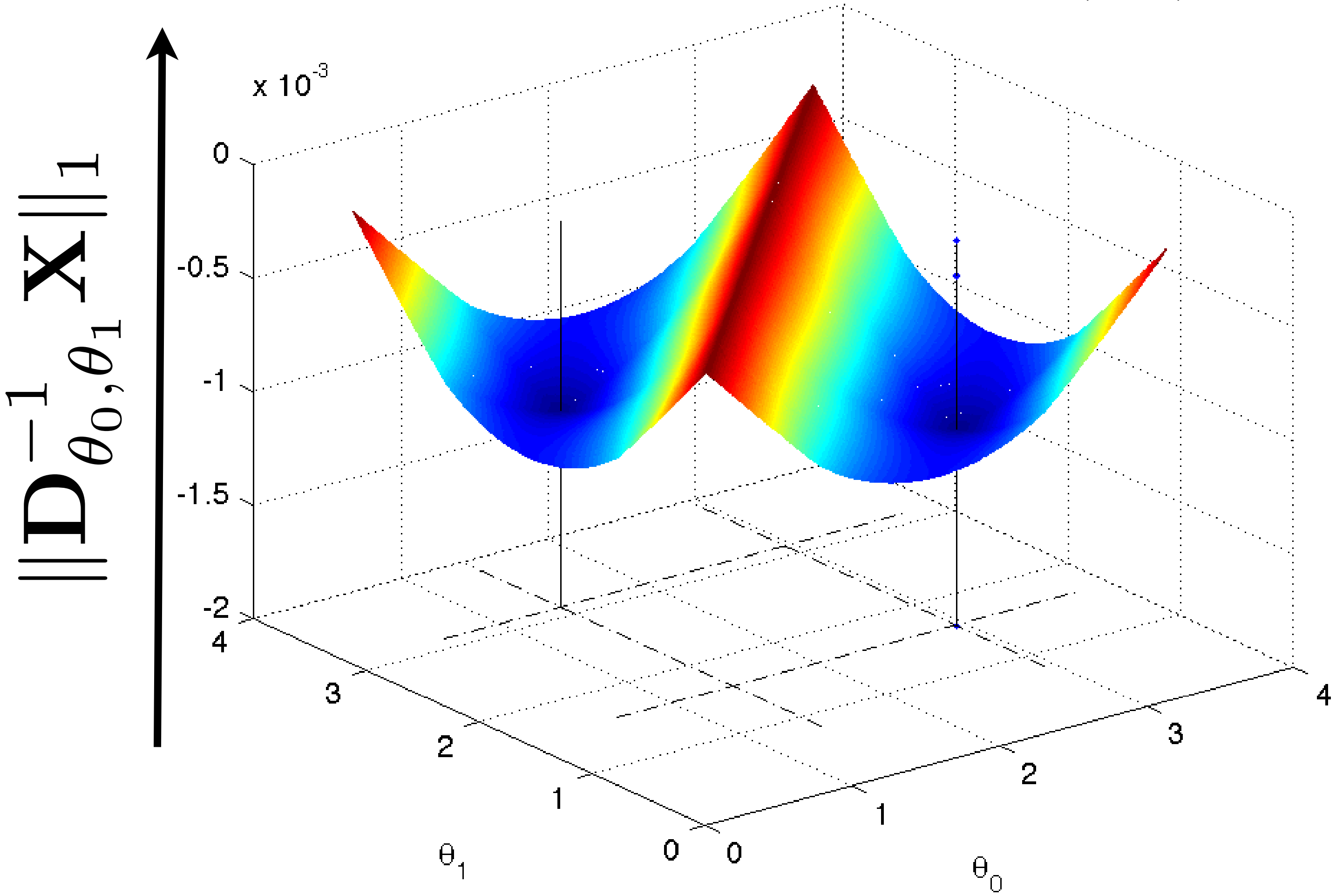


Numerical Example (2D)

$$\mathbf{X} = \mathbf{D}_0 Z_0$$



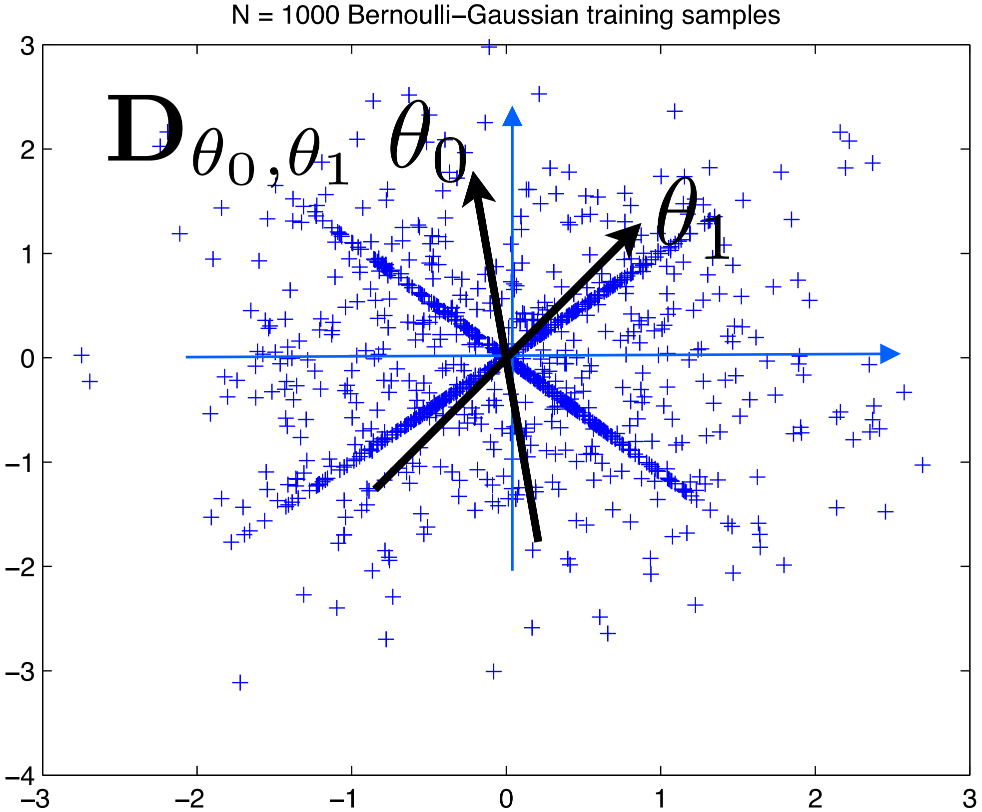
$F_{\mathbf{X}}(\mathbf{D})$

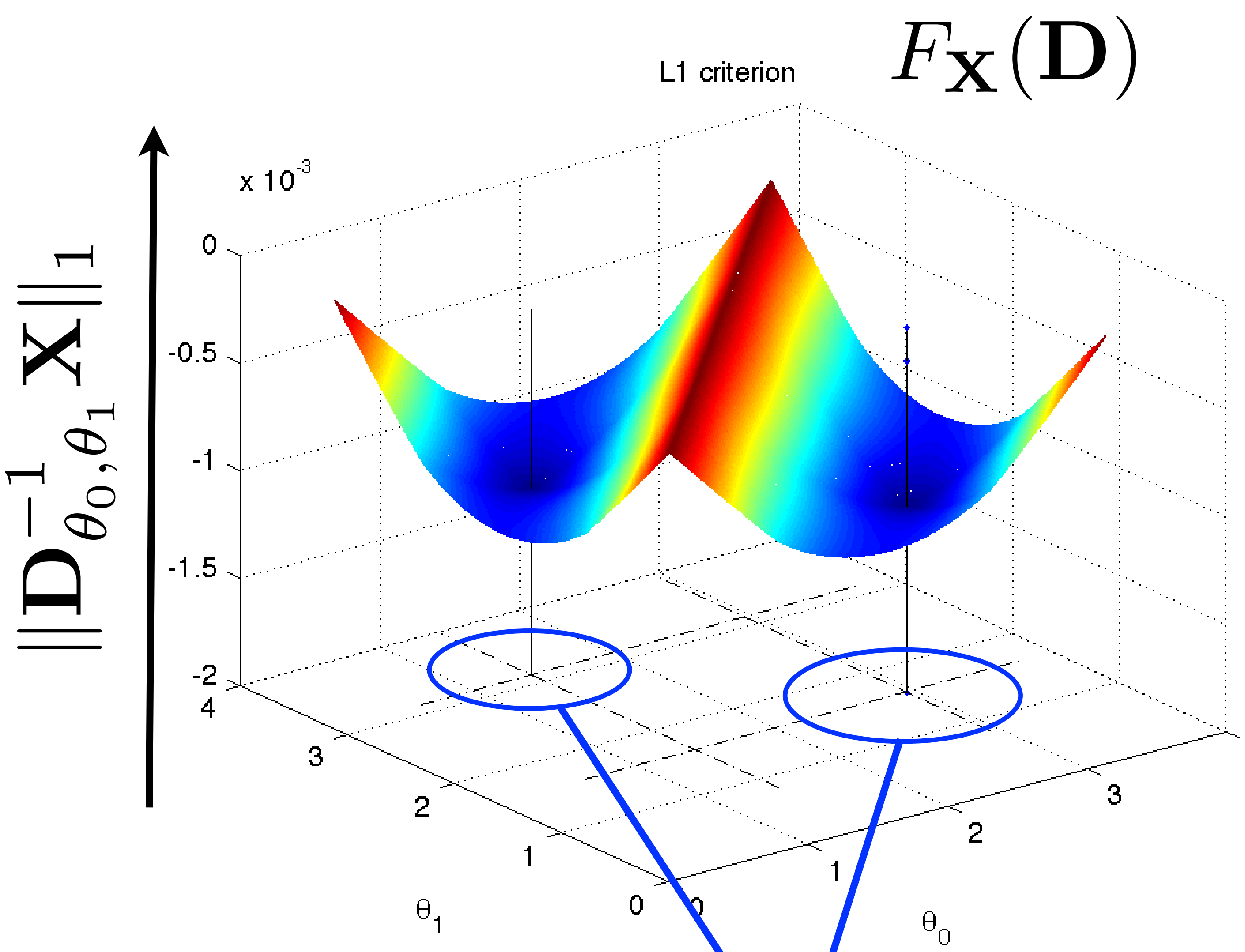


Symmetry = permutation ambiguity

Numerical Example (2D)

$$\mathbf{X} = \mathbf{D}_0 Z_0$$



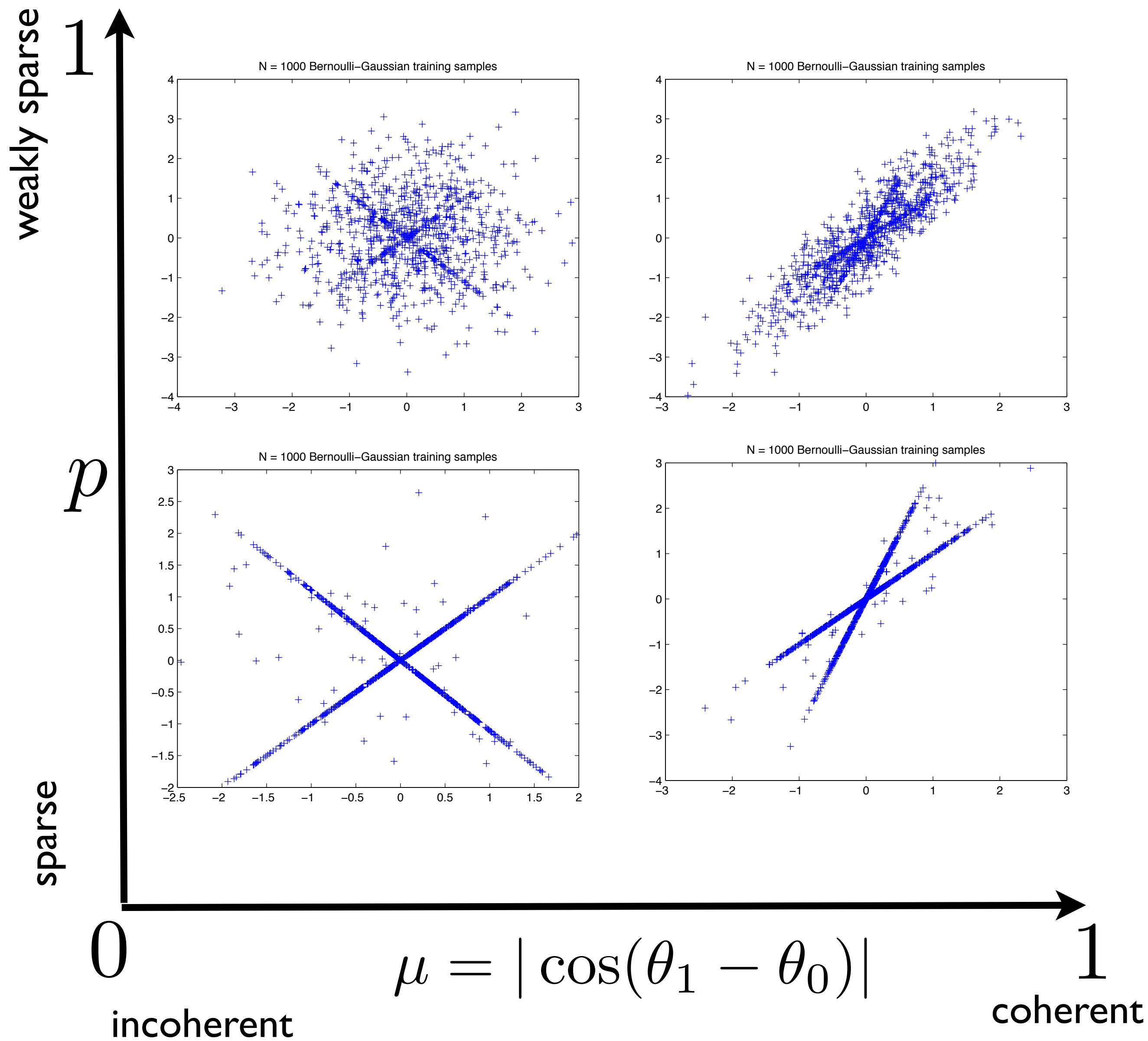


Empirical observations

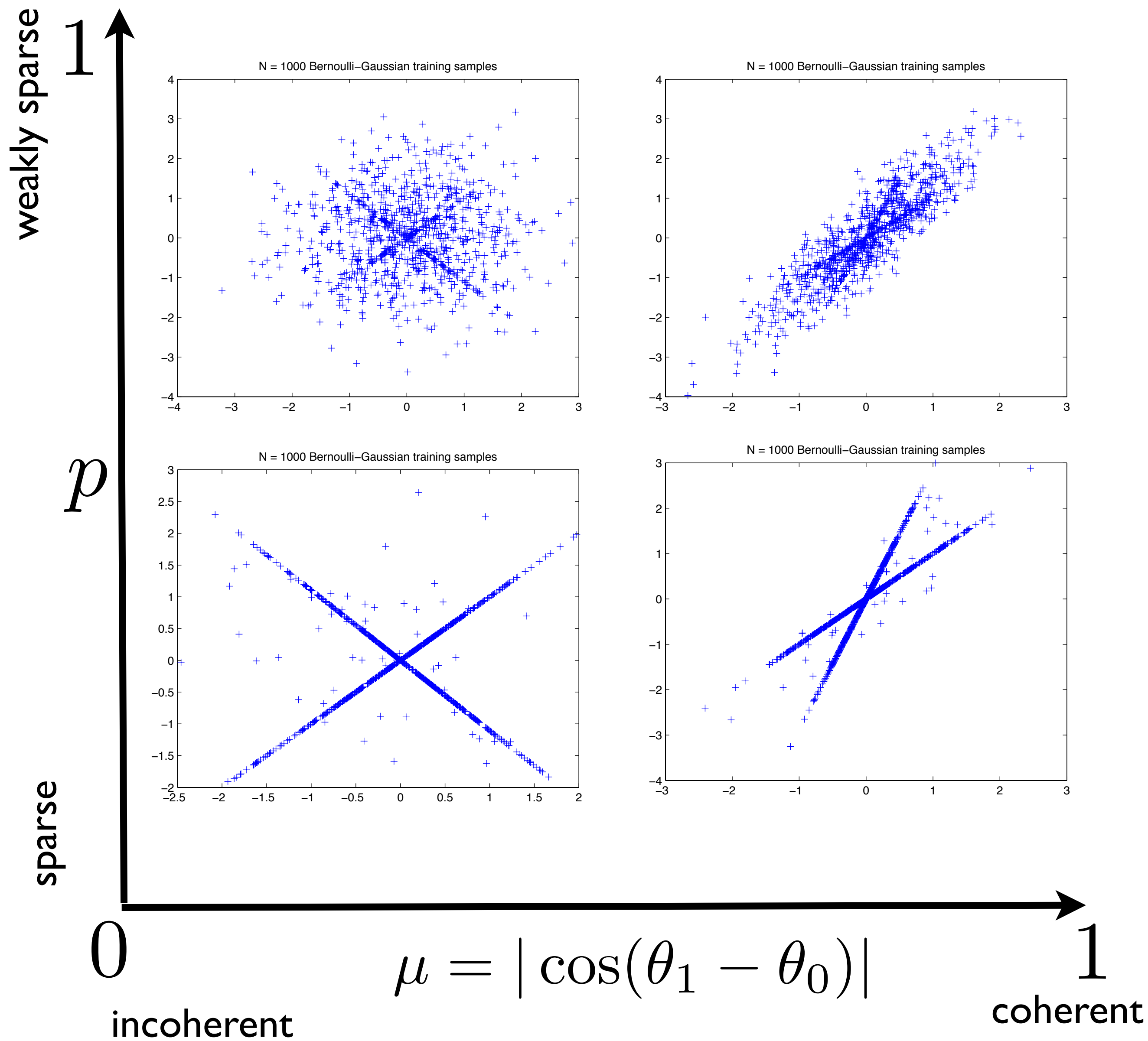
a) Global minima match angles of the original basis

b) There is no other local minimum.

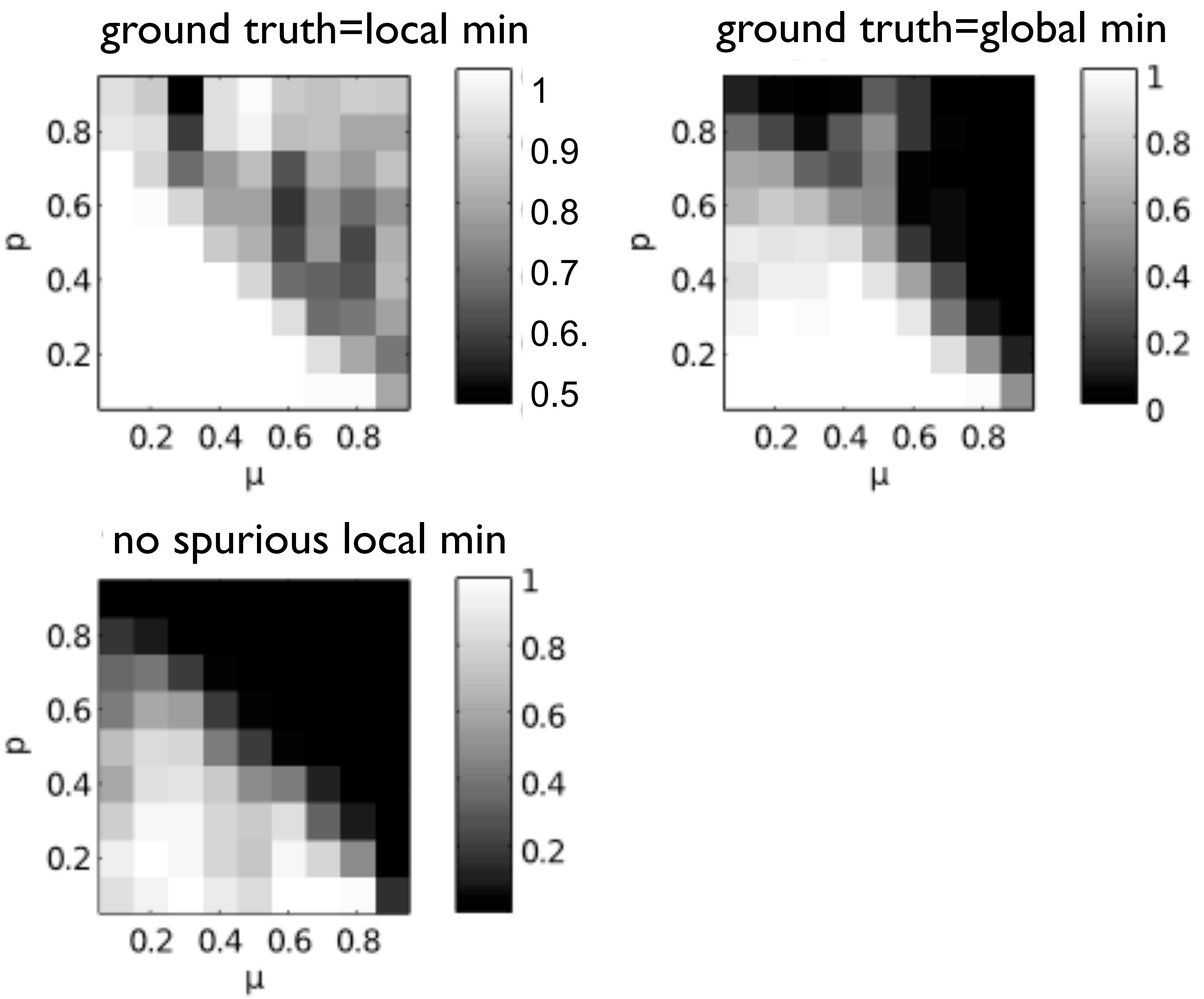
Sparsity *vs* Coherence (2D)



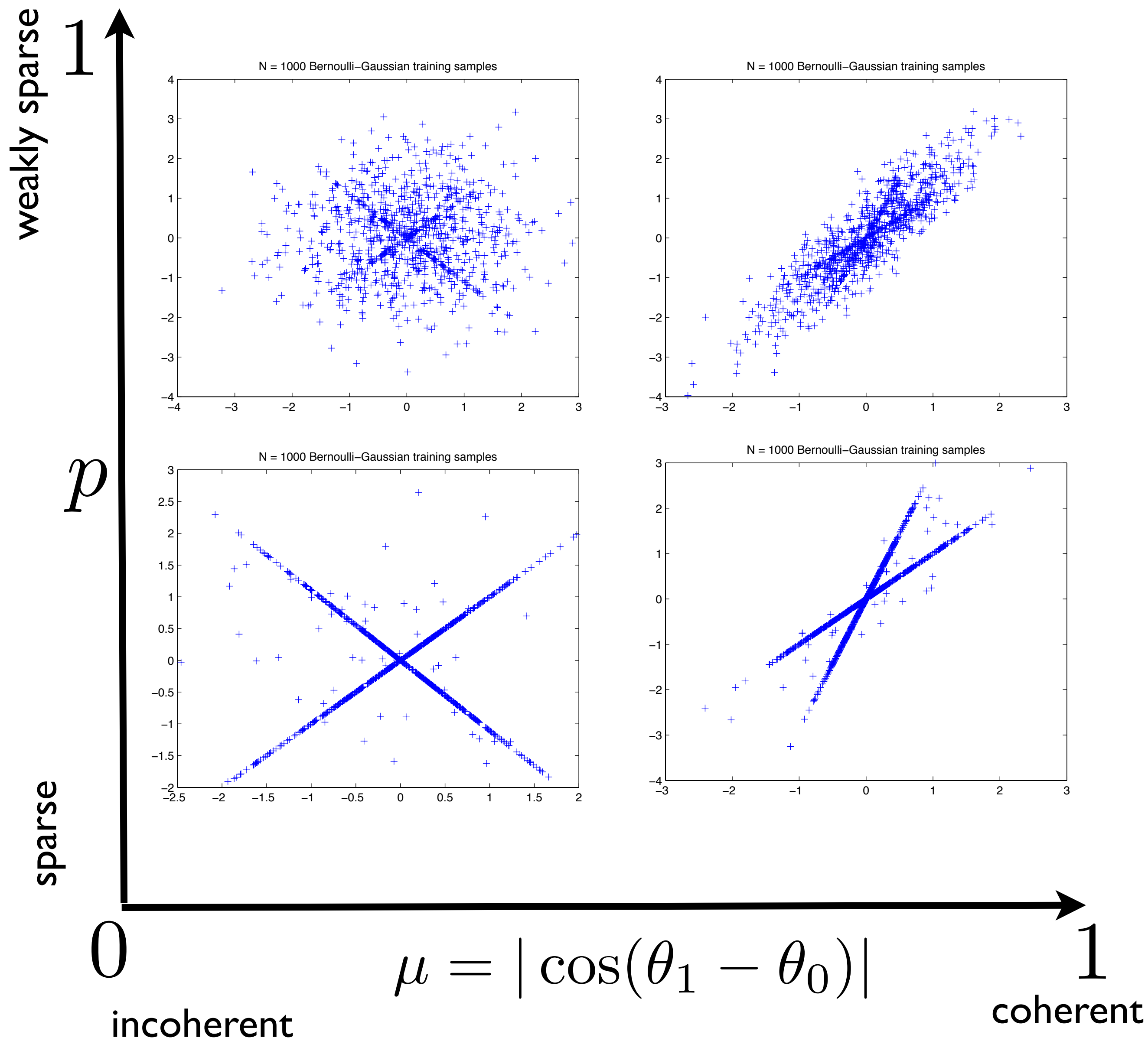
Sparsity *vs* Coherence (2D)



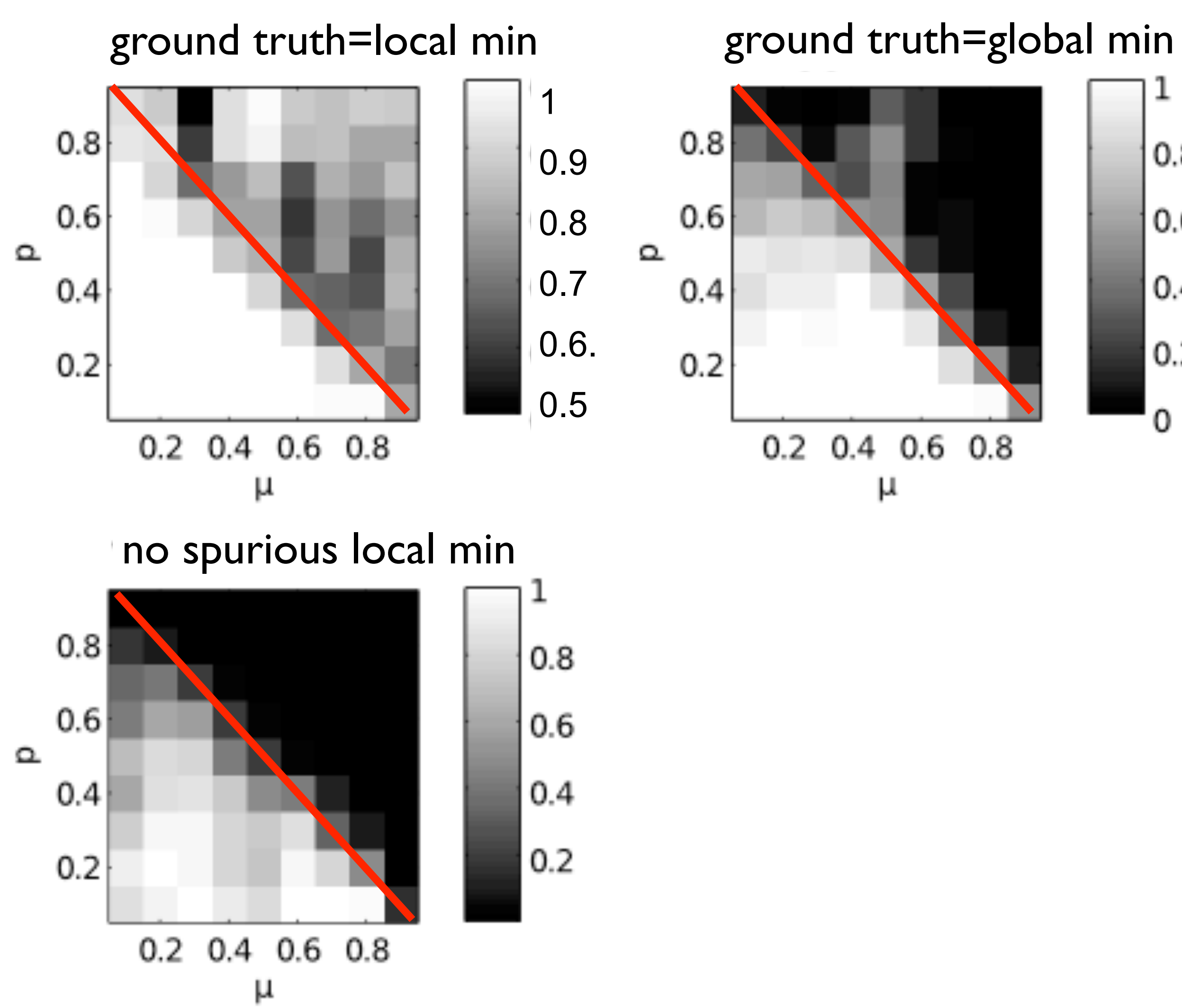
Empirical probability of success



Sparsity *vs* Coherence (2D)



Empirical probability of success



Rule of thumb: perfect recovery if:

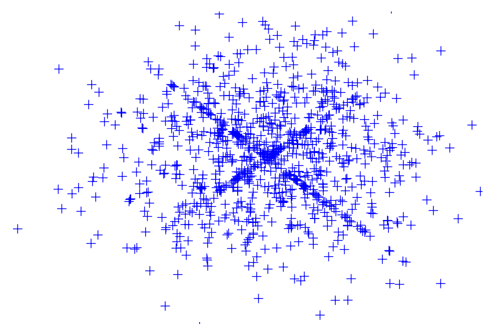
a) Incoherence $\mu < 1 - p$

b) Enough training samples (N large enough)

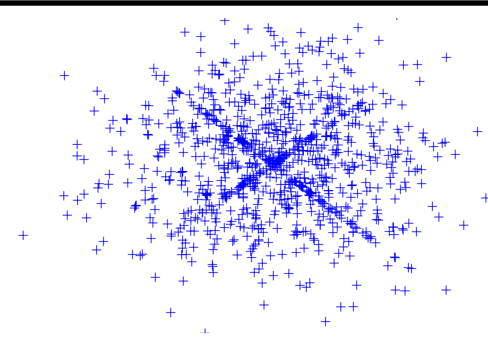
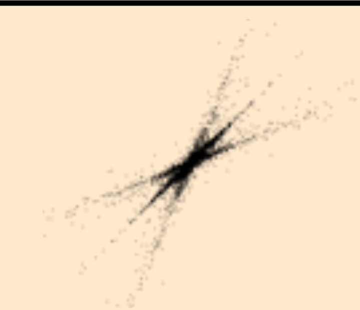
Empirical Findings

- **Stable & robust dictionary identification**
  - ✓ Global minima often match ground truth
  - ✓ Often, there is no spurious local minimum
- **Role of parameters ?**
  - ✓ *sparsity* level ?
  - ✓ *incoherence* of **D** ?
  - ✓ *noise* level ?
  - ✓ presence / nature of *outliers* ?
  - ✓ *sample complexity* (number of training samples) ?

Identifiability Analysis: Overview

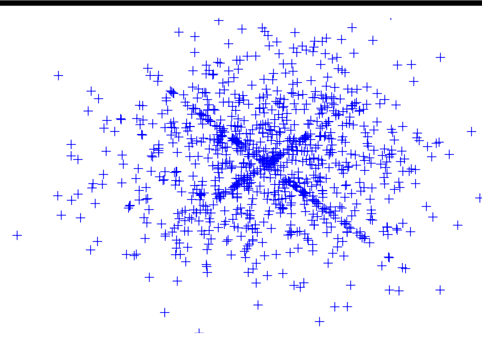
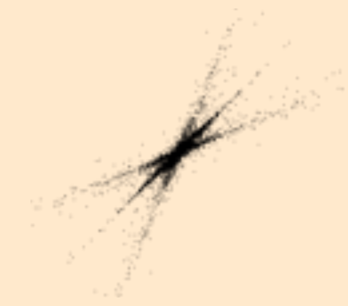
	[G. & Schnass 2010]	[Geng & al 2011]
signal model		
overcomplete (d<K)	no	**yes**
outliers	**yes**	no
noise	no	
cost function	$\min_{\mathbf{D},Z} \|Z\|_1 \text{ s.t. } \mathbf{D}Z = \mathbf{X}$	

Identifiability Analysis: Overview

	[G. & Schnass 2010]	*[Geng & al 2011]*	*[Jenatton, Bach & G.]*
signal model			
overcomplete (d<K)	no	**yes**	**yes**
outliers	**yes**	no	**yes**
noise	no		**yes**
cost function	$\min_{\mathbf{D},Z} \|Z\|_1 \text{ s.t.} \mathbf{D}Z = \mathbf{X}$		$\min F_{\mathbf{X}}(\mathbf{D})$

$$\phi(z) = \lambda \|z\|_1$$

Identifiability Analysis: Overview

	[G. & Schnass 2010]	*[Geng & al 2011]*	*[Jenatton, Bach & G.]*
signal model			
overcomplete (d<K)	no	**yes**	**yes**
outliers	**yes**	no	**yes**
noise	no		**yes**
cost function	$\min_{\mathbf{D},Z} \|Z\|_1 \text{ s.t.} \mathbf{D}Z = \mathbf{X}$		$\min F_{\mathbf{X}}(\mathbf{D})$

$$\phi(z) = \lambda \|z\|_1$$

See also: *[Spielman&al 2012, Agarwal & al 2013/2014, Arora & al 2013/2014, Schnass 2013, Schnass 2014]*

Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...	Source localization, neural coding ...
✓ No «ground truth dictionary»	✓ Ground truth $\mathbf{x} = \mathbf{D}_0 z + \varepsilon$
✓ Goal = performance generalization	✓ Goal = dictionary estimation
$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$	$\lVert\hat{\mathbf{D}}_N - \mathbf{D}_0\rVert_F$
• ***«How many training samples ?»***	• ***What recovery conditions ?***
• **Excess risk analysis** (~Machine Learning)	• ***Identifiability analysis*** (~Signal Processing)
✦ *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*	✦ *[Independent Component Analysis, e.g. book Comon & Jutten 2011]*

Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

✓ Compression, denoising, calibration, inverse problems ...

✓ No «ground truth dictionary»

✓ Goal = performance generalization

$$\mathbb{E}F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E}F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***
- **Excess risk analysis** ✓
  (~Machine Learning)
  - *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*

Source localization, neural coding ...

✓ Ground truth $\mathbf{x} = \mathbf{D}_0 z + \varepsilon$

✓ Goal = dictionary estimation

$$\|\hat{\mathbf{D}}_N - \mathbf{D}_0\|_F$$

- ***What recovery conditions ?***
- ***Identifiability analysis***
  (~Signal Processing)
  - *[Independent Component Analysis, e.g. book Comon & Jutten 2011]*

Theoretical Guarantees ?

- **Given N training samples in *X*:** $\hat{\mathbf{D}}_N \in \arg\min_{\mathbf{D}} F_{\mathbf{X}}(\mathbf{D})$

Compression, denoising, calibration, inverse problems ...

- ✓ No «ground truth dictionary»
- ✓ Goal = performance generalization

$$\mathbb{E} F_{\mathbf{X}}(\hat{\mathbf{D}}_N) \leq \min_{\mathbf{D}} \mathbb{E} F_{\mathbf{X}}(\mathbf{D}) + \eta_N$$

- ***«How many training samples ?»***
- **Excess risk analysis** ✓ (~Machine Learning)
  - *[Maurer and Pontil, 2010; Vainsencher & al., 2010; Mehta and Gray, 2012; G. & al 2013]*

Source localization, neural coding ...

- ✓ Ground truth $\mathbf{x} = \mathbf{D}_0 z + \varepsilon$
- ✓ Goal = dictionary estimation

$$\|\hat{\mathbf{D}}_N - \mathbf{D}_0\|_F$$

- ***What recovery conditions ?***
- ***Identifiability analysis*** (~Signal Processing)
  - *[Independent Component Analysis, e.g. book Comon & Jutten 2011]*

«Ground Truth» = Sparse Signal Model

- **Random support** $J \subset [1, K], \ \sharp J = s$
- **Bounded** coefficient vector + bounded from below

$$\mathbb{P}(\|z_J\|_2 > M_z) = 0 \qquad \mathbb{P}(\min_{j \in J} |z_i| < \underline{z}) = 0$$

- **Bounded white noise**

$$\mathbb{P}(\|\varepsilon\|_2 > M_\epsilon) = 0$$

  - ✓ (+ second moment assumptions)

$$\mathbf{x} = \sum_{i \in J} z_i \mathbf{d}_i + \varepsilon = \mathbf{D}_J z_J + \varepsilon$$

«Ground Truth» = Sparse Signal Model

- **Random support** $J \subset [1, K], \ \sharp J = s$
- **Bounded** coefficient vector + bounded from below

$$\mathbb{P}(\|z_J\|_2 > M_z) = 0 \qquad \mathbb{P}(\min_{j \in J} |z_i| < \underline{z}) = 0$$

- **Bounded white noise**

$$\mathbb{P}(\|\varepsilon\|_2 > M_\epsilon) = 0$$

  ✓ (+ second moment assumptions)

$$\mathbf{x} = \sum_{i \in J} z_i \mathbf{d}_i + \varepsilon = \mathbf{D}_J z_J + \varepsilon$$

NB: z *not* required to have i.i.d. entries

Theorem: Robust Local Identifiability

- **Assume** *[Jenatton, Bach & G. 2012]*
  - dictionary with small *coherence* $\mu(\mathbf{D}_0) = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| \in [0,1]$
  - *s*-sparse coefficient model (no outlier, no noise) $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$

Theorem: Robust Local Identifiability

- **Assume** *[Jenatton, Bach & G. 2012]*
  - dictionary with small *coherence* $\mu(\mathbf{D}_0) = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| \in [0,1]$
  - *s*-sparse coefficient model (no outlier, no noise) $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$
- **Then**: consider $F_{\mathbf{X}}(\mathbf{D}) = \min_Z \frac{1}{2}\|\mathbf{X} - \mathbf{D}Z\|_F^2 + \lambda\|Z\|_{1,1}$

✓ for any small enough λ, with high probability on $\mathbf{X}$, there is a **local minimum** $\hat{\mathbf{D}}$ of $F_{\mathbf{X}}(\mathbf{D})$ such that

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu |||\mathbf{D}_0|||_2)$$

Theorem: Robust Local Identifiability

- **Assume** *[Jenatton, Bach & G. 2012]*
  - ✦ dictionary with small *coherence* $\mu(\mathbf{D}_0) = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| \in [0, 1]$
  - ✦ *s*-sparse coefficient model (no outlier, no noise) $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$
- **Then**: consider $F_{\mathbf{X}}(\mathbf{D}) = \min_{Z} \frac{1}{2}\|\mathbf{X} - \mathbf{D}Z\|_F^2 + \lambda \|Z\|_{1,1}$

✓ for any small enough λ, with high probability on $\mathbf{X}$, there is a **local minimum** $\hat{\mathbf{D}}$ of $F_{\mathbf{X}}(\mathbf{D})$ such that

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu |||\mathbf{D}_0|||_2)$$

- + stability to **noise**
- + **finite sample** results
- + robustness to **outliers**

Example 1: Orthonormal Dictionary

- **Coherence** $\mu(\mathbf{D}_0) = 0$
- **No sparsity constraint** $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$

Example 1: Orthonormal Dictionary

- **Coherence** $\mu(\mathbf{D}_0) = 0$
- **No sparsity constraint** $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$
- **Asymptotic guarantee:**
  - ✓ for $M_\epsilon < \lambda \leq \underline{z}/4$, with high probability on $\mathbf{X}$, there is a local minimum such that

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu |||\mathbf{D}_0|||_2)$$

Example 1: Orthonormal Dictionary

- **Coherence** $\mu(\mathbf{D}_0) = 0$
- **No sparsity constraint** $s \lesssim \frac{1}{\mu(\mathbf{D}_0)\|\|\mathbf{D}_0\|\|_2}$
- **Asymptotic guarantee:**

✓ for $M_\epsilon < \lambda \leq \underline{z}/4$, with high probability on $\mathbf{X}$, there is a local minimum such that

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu \|\|\mathbf{D}_0\|\|_2) \quad \mathbf{=0}$$

✓ **exact recovery**

Example 1: Orthonormal Dictionary

- **Coherence** $\mu(\mathbf{D}_0) = 0$
- **No sparsity constraint** $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$
- **Asymptotic guarantee:**
  - ✓ for $M_\epsilon < \lambda \leq \underline{z}/4$, with high probability on $\mathbf{X}$, there is a local minimum such that

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu |||\mathbf{D}_0|||_2) \quad \mathbf{=0}$$

  - ✓ **exact recovery**
- **Noiseless: finite sample** results with $N = \Omega(d^4)$

$$\mathbf{D} \in \mathbb{R}^{d \times d}$$

Example 1: Orthonormal Dictionary

- **Coherence** $\mu(\mathbf{D}_0) = 0$
- **No sparsity constraint** $s \lesssim \frac{1}{\mu(\mathbf{D}_0)|||\mathbf{D}_0|||_2}$
- **Asymptotic guarantee:**
  - ✓ for $M_\epsilon < \lambda \leq \underline{z}/4$, with high probability on $\mathbf{X}$, there is a local minimum such that

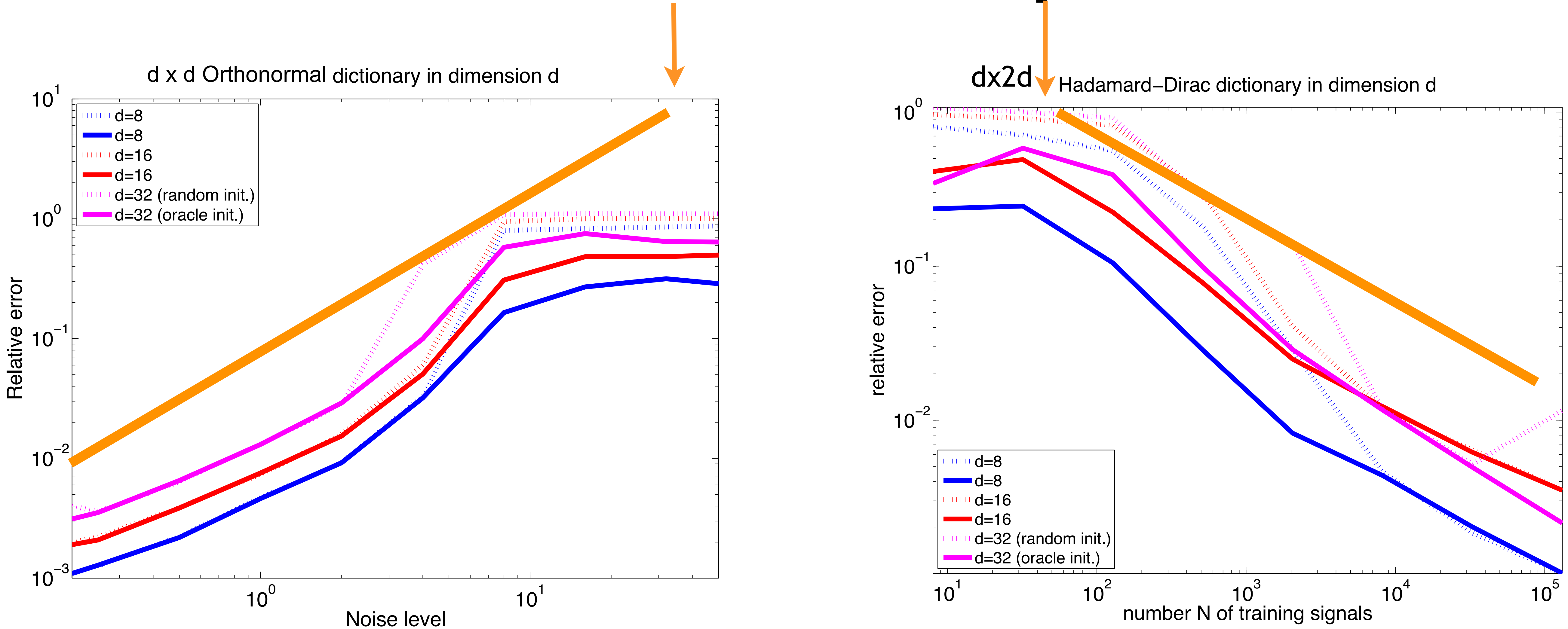
$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F \leq O(\lambda s \mu |||\mathbf{D}_0|||_2) \quad \mathbf{=0}$$

  - ✓ **exact recovery**
- **Noiseless: finite sample** results with $N = \Omega(d^4)$
- +Robustness to **outliers** $\mathbf{D} \in \mathbb{R}^{d \times d}$

Example 2: Guarantees vs Observations

- **Robustness to noise**
- **Sample complexity**

Predicted slope

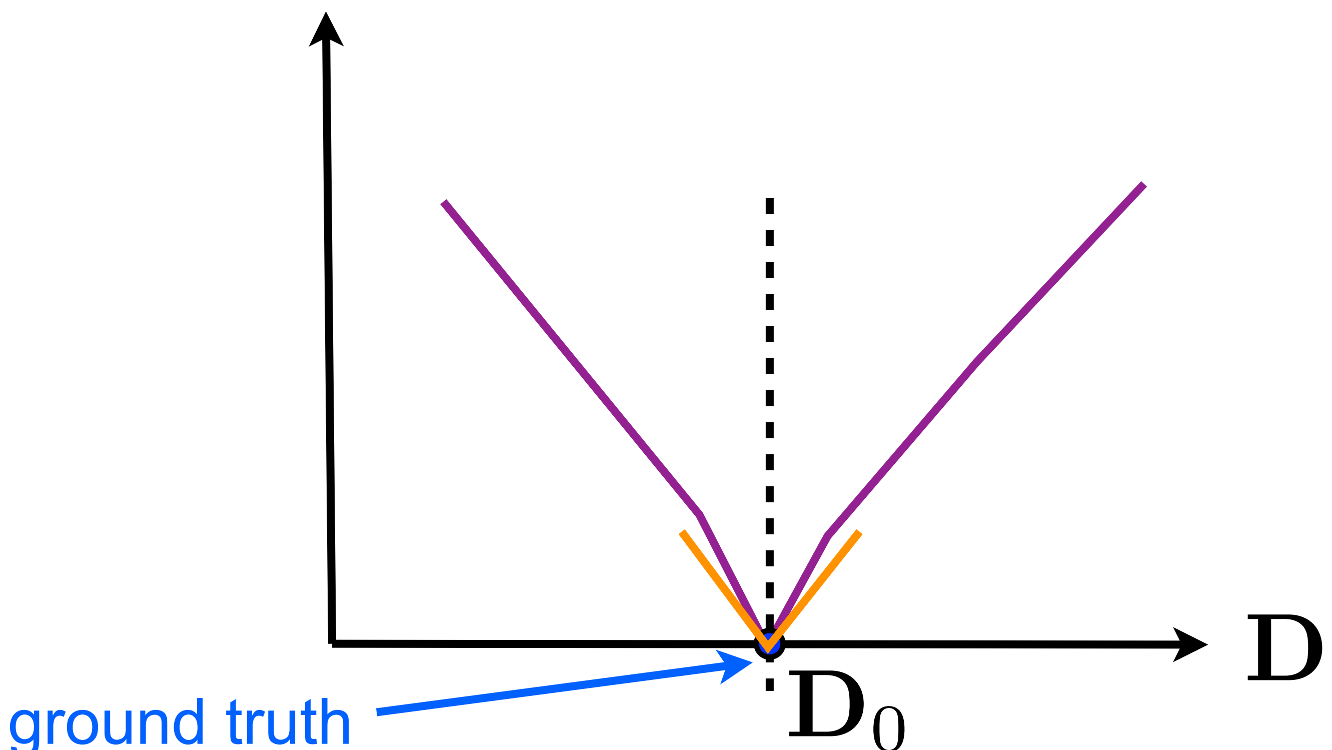


Flavor of the proof

Characterizing Local Minima (1)

- **Noiseless setting**
  - ✓ Minimum ***exactly*** at ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

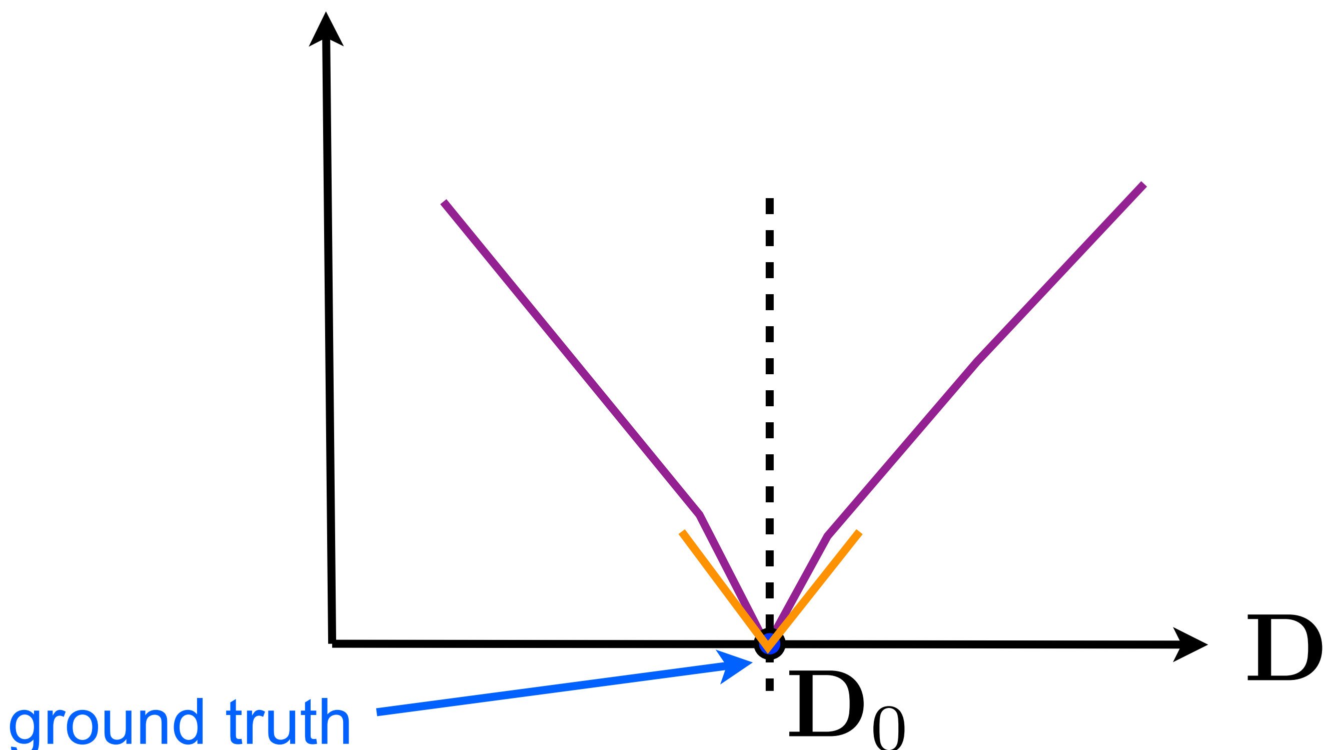


  - ✓ one-sided directional derivatives

Characterizing Local Minima (1)

- **Noiseless setting**
  - ✓ Minimum ***exactly*** at ground truth

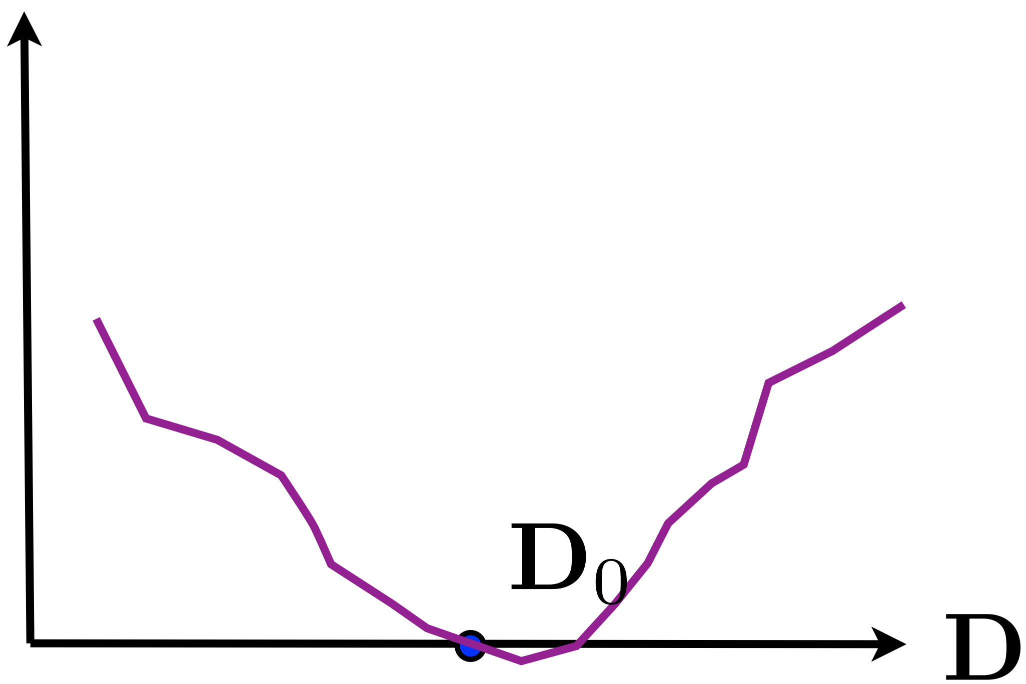
$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



  - ✓ one-sided directional derivatives

- **Noisy setting**
  - ✓ Minimum ***close to*** ground truth

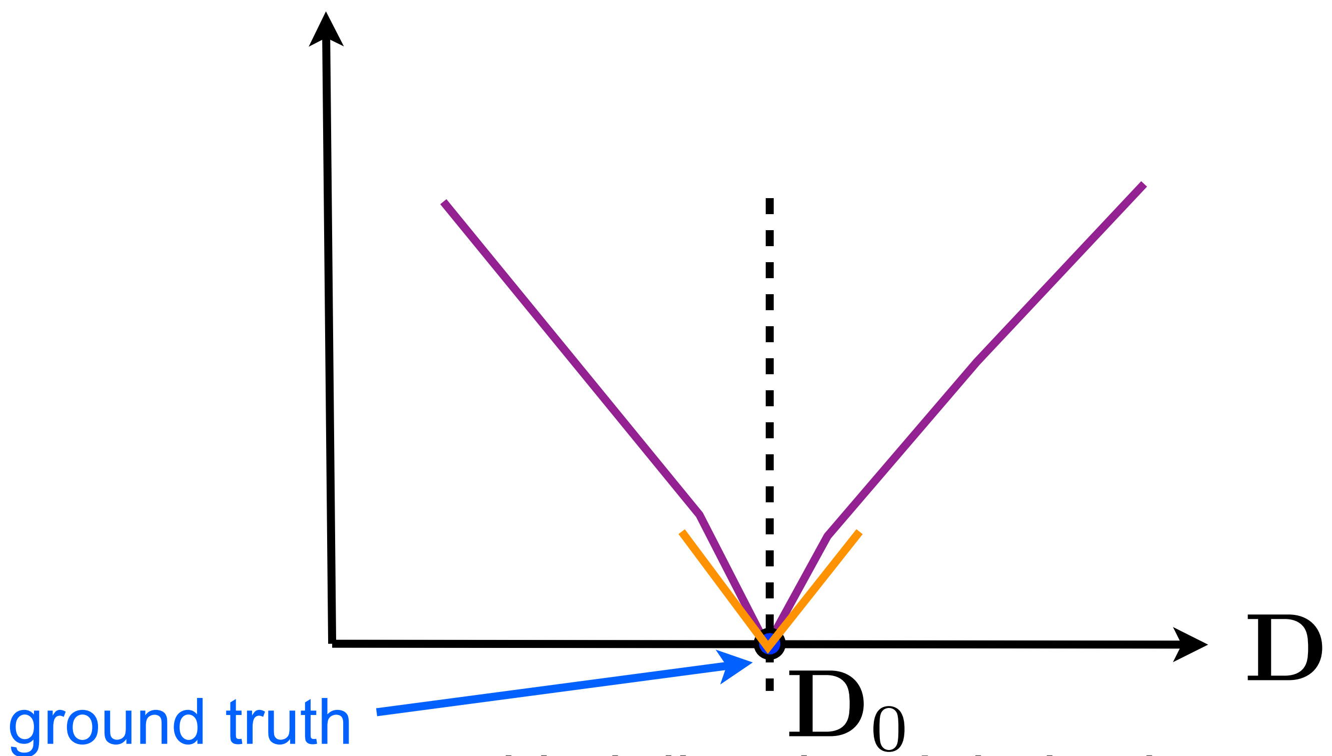
$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



Characterizing Local Minima (1)

- **Noiseless setting**
  - ✓ Minimum ***exactly*** at ground truth

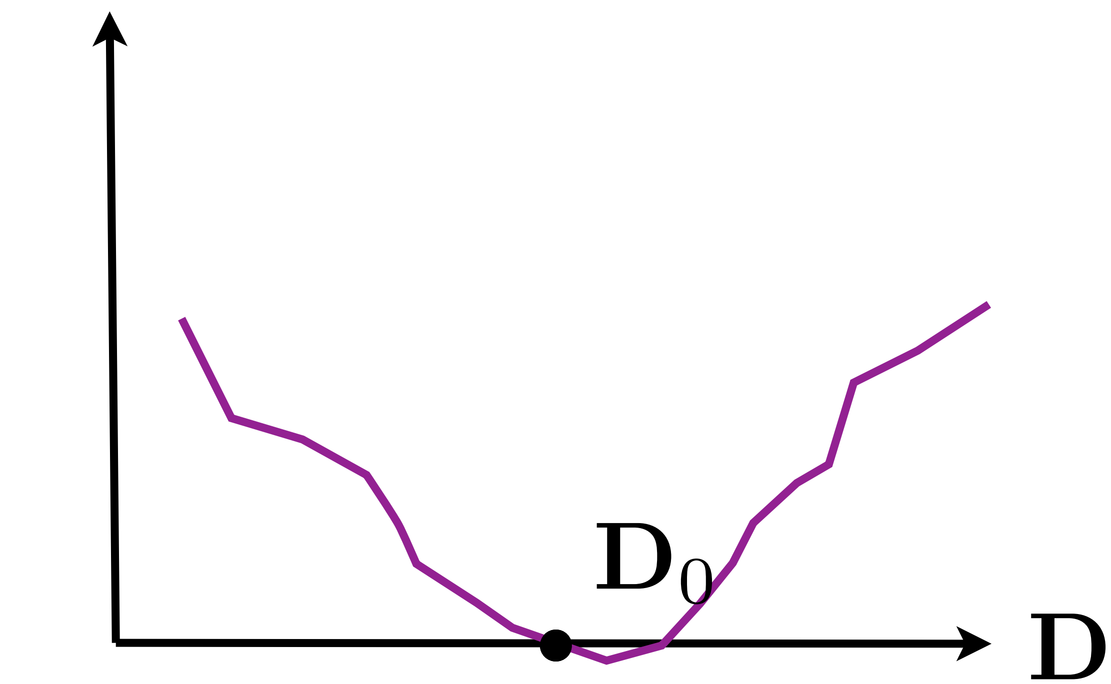
$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



  - ✓ one-sided directional derivatives

- **Noisy setting**
  - ✓ Minimum ***close to*** ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$

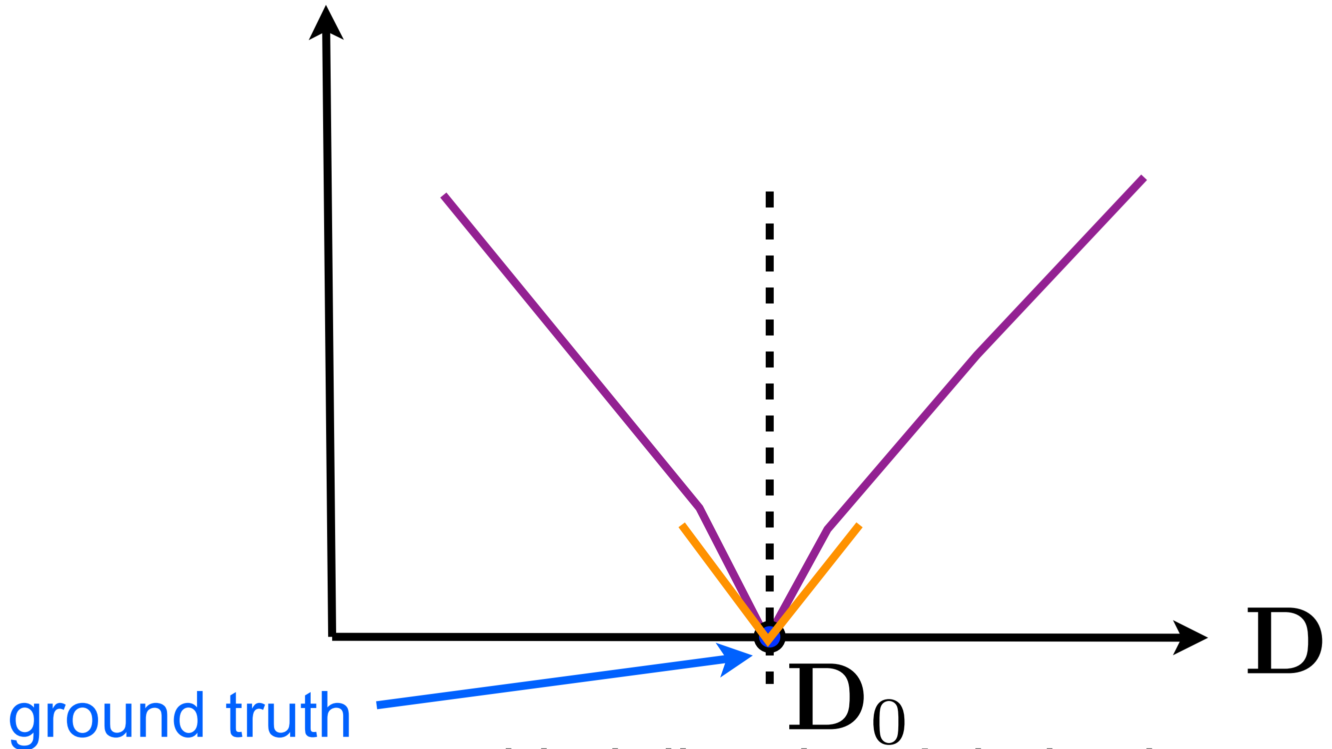


  - ✓ Zero at ground truth

Characterizing Local Minima (1)

- **Noiseless setting**
  - ✓ Minimum ***exactly*** at ground truth

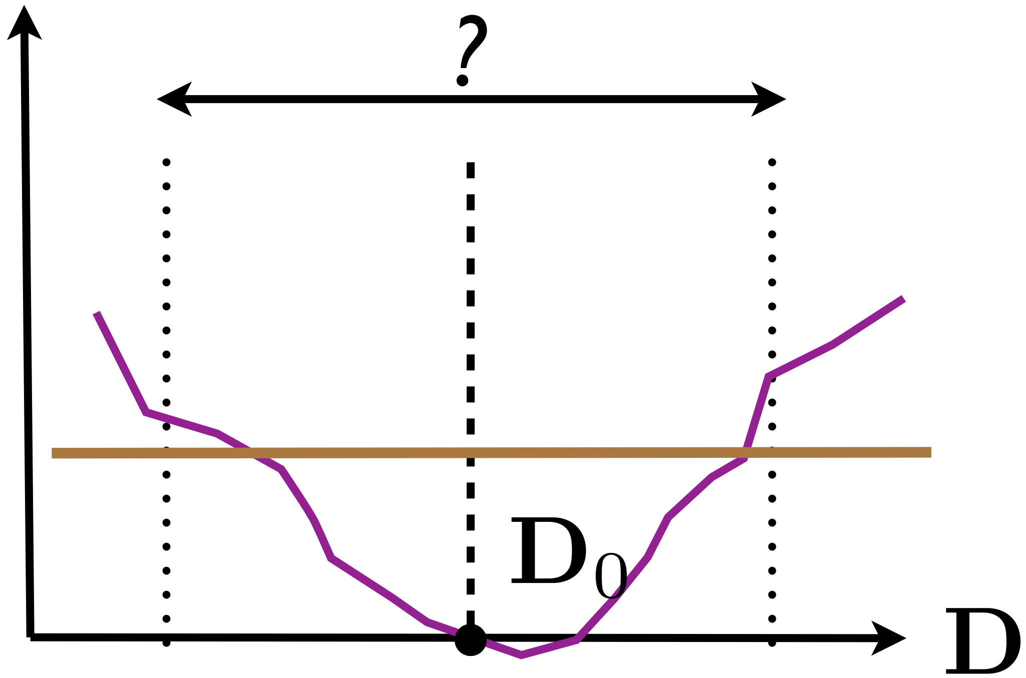
$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



  - ✓ one-sided directional derivatives

- **Noisy setting**
  - ✓ Minimum ***close to*** ground truth

$$F_{\mathbf{X}}(\mathbf{D}) - F_{\mathbf{X}}(\mathbf{D}_0)$$



  - ✓ Zero at ground truth
  - ✓ Lower bound at radius *r*

Leveraging Sparse Recovery Results

- **Problem 1**: implicit definition

$$f_{\mathbf{x}_n}(\mathbf{D}) = \boxed{\min_{z_n}} \frac{1}{2}\|\mathbf{x}_n - \mathbf{D}z_n\|_2^2 + \lambda\|z_n\|_1$$

- **Approach**: *explicit* expression & sparse recovery ***stable to dictionary perturbations and noise***
  - adaptation from [Fuchs, 2005; Zhao and Yu, 2006; Wainwright, 2009]

$$f_{\mathbf{x}}(\mathbf{D}) = \phi_{\mathbf{x}}(\mathbf{D}|\mathrm{sign}(z_0)) \qquad \mathbf{x} = \mathbf{D}_0\mathbf{z}_0 + \varepsilon$$

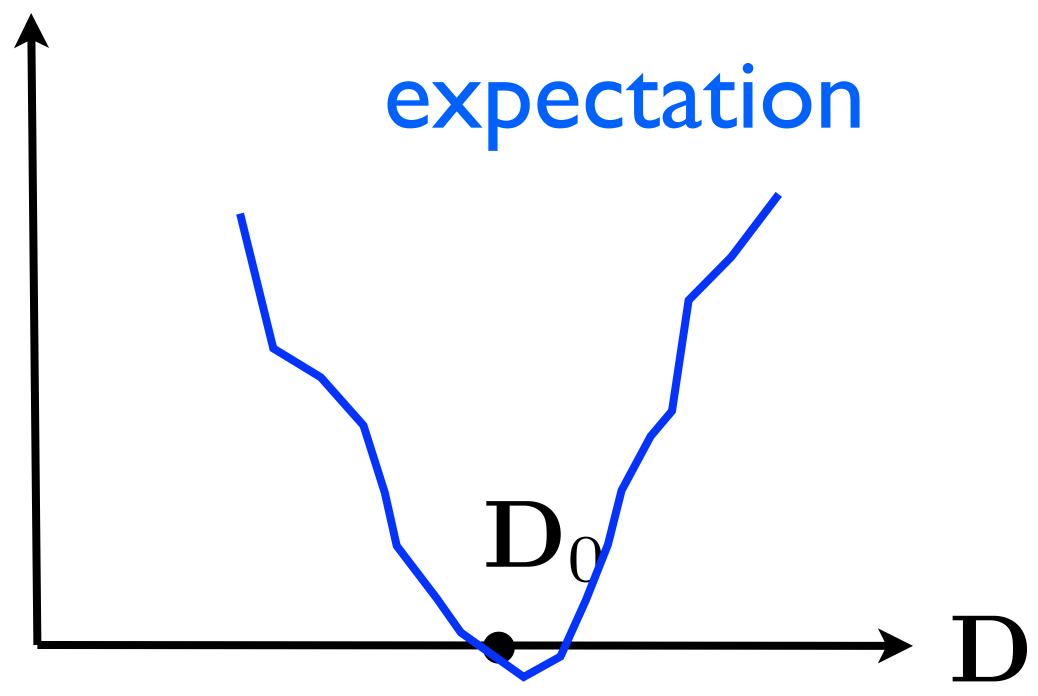
  - uses «guess» of minimizer

$$\hat{z} = \mathbf{D}_J^{+}\mathbf{x} - \lambda(\mathbf{D}_J^{\top}\mathbf{D}_J)^{-1}\mathrm{sign}(z_0)$$

Step 1: Asymptotic Regime

- **Goal: control *expectation***

$$\mathbb{E} f_{\mathbf{x}}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D}_0)$$



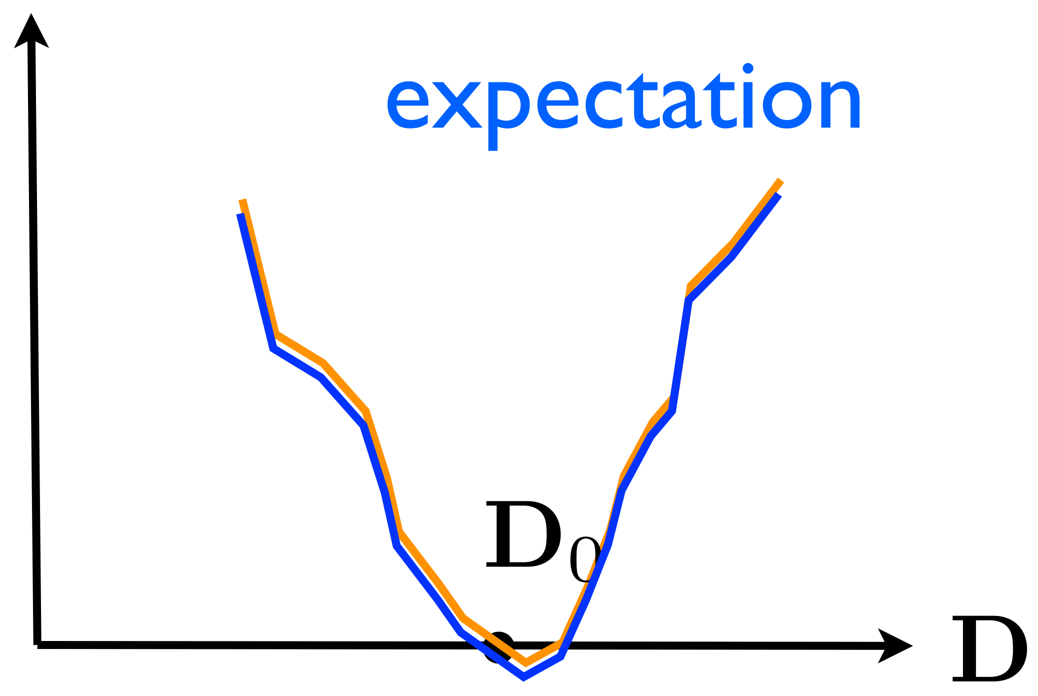
Step 1: Asymptotic Regime

- **Goal: control *expectation***

$$\mathbb{E} f_{\mathbf{x}}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D}_0)$$

- **Using incoherence**
  - ✓ **more explicit form**

$$\mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}|\mathrm{sign}(z_0)) - \mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}_0|\mathrm{sign}(z_0))$$



Step 1: Asymptotic Regime

- **Goal: control *expectation***

  $$\mathbb{E} f_{\mathbf{x}}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D}_0)$$

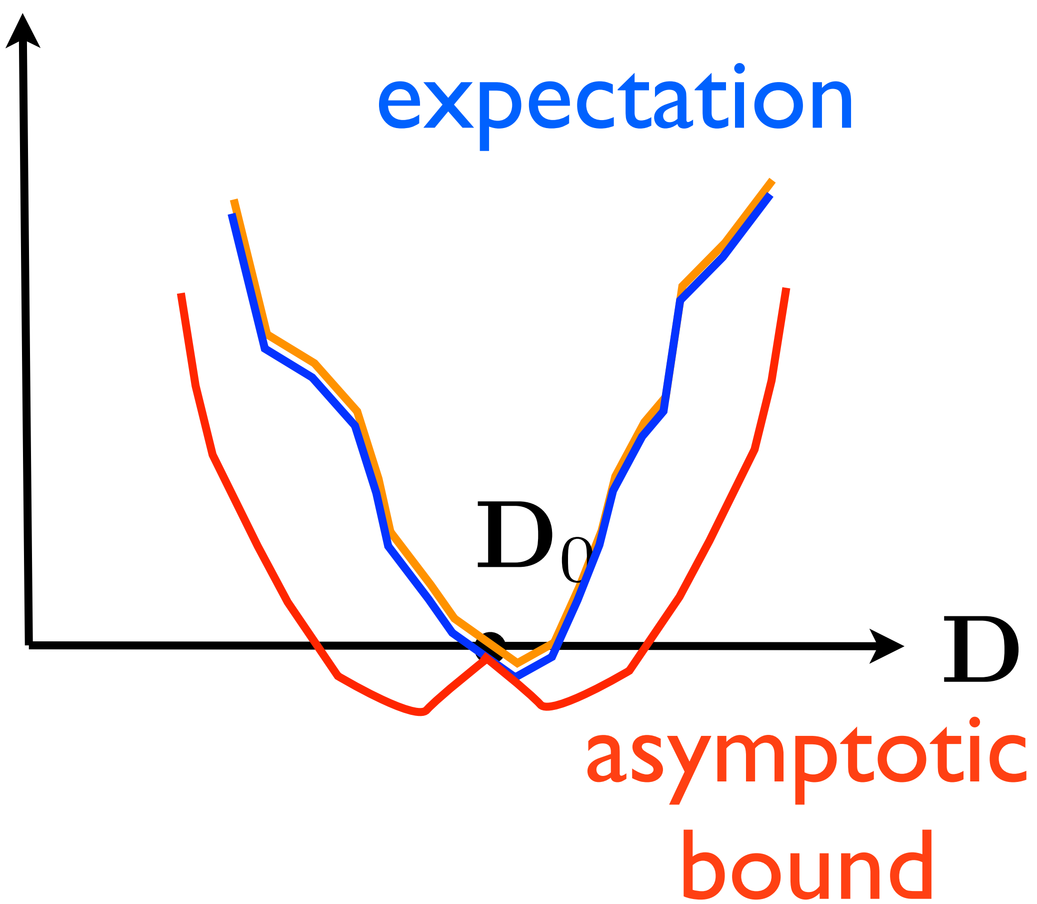
- **Using incoherence**
  - ✓ more explicit form

  $$\mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}|\mathrm{sign}(z_0)) - \mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}_0|\mathrm{sign}(z_0))$$

  - ✓ lower bound

  $$a\|\mathbf{D}-\mathbf{D}_0\|_F(\|\mathbf{D}-\mathbf{D}_0\|_F - r_0)$$

    - ✦ where $r_0 = O(\lambda s \mu |\!|\!| \mathbf{D}_0 |\!|\!|_2)$



Step 1: Asymptotic Regime

- **Goal: control *expectation***

$$\mathbb{E} f_{\mathbf{x}}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D}_0)$$

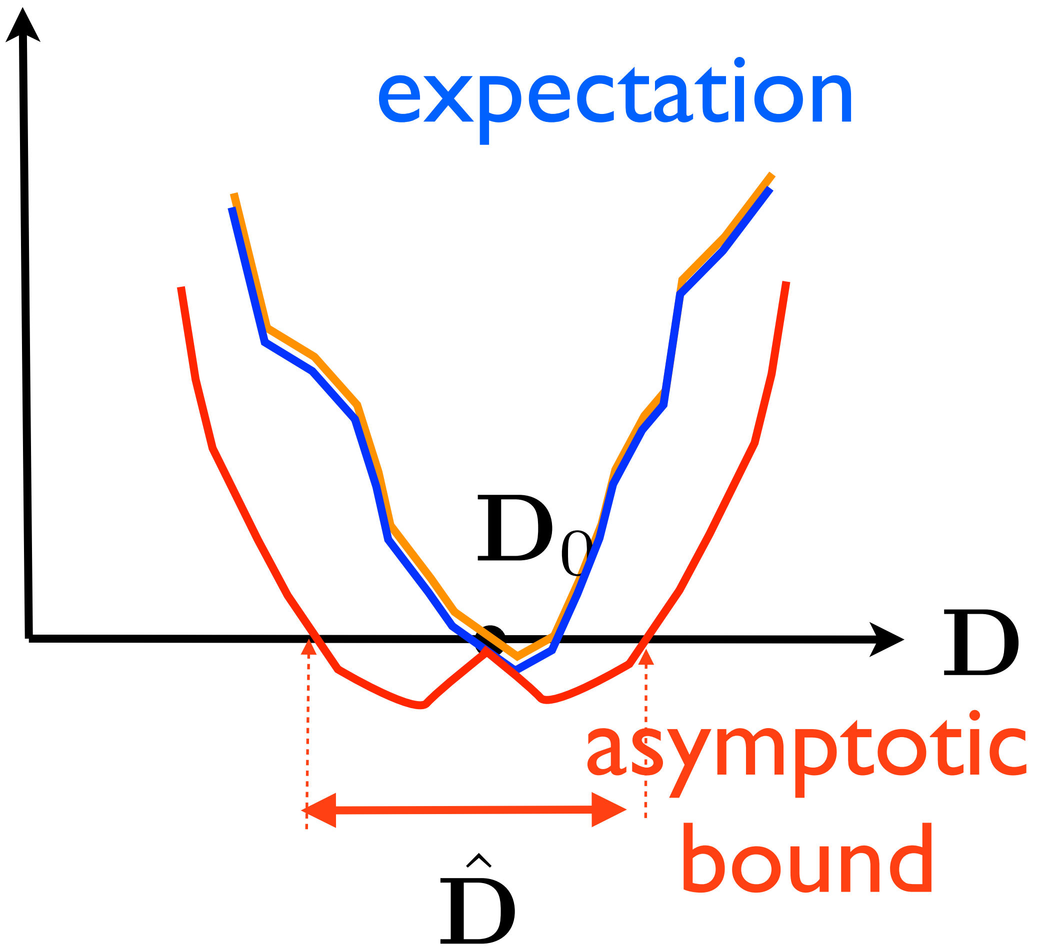
- **Using incoherence**
  - ✓ **more explicit form**

$$\mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}|\mathrm{sign}(z_0)) - \mathbb{E}\phi_{\mathbf{x}}(\mathbf{D}_0|\mathrm{sign}(z_0))$$

  - ✓ **lower bound**

$$a\|\mathbf{D} - \mathbf{D}_0\|_F(\|\mathbf{D} - \mathbf{D}_0\|_F - r_0)$$

    - ✦ **where** $r_0 = O(\lambda s \mu \|\!|\mathbf{D}_0\|\!|_2)$



- **Asymptotically:**
  - ✓ there is a local minimum **within radius *r0***

Step 2: Finite Sample Analysis

- **Sample complexity result**

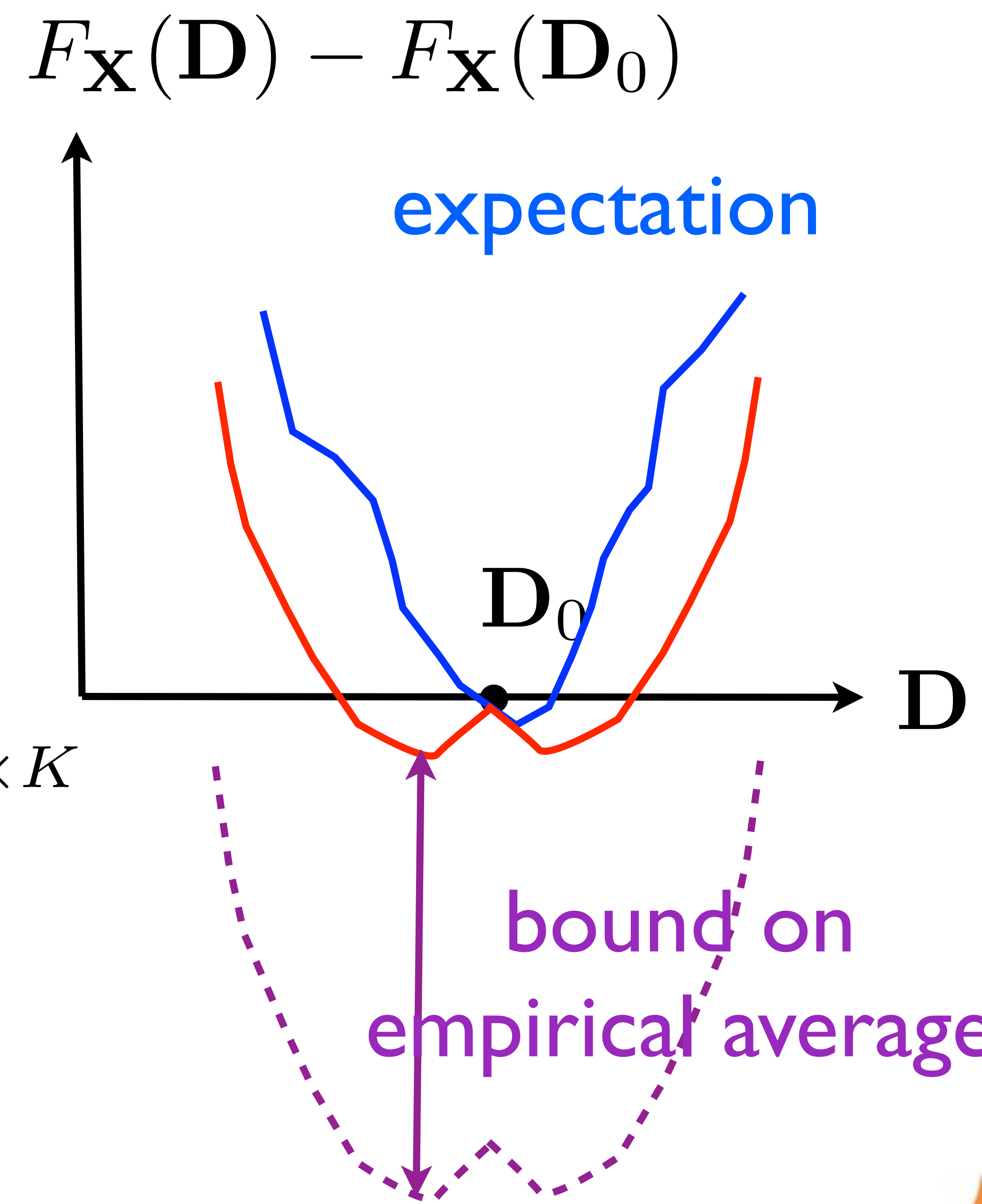
$$\sup_{\mathbf{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E}f_{\mathbf{x}}(\mathbf{D})| \leq \eta_N$$

- ***Naive* version: whp**

$$= O\left(\sqrt{\frac{\log N}{N}}\right)$$

✓ local min with $\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F < r$ if

$$N = \Omega(dK^3 r^{-2}) \qquad \mathbf{D} \in \mathbb{R}^{d \times K}$$



Step 2: Finite Sample Analysis

- **Sample complexity result**

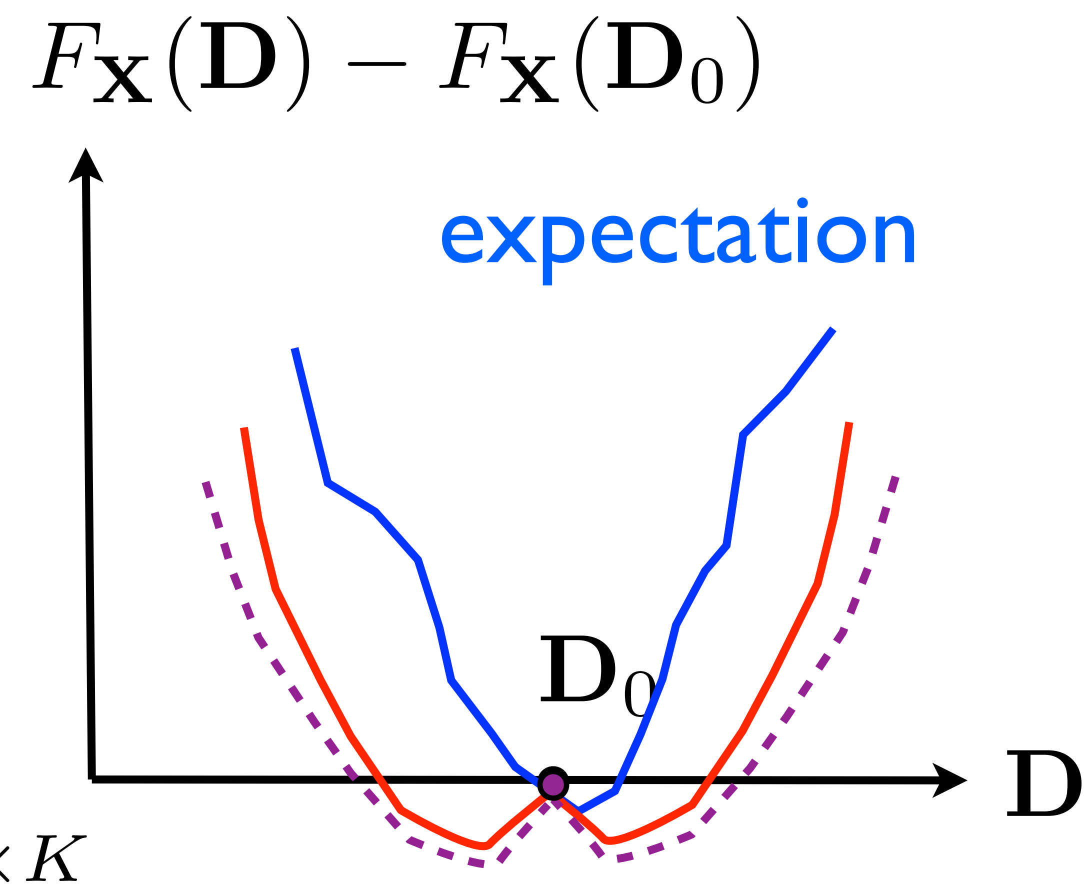
$$\sup_{\mathbf{D}} |F_{\mathbf{X}}(\mathbf{D}) - \mathbb{E} f_{\mathbf{x}}(\mathbf{D})| \leq \eta_N$$

- ***Naive* version: whp**

$$= O\left(\sqrt{\frac{\log N}{N}}\right)$$

✓ **local min with** $\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F < r$ **if**

$$N = \Omega(dK^3 r^{-2}) \qquad \mathbf{D} \in \mathbb{R}^{d \times K}$$



- ***Refined* version:** [Rademacher averages & Slepian's lemma]

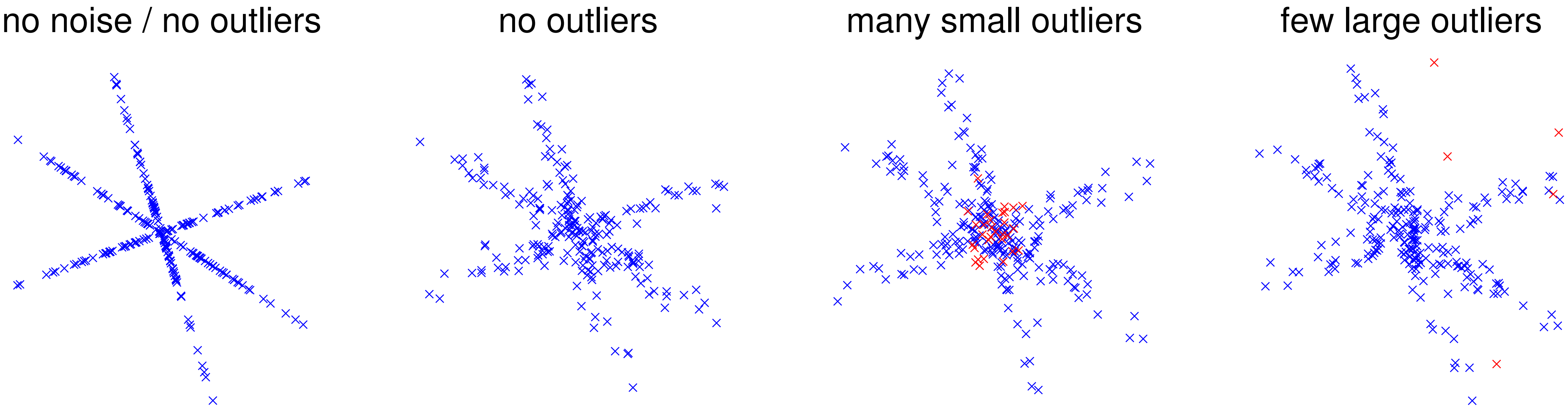
$$= O(r_0^2/\sqrt{N})$$

bound on empirical average

✓ **local minimum if** $N = \Omega(dK^3)$

Outliers ?

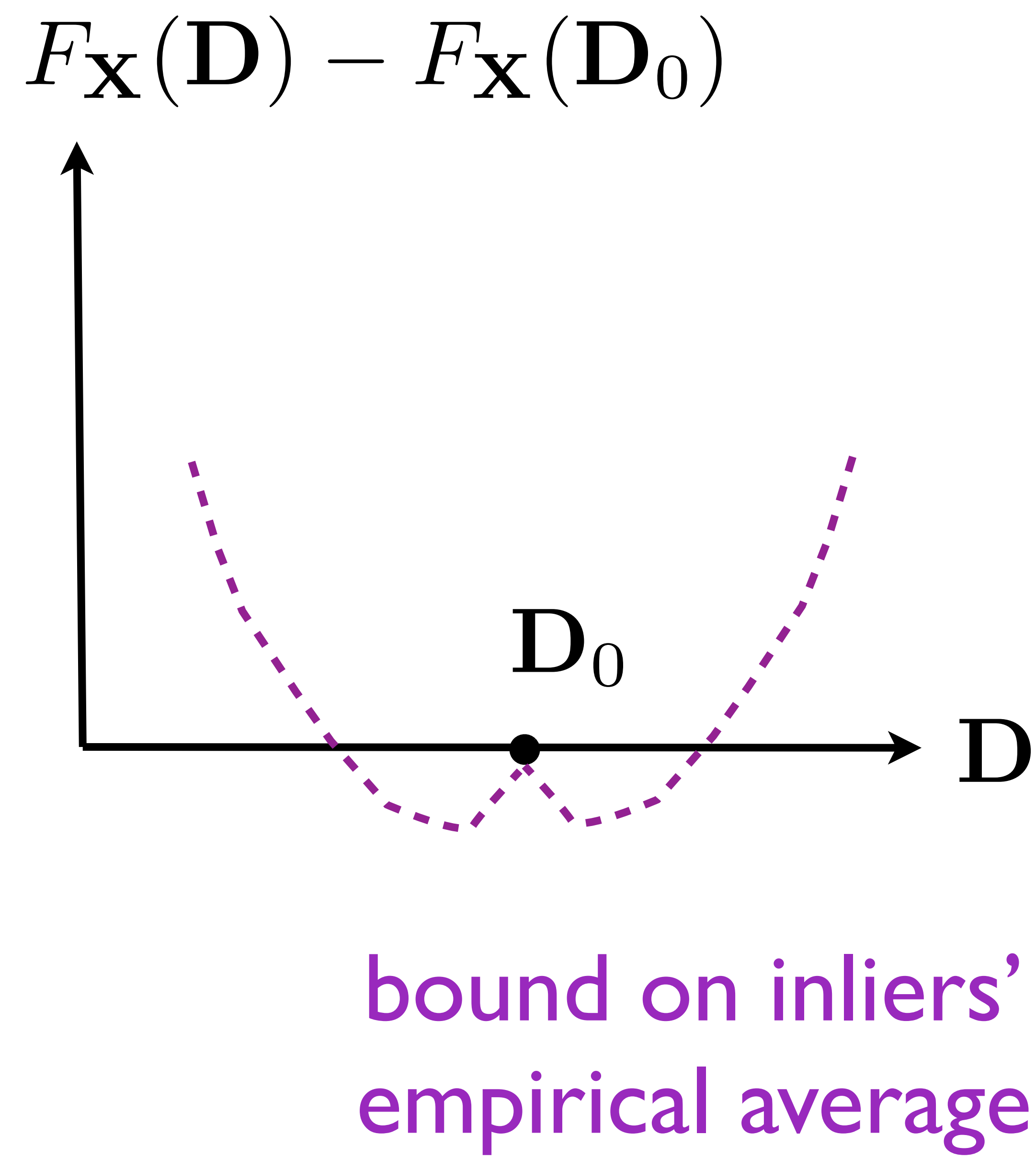
- **Inliers**: sparse signal model $\mathbf{x} = \sum_{i \in J} z_i \mathbf{d}_i + \varepsilon = \mathbf{D}_J z_J + \varepsilon$
- **Outliers**: anything else, ***even adversarial***



- Wlog, decomposition of training set $\mathbf{X} = [\mathbf{X}_{\mathrm{in}}, \mathbf{X}_{\mathrm{out}}]$

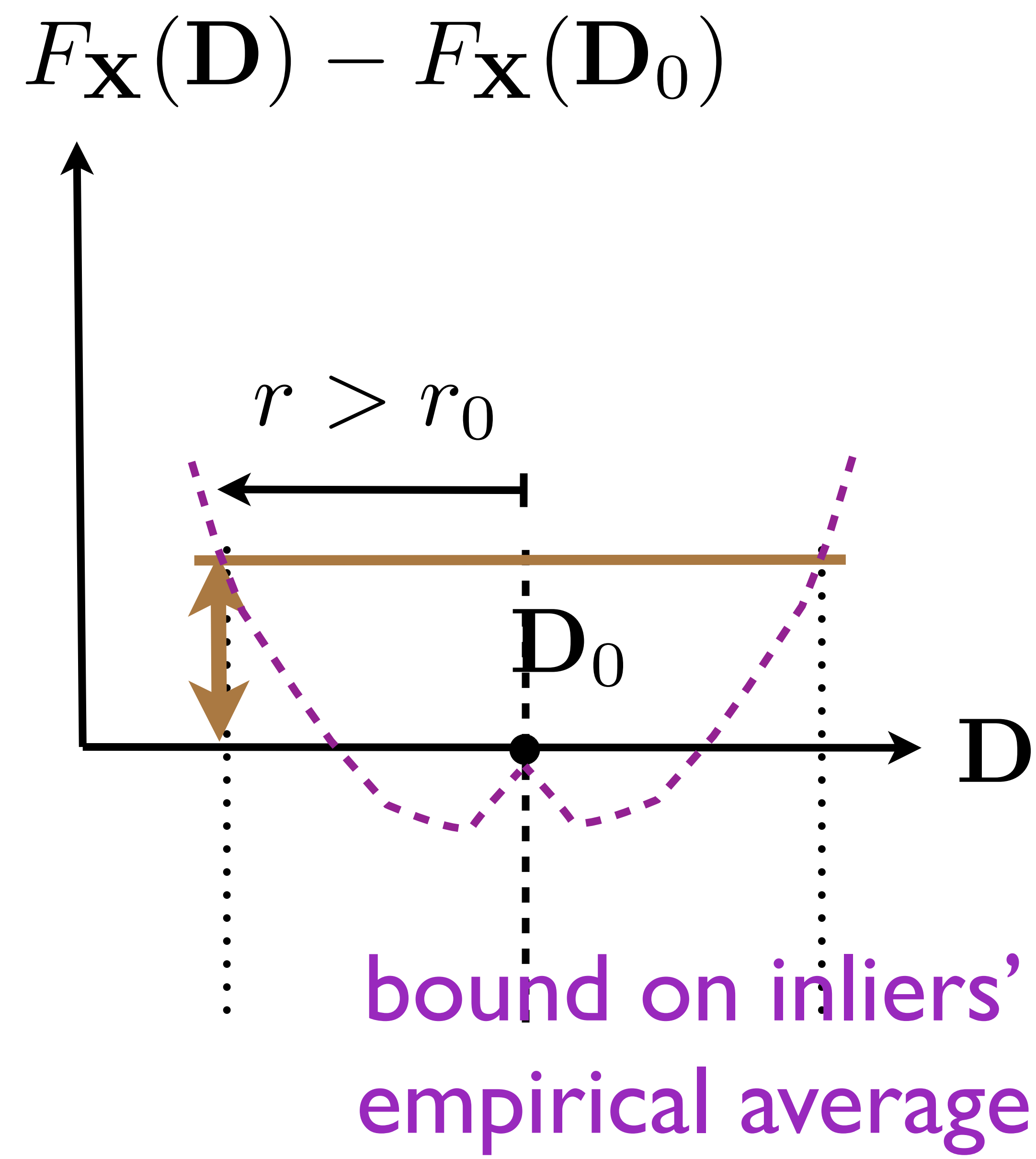
Step 3: Robustness to Outliers

- ***Inliers* sample complexity**



Step 3: Robustness to Outliers

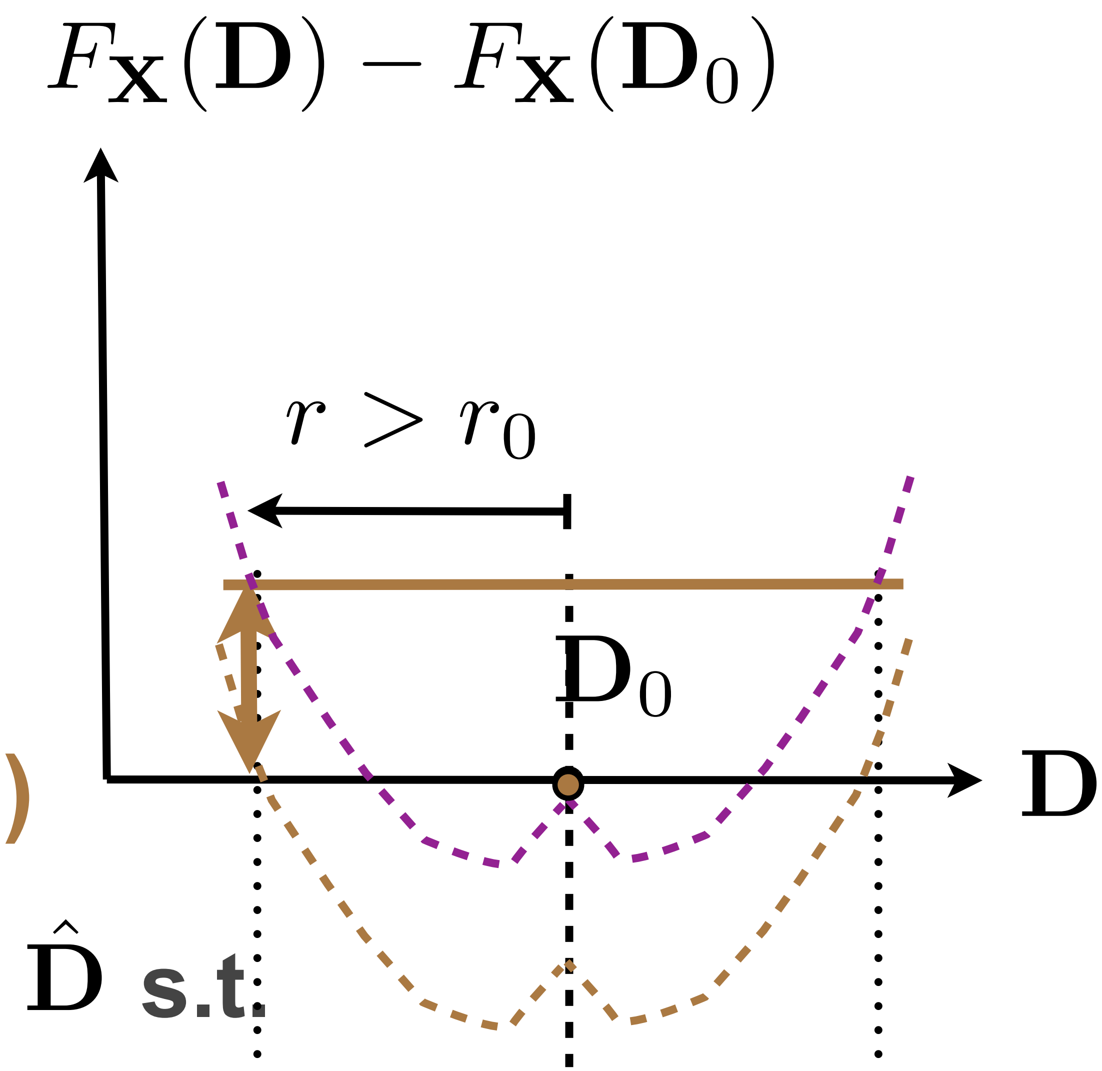
- ***Inliers* sample complexity**
- **«Room left» for outliers**



Step 3: Robustness to Outliers

- *Inliers* **sample complexity**
- **«Room left» for outliers**
- **If** $\sum_{n \in \text{outlier}} \|\mathbf{x}_n\|_2 \leq C(r) N_{\text{inlier}}$

(admissible «energy» of outliers)

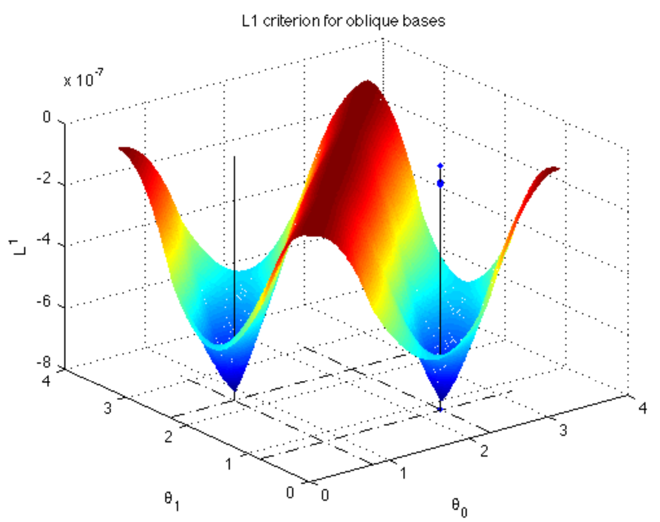


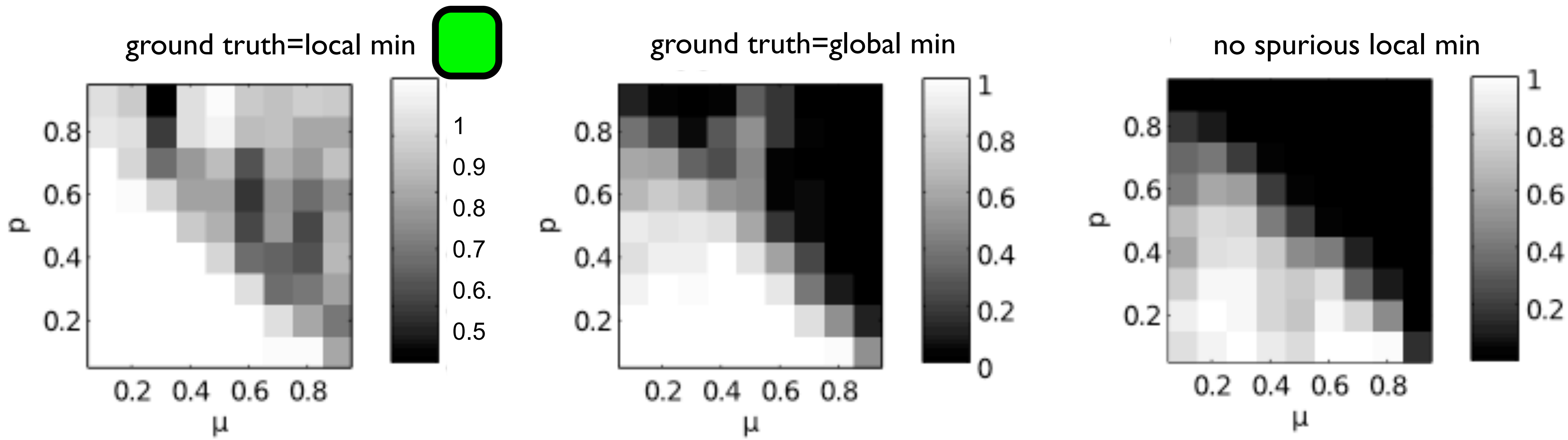
✓ **then whp there is a local min $\hat{\mathbf{D}}$ s.t.**

$$\|\hat{\mathbf{D}} - \mathbf{D}_0\|_F < r$$

From Local to Global Guarantees ?

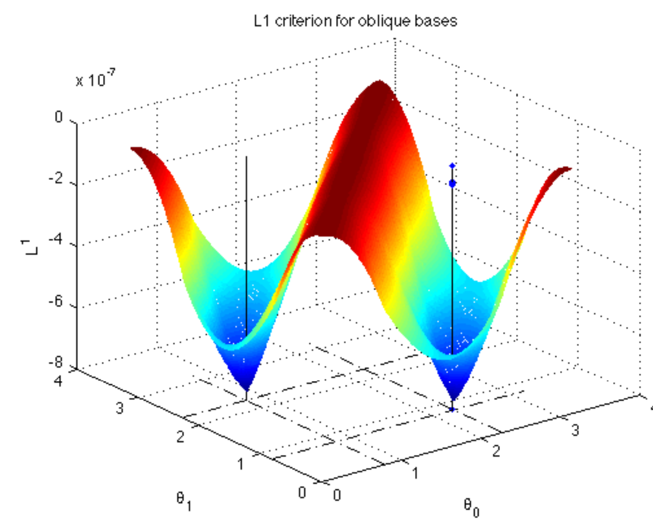
$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}\in\mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$

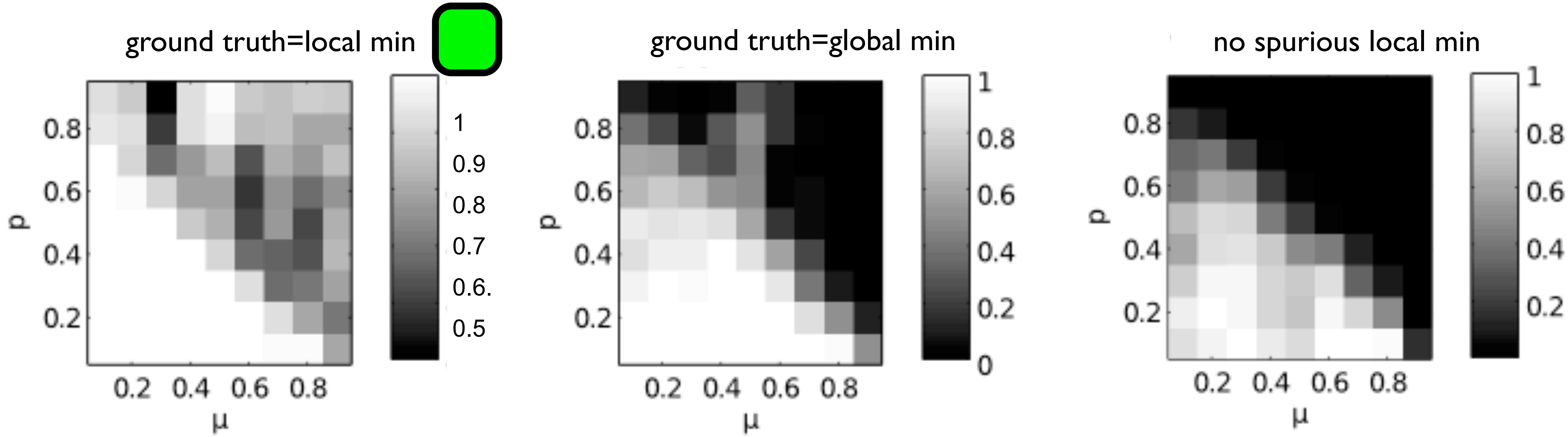




From Local to Global Guarantees ?

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}\in\mathcal{D}} F_{\mathbf{X}}(\mathbf{D})$$





See also: *[Spielman&al 2012, Agarwal & al 2013/2014, Arora & al 2013/2014]*

Recent results $\mathbf{D} \in \mathbb{R}^{m \times p}$

POLYNOMIAL ALGORITHMS

Reference	Overcomplete	Noise	Outliers	Global min / algorithm	Polynomial algorithm	Exact (no noise, no outlier, n finite)	Sample complexity	Admissible sparsity for exact recovery	Coefficient model (main characteristics)
Georgiev et al. [2005] *Combinatorial approach*	✓	✗	✗	✓	✗	✓	$m\binom{p}{m-1}$	$k = m-1$, $\underline{\delta}_m(\mathbf{D}^o) < 1$	Combinatorial
Aharon et al. [2006] *Combinatorial approach*	✓	✗	✗	✓	✗	✓	$(k+1)\binom{p}{k}$	$\underline{\delta}_{2k}(\mathbf{D}^o) < 1$	Combinatorial
Gribonval and Schnass [2010] ℓ^1 *criterion*	✗	✗	✗	✗	✗	✓	$\frac{m^2 \log m}{k}$	$\frac{k}{m} < 1 - \\|\mathbf{D}^\top \mathbf{D} - \mathbf{I}\\|_{2,\infty}$	Bernoulli(k/p) -Gaussian
Geng et al. [2011] ℓ^1 *criterion*	✓	✗	✗	✗	✗	✓	kp^3	$O(1/\mu_1(\mathbf{D}^o))$	k-sparse -Gaussian
Spielman et al. [2012] ℓ^0 *criterion* / *ER-SpUD (randomized)*	✗	✗	✗	✓ / ✓	✗ / ✓	✓ / ✓	$m \log m$ / $m^2 \log^2 m$	$O(m)$ / $O(\sqrt{m})$	Bernoulli(k/p) -Gaussian or -Rademacher
Schnass [2013] *K-SVD criterion* (NB: tight frames only)	✓	✓	✗	✗	✗	$\\|\hat{\mathbf{D}} - \mathbf{D}^o\\|_{2,\infty} = O(pn^{-1/4})$	$\frac{mp^3}{r^2}$	$O(1/\mu_1(\mathbf{D}^o))$	"Symmetric decaying": $\boldsymbol{\alpha}_j = \epsilon_j \mathbf{a}_{\sigma(j)}$
Arora et al. [2013] *Graphs & clustering*	✓	✗	✗	✓	✓	$\\|\hat{\mathbf{D}} - \mathbf{D}^o\\|_{2,\infty} \leqslant r$	$\max\left(\frac{p^2 \log p}{k^2}, \frac{p \log p}{r^2}\right)$	$O\left(\min\left(\frac{1}{\mu_1(\mathbf{D}^o)\log m}, p^{1/2-\epsilon}\right)\right)$	k-sparse $1 \leqslant \|\alpha_j\| \leqslant C$
Agarwal et al. [2013b] *Clustering & ℓ^1*	✓	✗	✗	✓	✗	✓	$p \log mp$	$O\left(\min(1/\sqrt{\mu_1(\mathbf{D}^o)}, m^{1/5}, p^{1/6})\right)$	k-sparse -Rademacher
Agarwal et al. [2013a] ℓ^1 *optim with AltMinDict*	✓	✗	✗	✓	✓	✓	$\frac{p^2}{k^2}$	$O\left(\min(1/\sqrt{\mu_1(\mathbf{D}^o)}, m^{1/9}, p^{1/8})\right)$	k-sparse - i.i.d., $\|\alpha_j\| \leqslant M$
Schnass [2014] *Response maxim. criterion*	✓	✓	✗	✗	✗	$\\|\hat{\mathbf{D}} - \mathbf{D}^o\\|_{2,\infty} \leqslant r$	$\frac{mp^3 k}{r^2}$	$O(1/\mu_1(\mathbf{D}^o))$	"Symmetric decaying"
This contribution *Regularized ℓ^1 criterion with penalty factor λ*	✓	✓	✓	✗	✗	$\\|\hat{\mathbf{D}} - \mathbf{D}^o\\|_F \leqslant r = O(\lambda)$ ✓for $\lambda \to 0$	mp^3	$\mu_k(\mathbf{D}^o) \leqslant 1/4$	k-sparse, $\underline{\alpha} \leqslant \|\alpha_j\|$, $\\|\boldsymbol{\alpha}\\|_2 \leqslant M_{\boldsymbol{\alpha}}$

To conclude ...

Summary

- **Dictionary learning**
  - ✓ widely used in image processing and machine learning
    - [Rubinstein, Bruckstein & Elad, *Dictionaries for Sparse Representation Modeling*,Proc. IEEE, vol. 98, no. 6, pp. 1045–1057, 2010.]
    - [Tosic & Frossard, *Dictionary Learning*, IEEE Sig Proc. Magazine, vol. 28, no. 2, pp. 27–38.]
- Empirically successful **heuristics** ...
  - ✓ batch / online algorithms (K-SVD & al)
- ... together with recent **statistical guarantees**
  - ✓ **sample complexity (also NMF, PCA, sparse PCA ...)**
    - [G. & al, *Sample Complexity of Dictionary Learning and other Matrix Factorizations,* arXiv:1312.3790, December 2013]
  - ✓ **local stability and robustness guarantees**
    - [G. & al, *Sparse and spurious: dictionary learning with noise and outliers,* arxiv 1407.2490*,* July 2014]

What's next ?

- **Towards scalable dictionary learning**
  - [Le Magoarou & G., *Learning computationally efficient dictionaries and their implementation as fast transforms,* http://hal.inria.fr/hal-01010577, June 2014]
- **Sharp** sample complexity ?
  - [Jung & al, *Performance Limits of Dictionary Learning for Sparse Coding,*arXiv:1402.4078, *2014]*
- **Global** identifiability guarantees ?
  - ✓ Empirically yes ... on simple synthetic data
  - ✓ Guarantees from cost functions to ***algorithms*** *?*
    - *http://arxiv.org/abs/1206.5882*
    - *http://arxiv.org/abs/1308.6273,*
    - *http://arxiv.org/abs/1309.1952v1*
- Beyond dictionaries and sparse approximation
  - ✓ *analysis sparsity, classification, clustering ...*

THANKS

PLEASE

projection, learning and sparsity for efficient data processing