IN550 Machine Learning

Decomposizione sparsa di segnali [sparse coding]

Vincenzo Bonifaci

Minimizzazione della norma ℓ_1

$$\begin{aligned} \text{minimize} \quad & |x_1| + |x_2| + \ldots + |x_m| \\ \text{s.t.} \quad & Ax = b \\ & x \in \mathbb{R}^m \end{aligned} \tag{L1}$$

per $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$

Minimizzazione della norma ℓ_1

$$\begin{aligned} \text{minimize} \quad & |x_1| + |x_2| + \ldots + |x_m| \\ \text{s.t.} \quad & Ax = b \\ & x \in \mathbb{R}^m \end{aligned} \tag{L1}$$

per $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$

Problema comune nell'elaborazione dei segnali e in statistica

Minimizzazione della norma ℓ_1

$$\begin{aligned} \text{minimize} \quad & |x_1| + |x_2| + \ldots + |x_m| \\ \text{s.t.} \quad & Ax = b \\ & x \in \mathbb{R}^m \end{aligned} \tag{L1}$$

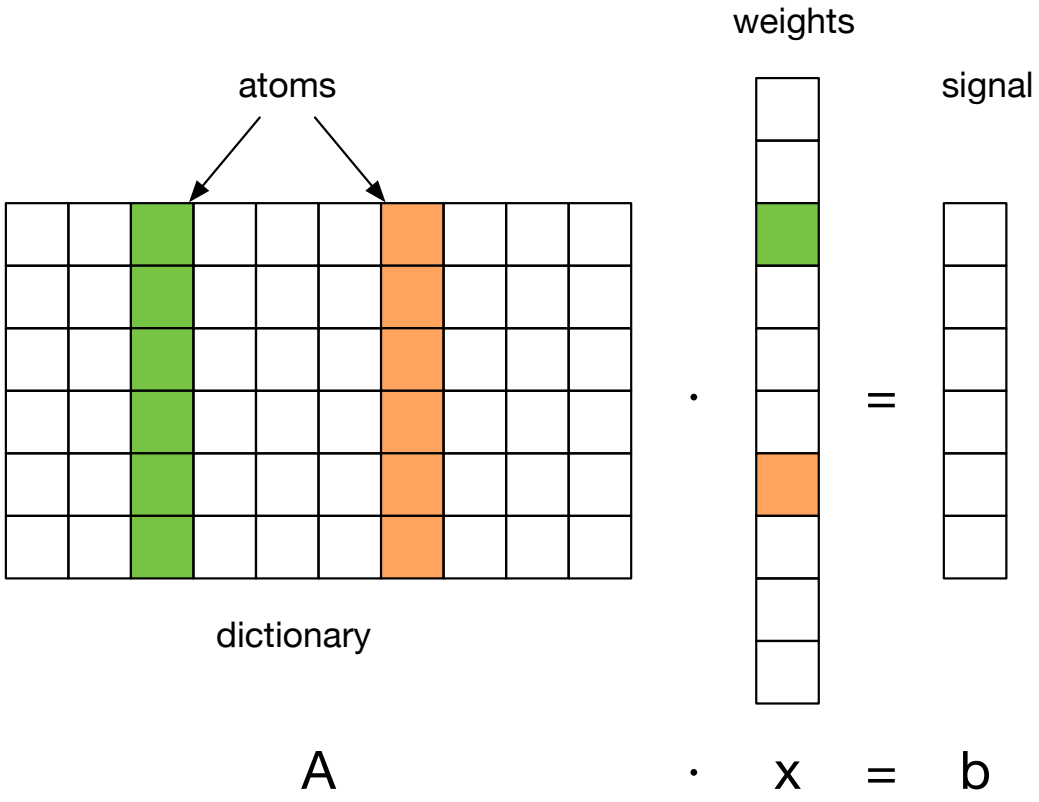
per $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$

Problema comune nell'elaborazione dei segnali e in statistica

Assunzione standard: A ha rango pieno, pari ad n

Decomposizione di segnali

Dati $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$: risolvere $Ax = b$ per $x \in \mathbb{R}^m$



Tipicamente $m > n$: il sistema è sottodeterminato

Regolarizzazione

Per selezionare una particolare soluzione: $\min_x J(x)$ s.t. $Ax = b$

$J(\cdot)$ è un criterio di regolarizzazione

Una scelta comune: $J(x) = \frac{1}{2} \|x\|_2^2$

Perché?

Regolarizzazione

Per selezionare una particolare soluzione: $\min_x J(x)$ s.t. $Ax = b$

$J(\cdot)$ è un criterio di regolarizzazione

Una scelta comune: $J(x) = \frac{1}{2} \|x\|_2^2$

Perché?

- Ha soluzione in forma chiusa: $\hat{x} = A^\top (AA^\top)^{-1} b$
- $\hat{x}$ può essere calcolato risolvendo 1 sistema lineare di equazioni

Ma la soluzione $\hat{x}$ è densa:
tipicamente $\hat{x}_i \neq 0$ per ogni i

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

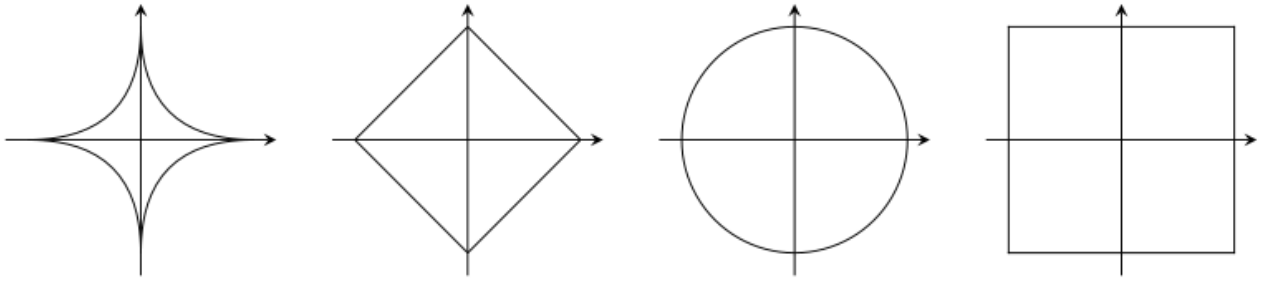
- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ (caso limite per $p \to \infty$)

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ (caso limite per $p \to \infty$)
- $\|x\|_0 \stackrel{\text{def}}{=}$ numero di componenti non-nulle di x (attenzione: non è una norma, nonostanza la notazione!)



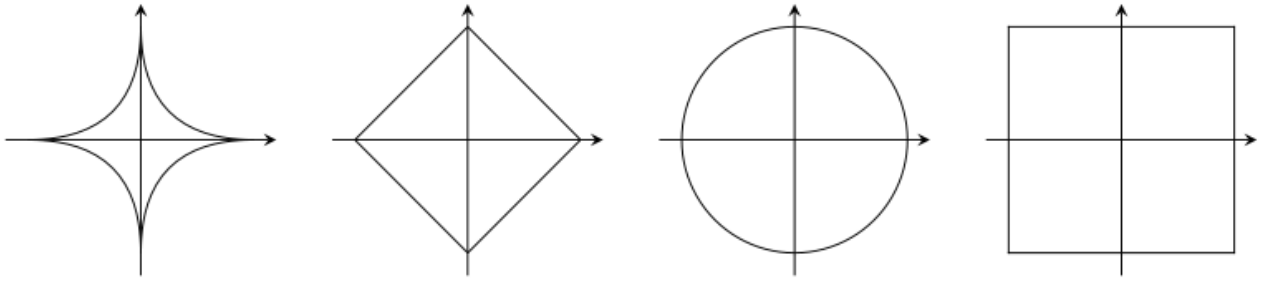
$p = \frac{1}{2}$ $\quad$ $p = 1$ $\quad$ $p = 2$ $\quad$ $p = \infty$

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ (caso limite per $p \to \infty$)
- $\|x\|_0 \stackrel{\text{def}}{=}$ numero di componenti non-nulle di x (attenzione: non è una norma, nonostanza la notazione!)



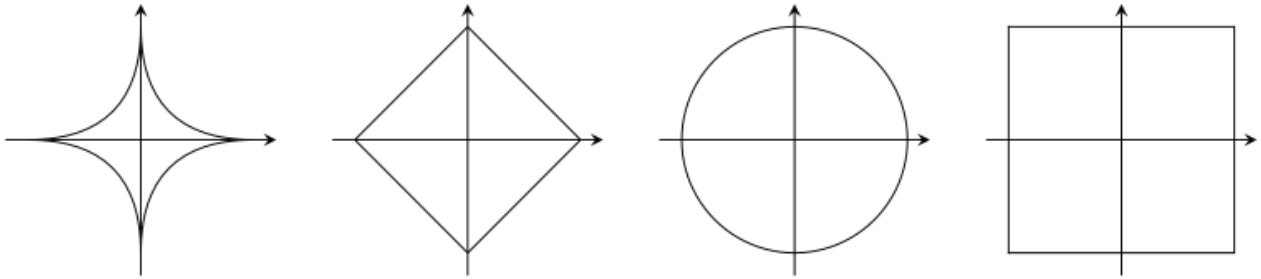
$p = \frac{1}{2}$ $\quad$ $p = 1$ $\quad$ $p = 2$ $\quad$ $p = \infty$

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ (caso limite per $p \to \infty$)
- $\|x\|_0 \stackrel{\text{def}}{=}$ numero di componenti non-nulle di x (attenzione: non è una norma, nonostanza la notazione!)



$p = \frac{1}{2}$ $\quad$ $p = 1$ $\quad$ $p = 2$ $\quad$ $p = \infty$

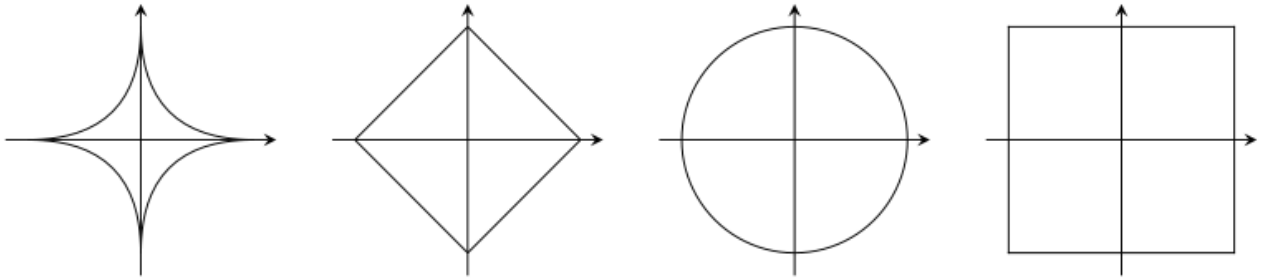
- ℓ_p è una norma sse $p \geq 1$

Regolarizzazione ℓ_p

Funzione ℓ_p: $\|x\|_p = \sqrt[p]{\sum_i |x_i|^p}$

Notazione:

- $\|x\|_\infty \stackrel{\text{def}}{=} \max_i |x_i|$ (caso limite per $p \to \infty$)
- $\|x\|_0 \stackrel{\text{def}}{=}$ numero di componenti non-nulle di x (attenzione: non è una norma, nonostanza la notazione!)

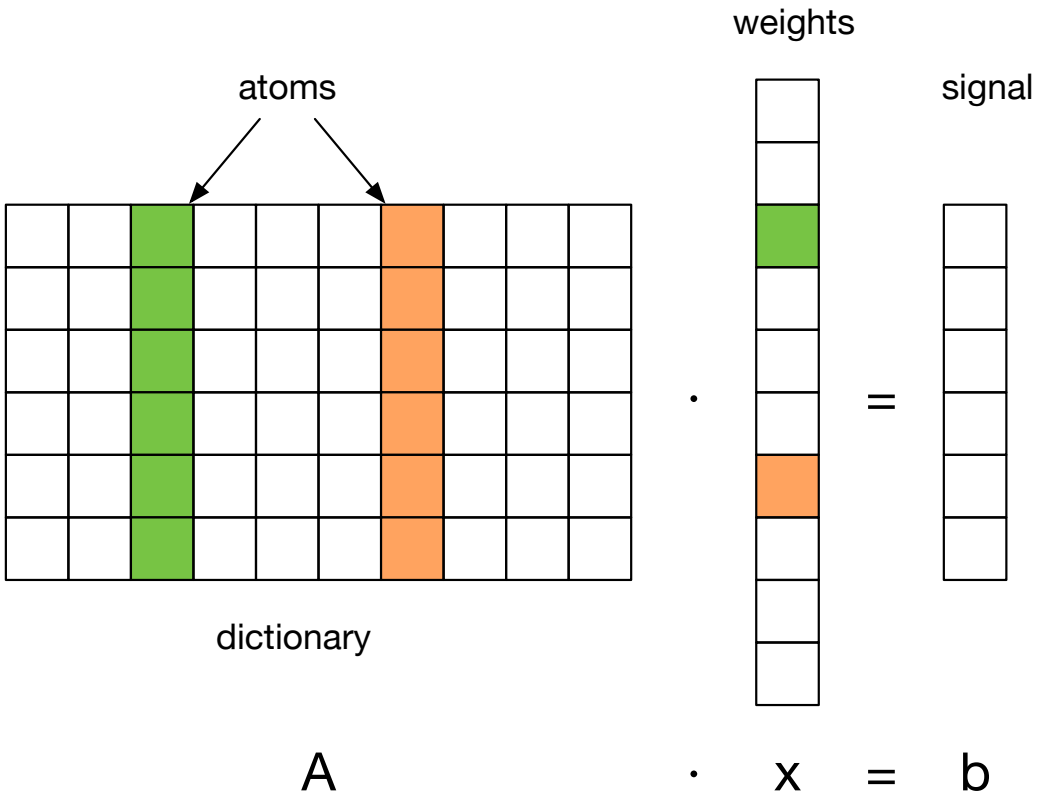


$p = \frac{1}{2}$ $\quad$ $p = 1$ $\quad$ $p = 2$ $\quad$ $p = \infty$

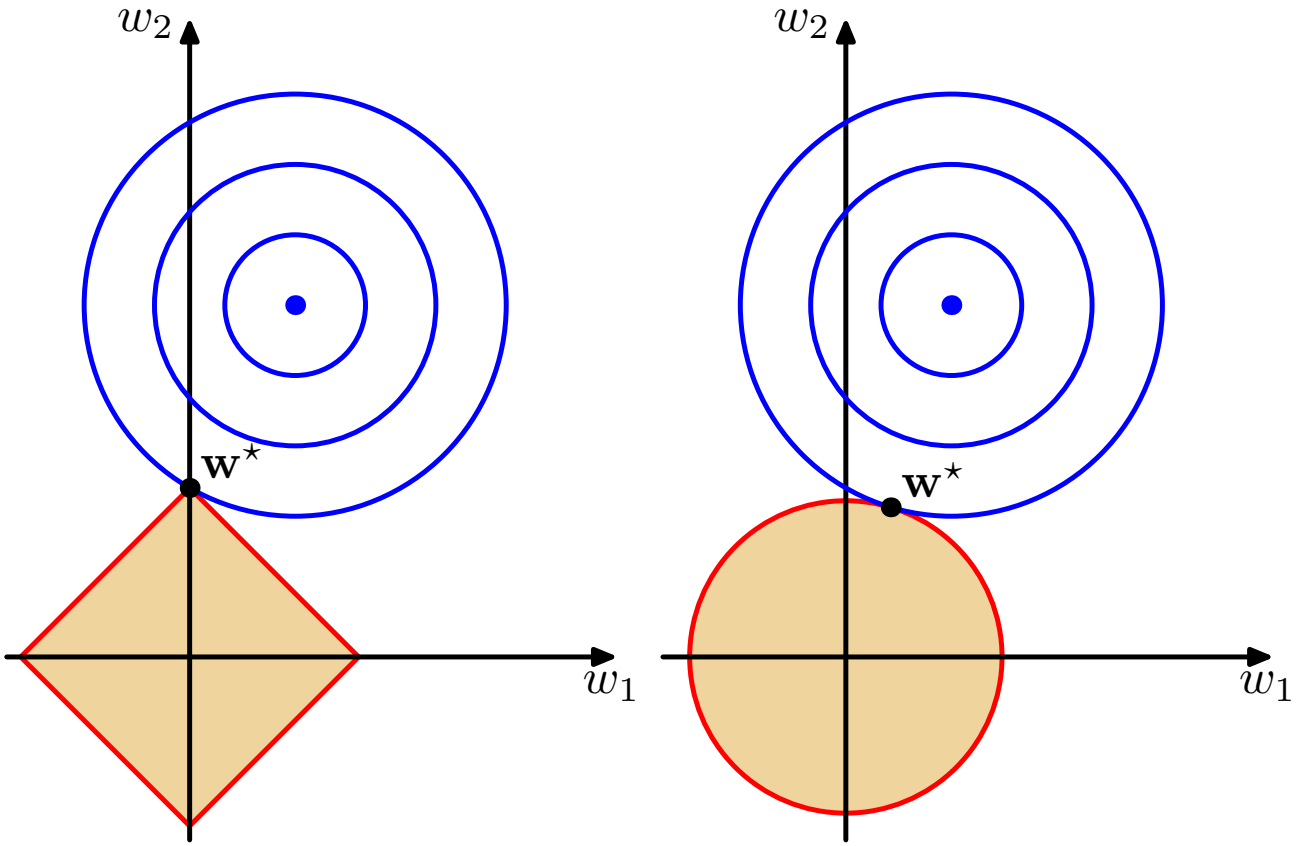
- ℓ_p è una norma sse $p \geq 1$
- ℓ_p è convessa sse $p \geq 1$

Decomposizione di segnali in norma ℓ_p

$$\min \|x\|_p \text{ s.t. } Ax = b$$



Effetto di usare ℓ_1 vs. ℓ_2



La funzione obiettivo $\|x\|_1$ promuove decomposizioni sparse

ℓ_1 promuove decomposizioni più sparse

$$\begin{aligned} \text{minimize} \quad & \|x\|_1 \\ \text{s.t.} \quad & Ax = b \\ & x \in \mathbb{R}^m \end{aligned} \tag{L1}$$

Teorema

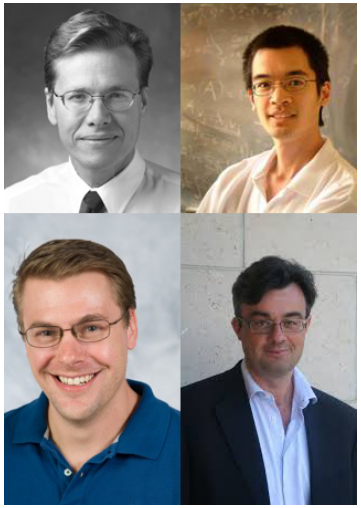
(L1) ammette un minimizzante con al più n componenti non-nulle.

Cio è interessante perché spesso $n \ll m$

Tipicamente, invece, i minimizzanti in norma ℓ_2 sono densi

Compressed sensing [rilevazione compressa]

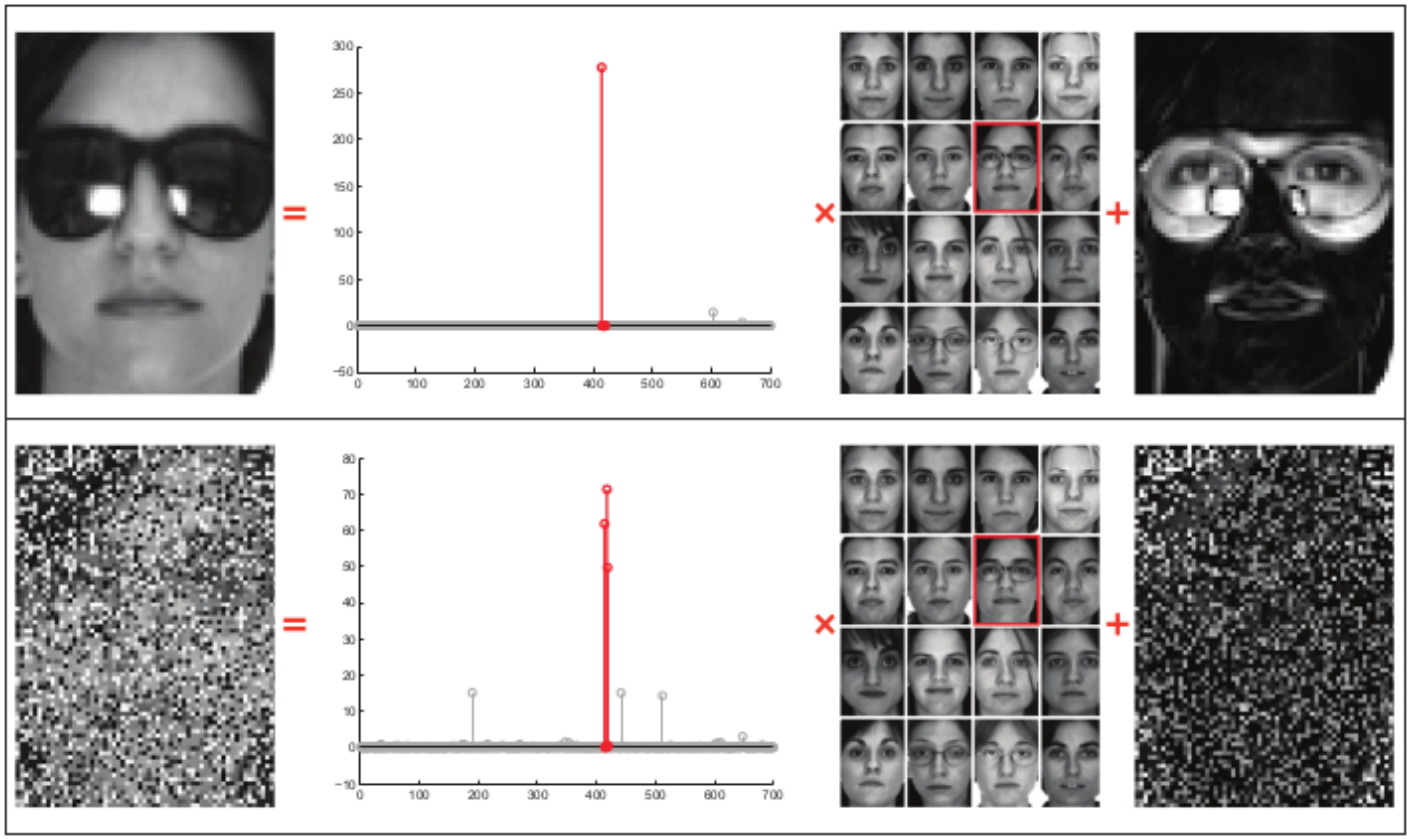
D. Donoho, T. Tao, J.K. Romberg, E. Candes, ed altri hanno dimostrato che un segnale K-sparso può essere ricostruito (in modo robusto) con $O(K)$ osservazioni



Un esempio di applicazione [Wright et al. 2009]

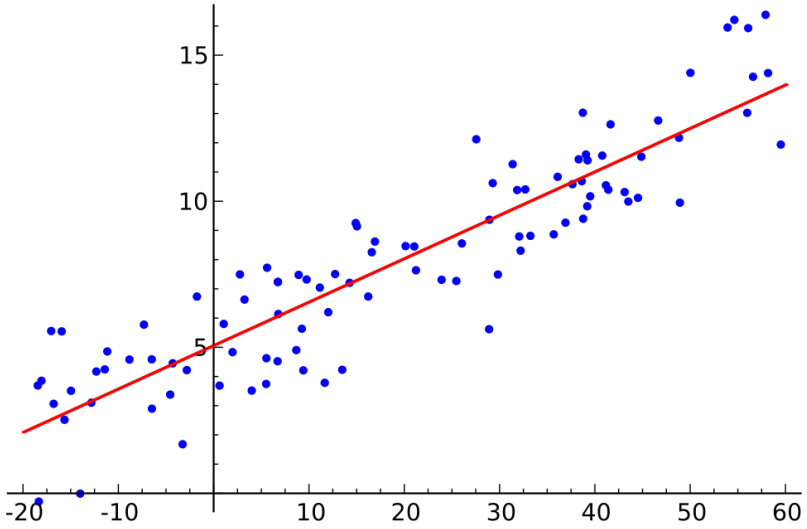
Face Recognition Breakthrough

By using sparse representation and compressed sensing, researchers have been able to demonstrate significant improvements in accuracy over traditional face-recognition techniques.



Problemi di regressione lineare

Dati di input $D \in \mathbb{R}^{m \times n}$, dati di output $y \in \mathbb{R}^m$ $(m \geq n)$



i	D_i	y_i
1	-14.0	0.01
2	2.5	7.52
...	...	
m	58.2	11.9

Fit di una funzione lineare: $\min_{w \in \mathbb{R}^n} \|y - Dw\|$

Cerchiamo soluzioni approssimate del sistema sovradeterminato $Dw = y$

Quale norma per i residui?

I residui sono le differenze $y - Dw$

- Metodo dei minimi quadrati: norma ℓ_2
- Deviazioni assolute minime (Least Absolute Deviations o LAD): norma ℓ_1

Quale norma per i residui?

I residui sono le differenze $y - Dw$

- Metodo dei minimi quadrati: norma ℓ_2
- Deviazioni assolute minime (Least Absolute Deviations o LAD): norma ℓ_1

minimize $(z - z_1)^2 + (z - z_2)^2 + (z - z_3)^2$
$\Rightarrow z =$ media degli z_i

Quale norma per i residui?

I residui sono le differenze $y - Dw$

- Metodo dei minimi quadrati: norma ℓ_2
- Deviazioni assolute minime (Least Absolute Deviations o LAD): norma ℓ_1

minimize $(z - z_1)^2 + (z - z_2)^2 + (z - z_3)^2$
$\Rightarrow z =$ media degli z_i

minimize $|z - z_1| + |z - z_2| + |z - z_3|$
$\Rightarrow z =$ mediana degli z_i

$\Rightarrow \ell_1$ può essere più robusta rispetto ad esempi anomali (outlier)

Breve digressione storica

Breve digressione storica

- Metodo dei minimi quadrati:

Breve digressione storica

- Metodo dei minimi quadrati: Adrien-Marie Legendre, 1805

Breve digressione storica

- Metodo dei minimi quadrati: Adrien-Marie Legendre, 1805
- Deviazioni assolute minime:

Breve digressione storica

- Metodo dei minimi quadrati: Adrien-Marie Legendre, 1805
- Deviazioni assolute minime: Ruggero Boscovich, 1760

Breve digressione storica

- Metodo dei minimi quadrati: Adrien-Marie Legendre, 1805
- Deviazioni assolute minime: Ruggero Boscovich, 1760

DE
LITTERARIA EXPEDITIONE
PER
PONTIFICIAM DITIONEM
AD DIMETIENDOS DUOS MERIDIANI GRADUS
ET CORRIGENDAM MAPPAM GEOGRAPHICAM
JUSSU, ET AUSPICIIS
BENEDICTI XIV.
PONT. MAX.
SUSCEPTA A PATRIBUS SOCIET. JESU
CHRISTOPHORO MAIRE
ET
ROGERIO JOSEPHO BOSCOVICH.
ROMÆ MDCCLV.

Boscovich, R. J. (1760). *De recentissimis graduum dimensionibus, et figura, ac magnitudine terrae inde derivanda*.

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

$$Ax = b$$

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

$$Ax = Ay$$

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

$$Ax = Ay \Leftrightarrow y - x \in \ker A$$

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

$$Ax = Ay \Leftrightarrow y - x \in \operatorname{im} D$$

Regressione LAD è equivalente al problema (L1)

Il problema delle deviazioni assolute minime

$$\min_{w \in \mathbb{R}^n} \|y - Dw\|_1$$

è equivalente a

$$\min_{x \in \mathbb{R}^m} \|x\|_1 \text{ s.t. } Ax = b$$

per opportune A, b

Dimostrazione

Si prendano A, b tali che $\ker A = \operatorname{im} D$ e $b = Ay$, allora

$$Ax = Ay \;\Leftrightarrow\; y - x \in \operatorname{im} D \;\Leftrightarrow\; x = y - Dw \text{ per qualche } w \in \mathbb{R}^n$$

(L1) è formulabile come problema di ottimizzazione lineare

(L1)

$$\begin{aligned} \min \ & \|x\|_1 \\ \text{s.t.} \ & Ax = b \\ & x \in \mathbb{R}^m \end{aligned}$$

(LP)

$$\begin{aligned} \min \ & 1^\top z \\ \text{s.t.} \ & z + x \geq 0 \\ & z - x \geq 0 \\ & Ax = b \\ & x, z \in \mathbb{R}^m \end{aligned}$$

m variabili, n vincoli $\Rightarrow$ $2m$ variabili, $n + 2m$ vincoli

(L1) è formulabile come problema di ottimizzazione lineare

(L1)

$$\begin{aligned} \min \ & \|x\|_1 \\ \text{s.t.} \ & Ax = b \\ & x \in \mathbb{R}^m \end{aligned}$$

(LP)

$$\begin{aligned} \min \ & 1^\top z \\ \text{s.t.} \ & z + x \geq 0 \\ & z - x \geq 0 \\ & Ax = b \\ & x, z \in \mathbb{R}^m \end{aligned}$$

m variabili, n vincoli $\Rightarrow$ $2m$ variabili, $n + 2m$ vincoli

Lasso

L'algoritmo LARS (Least Angle Regression) affronta il problema Lasso:

$$\begin{aligned} \text{minimize} \quad & \|y - Dw\|_2^2 + \alpha \|w\|_1 \\ \text{s.t.} \quad & w \in \mathbb{R}^n \end{aligned}$$

α: iperparametro di regolarizzazione

Un α maggiore incoraggia maggiore sparsità, a scapito di un fit meno fedele

In scikit-learn, visto come regressione regolarizzata:

```
model = sklearn.linear_model.Lasso(alpha=1.0)
model.fit(D, y)
```

Visto come decomposizione di segnali $Ax \approx b$:

```
sklearn.decomposition.SparseCoder(dictionary=A,
  transform_algorithm='lasso_lars', transform_alpha=1.0)
x = coder.transform(b)
```

Orthogonal Matching Pursuit (OMP)

OMP è un algoritmo avido (euristico) che affronta il seguente problema:

$$\begin{aligned} \text{minimize} \quad & \|y - Dw\|_2^2 \\ \text{s.t.} \quad & \|w\|_0 \leq k \\ & w \in \mathbb{R}^n \end{aligned}$$

k: numero massimo di componenti sparse desiderato

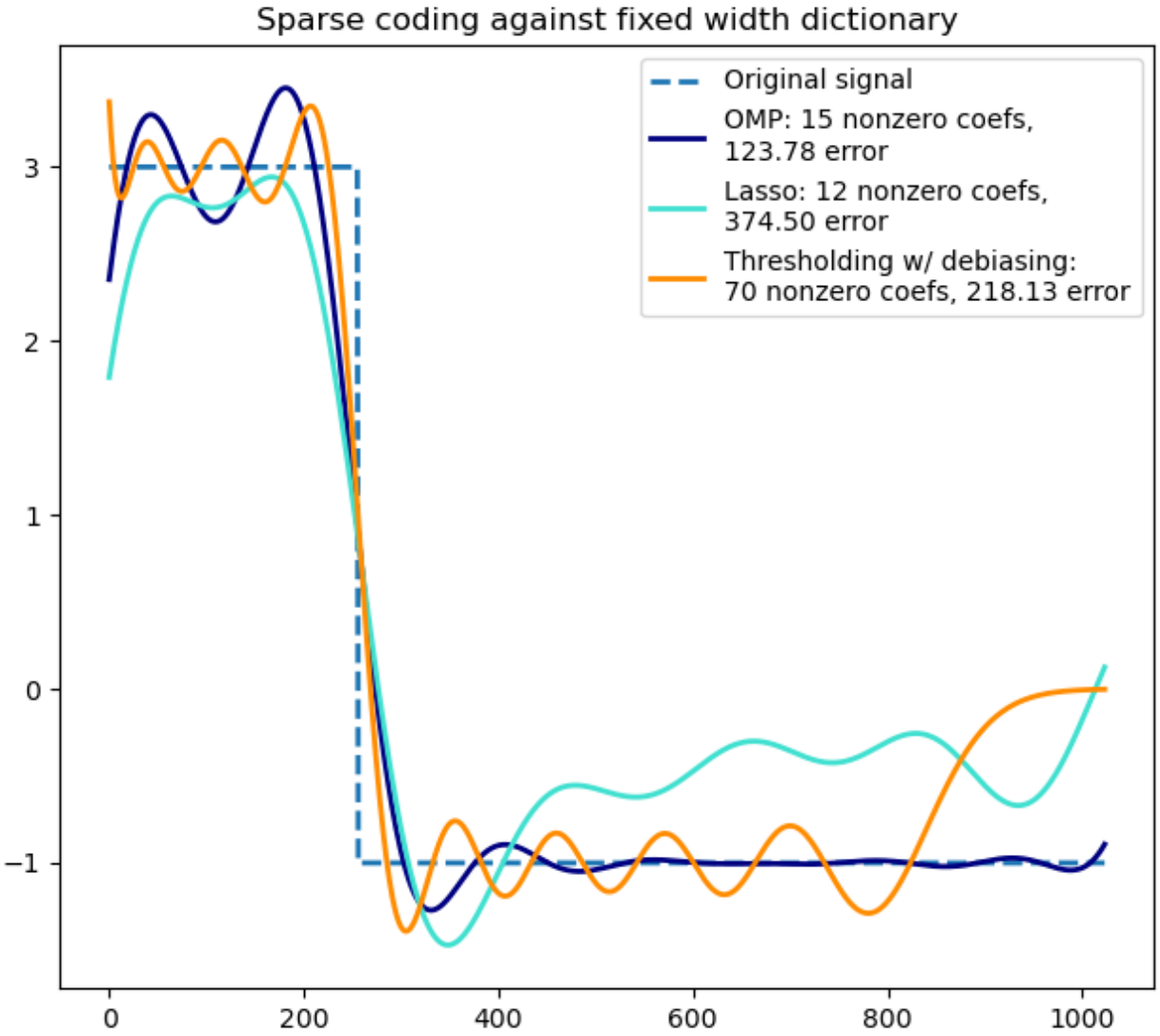
In scikit-learn, visto come regressione vincolata:

```
model = sklearn.linear_model.OrthogonalMatchingPursuit(
  n_nonzero_coefs=k)
model.fit(D, y)
```

Visto come decomposizione di segnali $Ax \approx b$:

```
coder = sklearn.decomposition.SparseCoder(dictionary=A,
  transform_algorithm='omp', transform_n_nonzero_coefs=k)
x = coder.transform(b)
```

Esempio


Sparse coding against fixed width dictionary
Original signal
OMP: 15 nonzero coefs, 123.78 error
Lasso: 12 nonzero coefs, 374.50 error
Thresholding w/ debiasing: 70 nonzero coefs, 218.13 error
3
2
1
0
−1
0
200
400
600
800
1000