Modelos dispersos para el procesamiento de señales y aprendizaje de máquinas

Alejandro Weinstein

Escuela de Ingeniería Biomédica
Universidad de Valparaíso

28 de Agosto de 2013

¿Qué es una señal dispersa?

Definición

Una señal $x \in \mathbb{R}^N$ es K-dispersa ($K < N$) si a lo más K coeficientes son distintos de zero. Formalmente

$$||x||_0 \leq K$$

Ejemplo: $N = 10$, $K = 2$

$$x = \begin{bmatrix} 0 & 0 & 3.14 & 0 & 0 & 0 & 0 & -2.71 & 0 & 0 \end{bmatrix}$$

En general nos interesan los casos en que $K \ll N$

¿Qué es una señal dispersa?

Definición

Una señal es aproximadamente K-dispersa si, dado $\epsilon > 0$

$$||x - x_K||_2 \leq \epsilon,$$

donde x_K es igual a x en las K posiciones más grande (en magnitud), y cero en el resto.

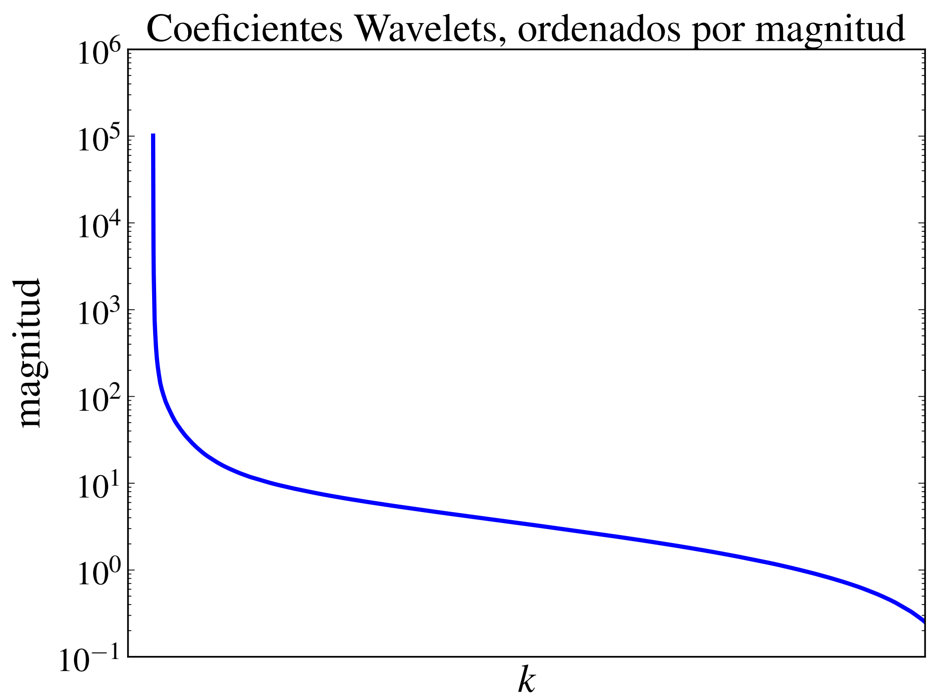
Ejemplo: $N = 10$, $K = 2$

$$x = \begin{bmatrix} -0.02 & 0.1 & 3.14 & -0.03 & -0.01 & 0.01 & -0.02 & -2.71 & 0.02 & 0.03 \end{bmatrix}$$
$$x_K = \begin{bmatrix} 0 & 0 & 3.14 & 0 & 0 & 0 & 0 & -2.71 & 0 & 0 \end{bmatrix}$$
$$||x - x_K||_2 = 0.12$$

¿Por qué señales dispersas?

Imagen original, 614400 pixels

Aproximacion usando 10 % de los coeficientes



http://www.flickr.com/photos/bibiweb/6931027375/

Usando la dispersión como modelo

"Esencialmente, todos los modelos son incorrectos, pero algunos modelos son útiles."

George Box

Compressive Sensing (CS)

Es posible diseñar sistemas de adquisición que tome un número de mediciones muy por debajo del tamaño de la señal

Útil cuando cada medición es:

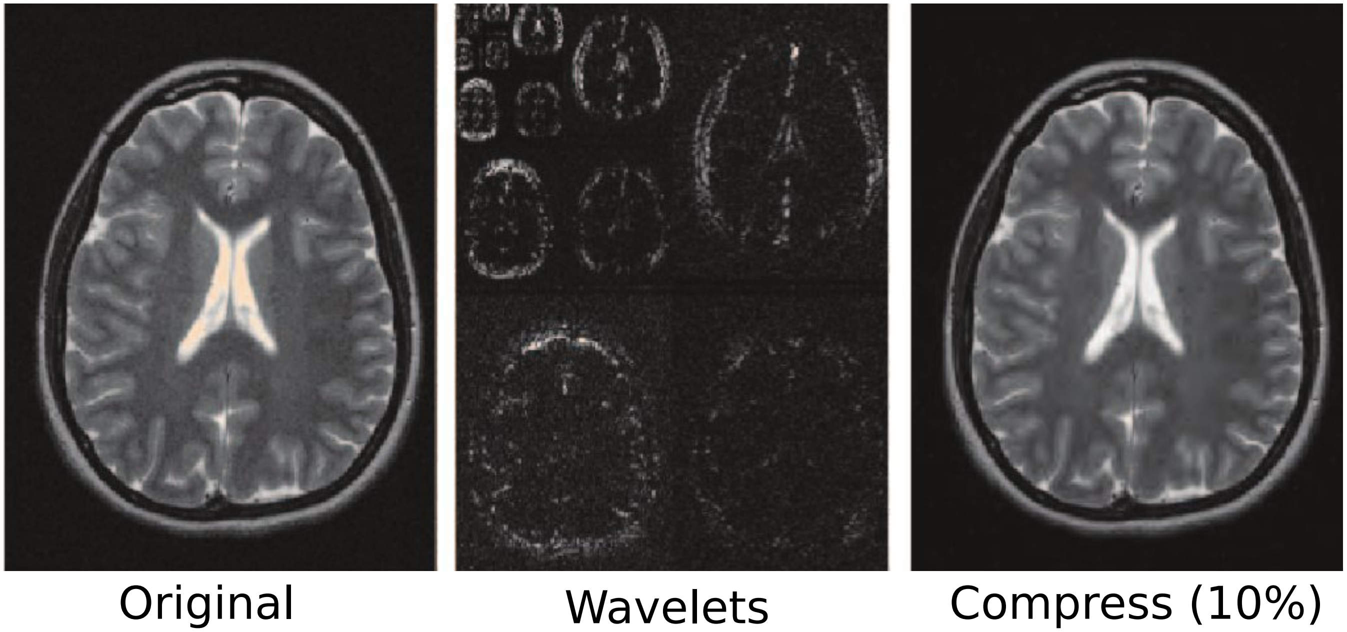
- Cara
- Lenta

E.J. Candès and M.B. Wakin, (2008). An Introduction To Compressive Sampling

Ejemplo: MRI

CS aprovecha dos hechos:

- Las imágenes médicas se pueden comprimir
- En MRI cada medición contiene información de todas las muestras

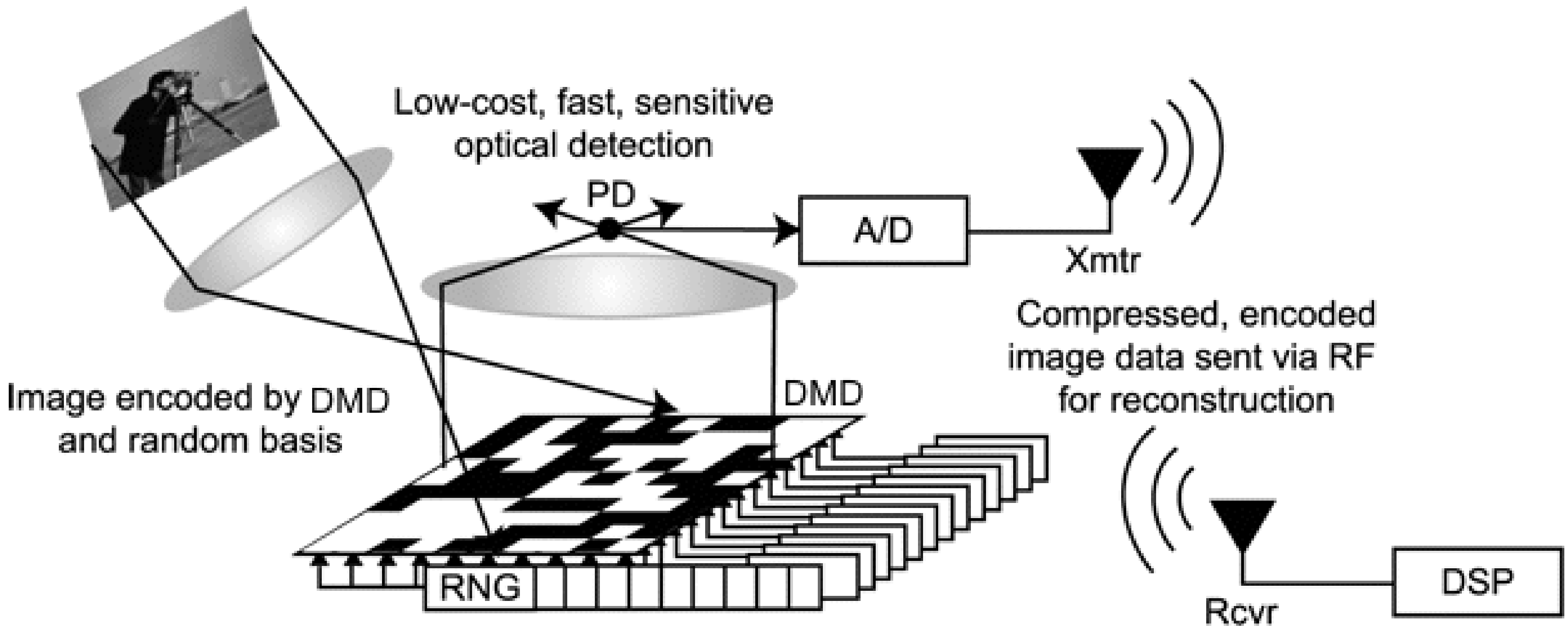


(M. Lustig et al., “Compressed Sensing MRI”)

Usando CS es posible reducir el tiempo de escaneo de una angiografía 3D de 22 a 4.4 segundos

M. Lustig et al., Sparse MRI: The application of compressed sensing for rapid MR imaging

Ejemplo: cámara de 1 pixel



http://dsp.rice.edu/cscamera

La cámara de 1 pixel seis años después

InView Technology compressive sensing workstation includes objective lens

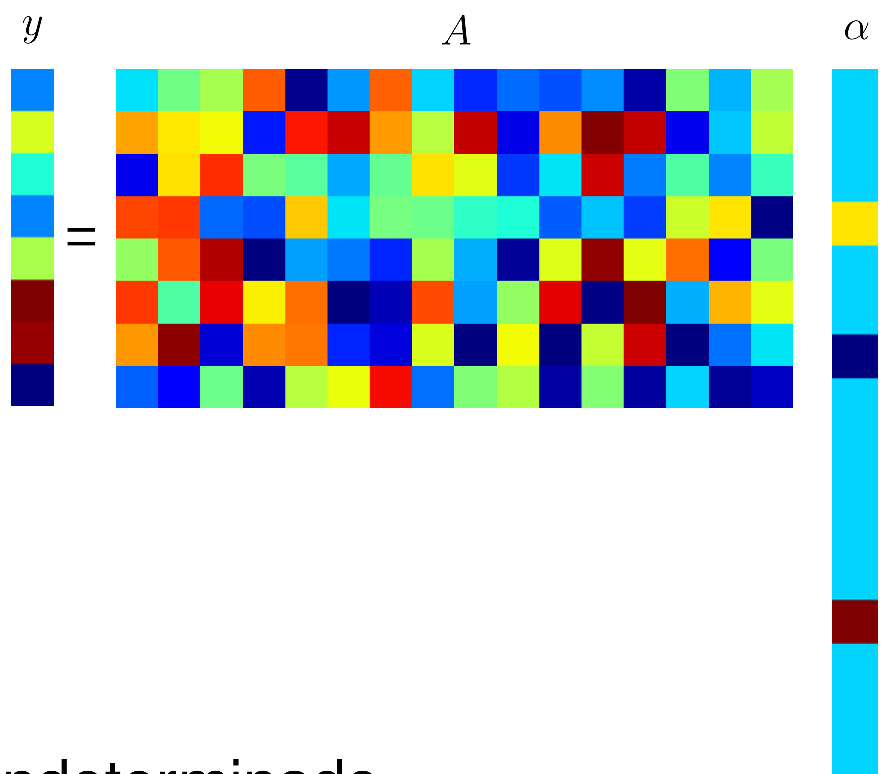
01/23/2013
Posted by Lee Mather
Associate Editor

The InView 220 includes a sensor and sensor controller for development and testing of compressive sensing algorithms. The sensor includes an objective lens, a 1920 × 1080 TI DMD with a mirror update rate of 10 kHz, a 4 × 8 SWIR detector array, and an analog-to-digital converter.
InView Technology
Austin, TX
http://www.laserfocusworld.com/articles/2013/01/inview-technology-compressive-sensing-workstation.html

¿Cómo funciona?

$$\left.\begin{array}{l} x = \Psi\alpha \\ y = \Phi x \end{array}\right\} \quad \Rightarrow \quad y = \underbrace{\Phi\Psi}_{A}\alpha$$

Dado y y A queremos recuperar α



- El sistema es indeterminado
- De las ∞ soluciones, elegimos la más dispersa

¿Cómo funciona?

Queremos resolver

$$\begin{array}{ll} \underset{\alpha}{\text{minimize}} & \|\alpha\|_0 \\ \text{subject to} & A\alpha = y \end{array}$$

pero es muy difícil (NP-hard)

En vez, relajamos el problema original y resolvemos

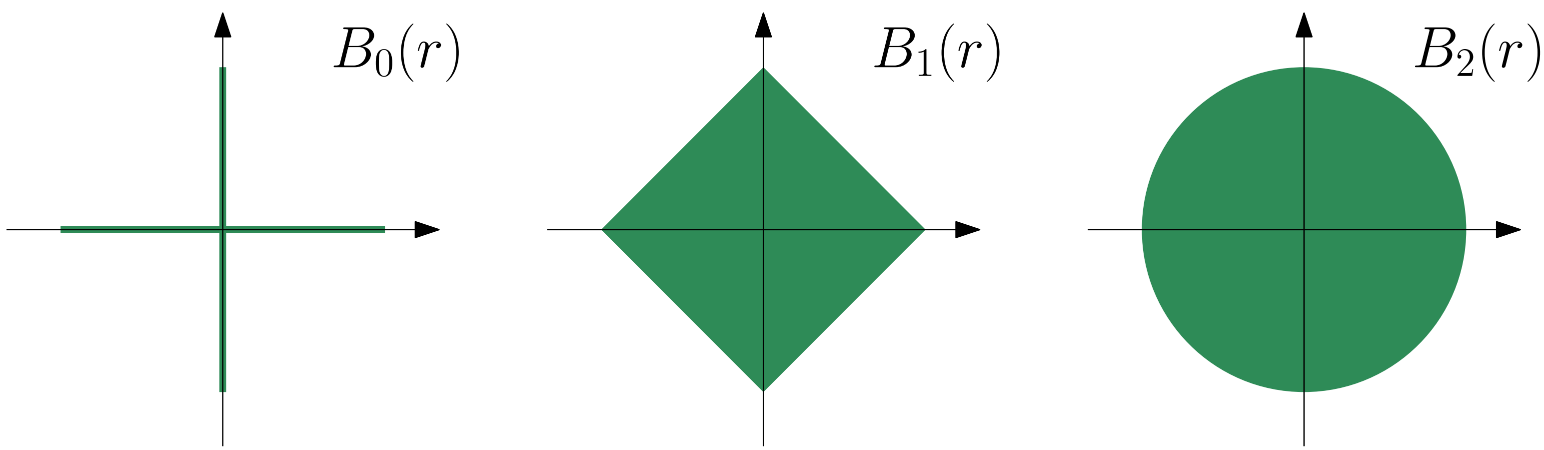
$$\begin{array}{ll} \underset{\alpha}{\text{minimize}} & \|\alpha\|_1 \\ \text{subject to} & A\alpha = y \end{array}$$

Existen condiciones bajo las cuales ambos problemas son equivalentes

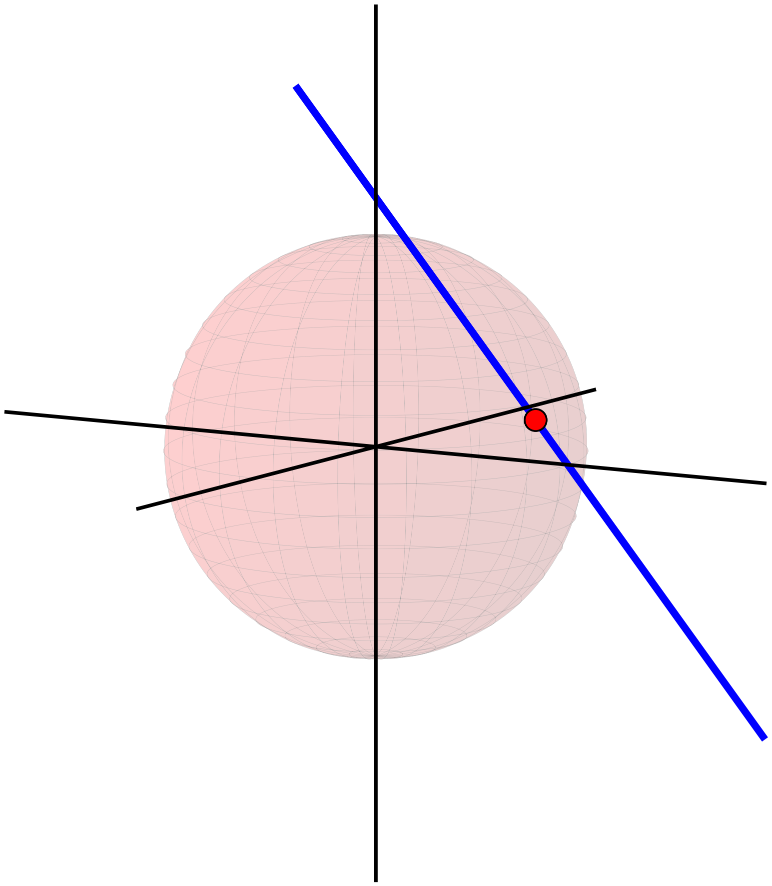
Interpretación geométrica

Norma p: $||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$

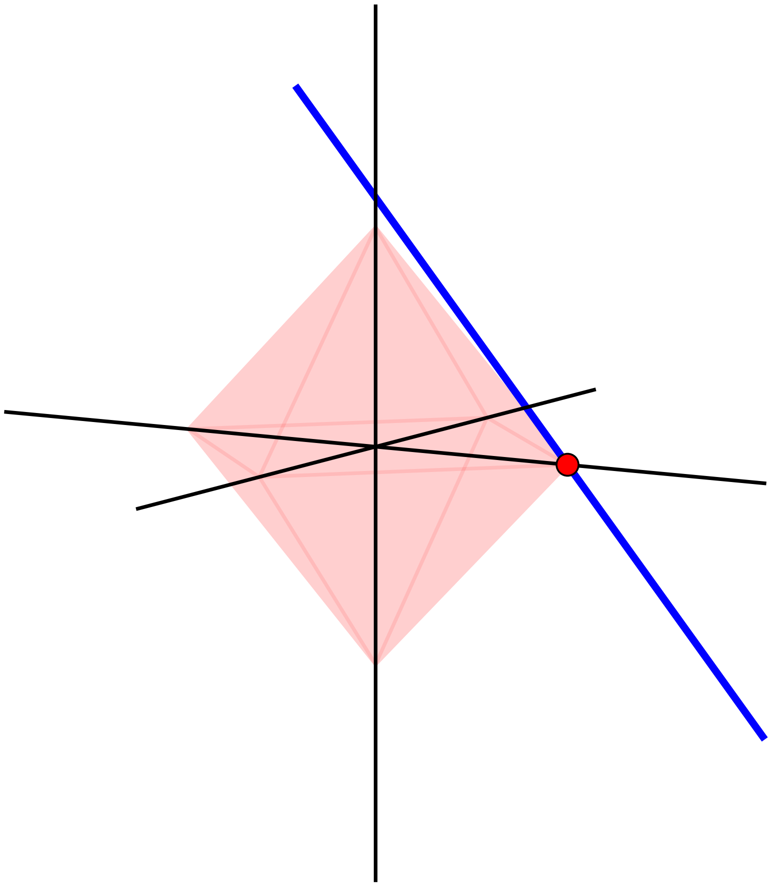
Bola p de diámetro r: $B_p(r) = \{x \in \mathbb{R}^n : ||x||_p < r\}$



Interpretación geométrica



$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \|\alpha\|_2 \\ \text{subject to} \quad & A\alpha = y \end{aligned}$$

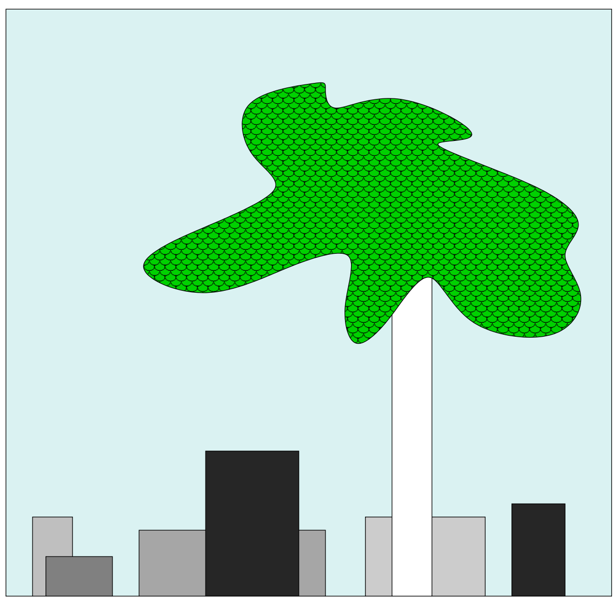


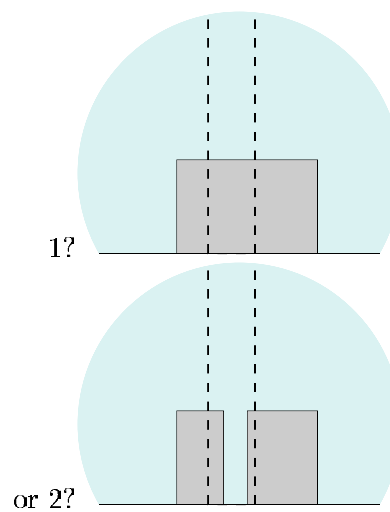
$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \|\alpha\|_1 \\ \text{subject to} \quad & A\alpha = y \end{aligned}$$

Otro punto de vista

Se puede pensar en CS como una instancia de la navaja de Occam:

> "Entre todas las explicaciones que se ajustan a los datos
> acepta la más simple"





David Mackay, "Information Theory, Inference, and Learning Algorithms"

CS en presencia de ruido

Si las observaciones son ruidosas:

$$y = \Phi x + \eta, \qquad ||\eta||_2 \leq \epsilon$$

podemos recuperar la señal resolviendo

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \|\alpha\|_1 \\ \text{subject to} \quad & ||A\alpha - y||_2 \leq \epsilon \end{aligned}$$

Garantía teórica

Bajo ciertas condiciones, es posible demostrar que

$$||x^{\star} - x||_2 \leq \frac{C_0 \, ||x - x_K||_1}{\sqrt{K}} + C_1 \epsilon,$$

donde

$x^{\star}$: Señal recuperada

C_0, C_1: Constantes

K: Nivel de dispersión

x_K: Mejor K-aproximación de x

E.J. Candès and M.B. Wakin, (2008). An Introduction To Compressive Sampling

Garantía teórica

$$\left.\begin{aligned} x &= \Psi\alpha \\ y &= \Phi x \end{aligned}\right\} \quad \Rightarrow \quad y = \underbrace{\Phi\Psi}_{A}\alpha$$

Ψ: Base de representación

Φ: Base de muestreo

Coherencia entre la base de representación y muestreo:

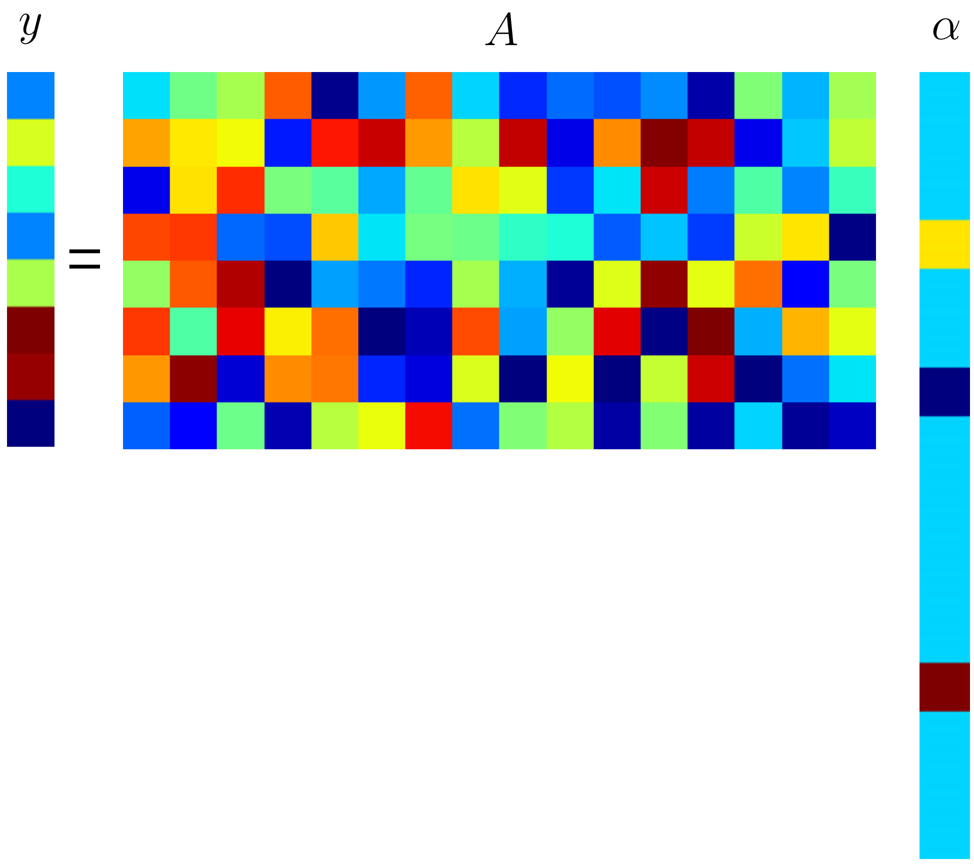
$$\mu(\Phi, \Psi) = \sqrt{n} \max_{1 \leq j,k \leq n} |\langle \phi_j, \psi_k \rangle|$$

E.J. Candès and M.B. Wakin, (2008). An Introduction To Compressive Sampling

Garantía teórica

Necesitamos pares Φ, Ψ con baja coherencia. Ejemplo:

- Tiempo y frecuencia
- Frecuencia y wavelets
- Aleatorio y cualquier base fija



Más allá de modelos dispersos

(a) ℓ_1/ℓ_2-norm ball without overlaps: $\Omega(\mathbf{w}) = \|\mathbf{w}_{\{1,2\}}\|_2 + |\mathbf{w}_3|$

(b) ℓ_1/ℓ_2-norm ball with overlaps: $\Omega(\mathbf{w}) = \|\mathbf{w}_{\{1,2,3\}}\|_2 + |\mathbf{w}_1| + |\mathbf{w}_2|$

- Señales dispersas agrupadas
- Minimizar la complejidad de Kolgomorov o Minimum Description Length (MDL)
- Dictionary learning

Bach et al., (2011). Structured sparsity through convex optimization

Aprendizaje de máquinas

El problema que uno resuelve en CS se puede escribir como

$$\underset{x}{\text{minimize}} \quad \frac{1}{2}\,||Ax - y||_2^2 + \lambda\|x\|_1$$

En aprendizaje de máquinas/estadística esta formulación se conoce como lasso (*least absolute shrinkage and selection operator*).

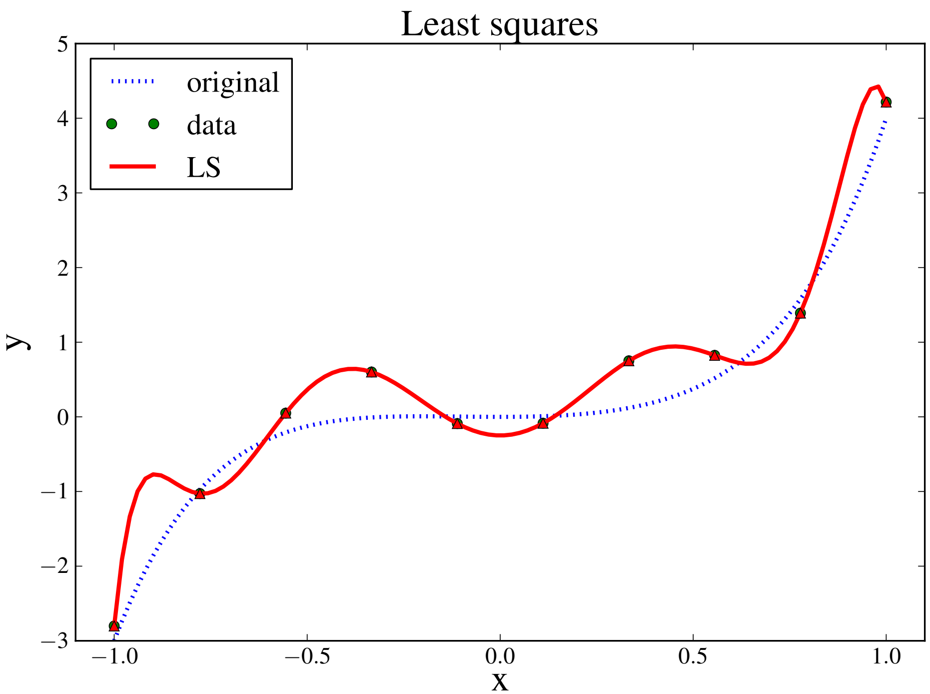
R. Tibshirani (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society.

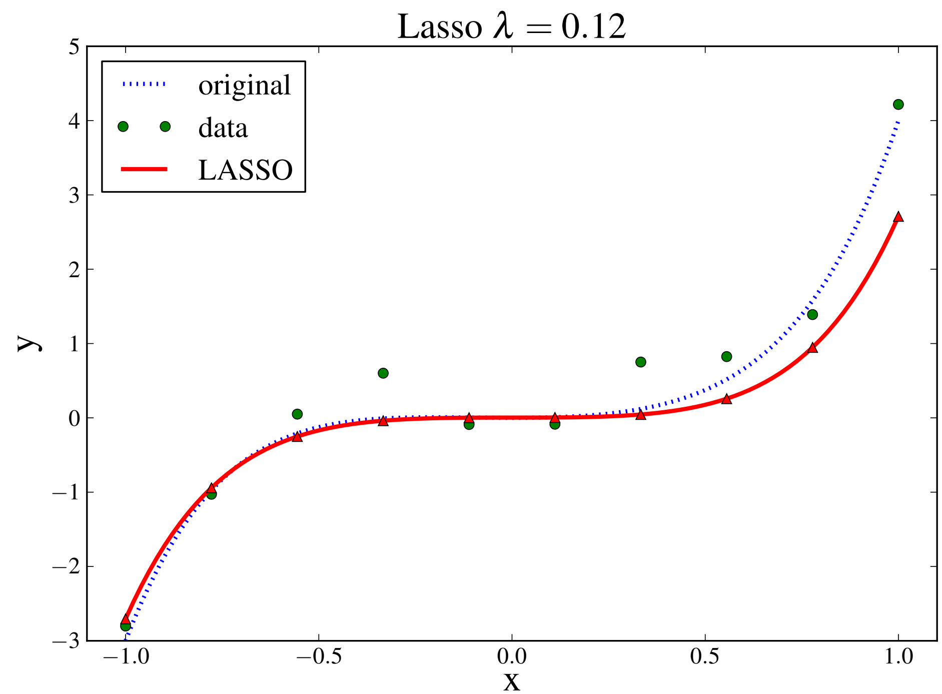
Lasso: Ejemplo

Observamos 10 datos en el intervalo $[-1, 1]$ generados por

$$y = 2x^5 + 1.5x^3 + 0.5x^2 + \eta, \eta \quad \sim \mathcal{N}(0, 0.4^2)$$

Asumimos que los datos siguen un polinomio de a lo mas grado 10





$$\hat{y}_{LS} = -54x^{10} + 3.1x^9 + 39x^8 - 3.2x^7 + 77x^6 + 2.4x^5 - 84x^4 + 3x^3 + 24x^2 - 0.01x - 0.8$$

$$\hat{y}_{LASSO} = 2.6x^5 + 1.5x^3 + 0.01$$

Graph Lasso

Dadas las muestras $y_1, \ldots, y_n \sim \mathcal{N}(\sigma_p, \mu)$, *Graph Lasso* estima la covarianza como

$$\hat{\Sigma}^{-1} = \underset{X \succ 0}{\operatorname{argmax}} \{ \underbrace{\log \det X - \operatorname{tr}(SX)}_{\text{log-likelihood}} - \underbrace{\lambda \, ||X||_1}_{\text{penalización}} \},$$

con

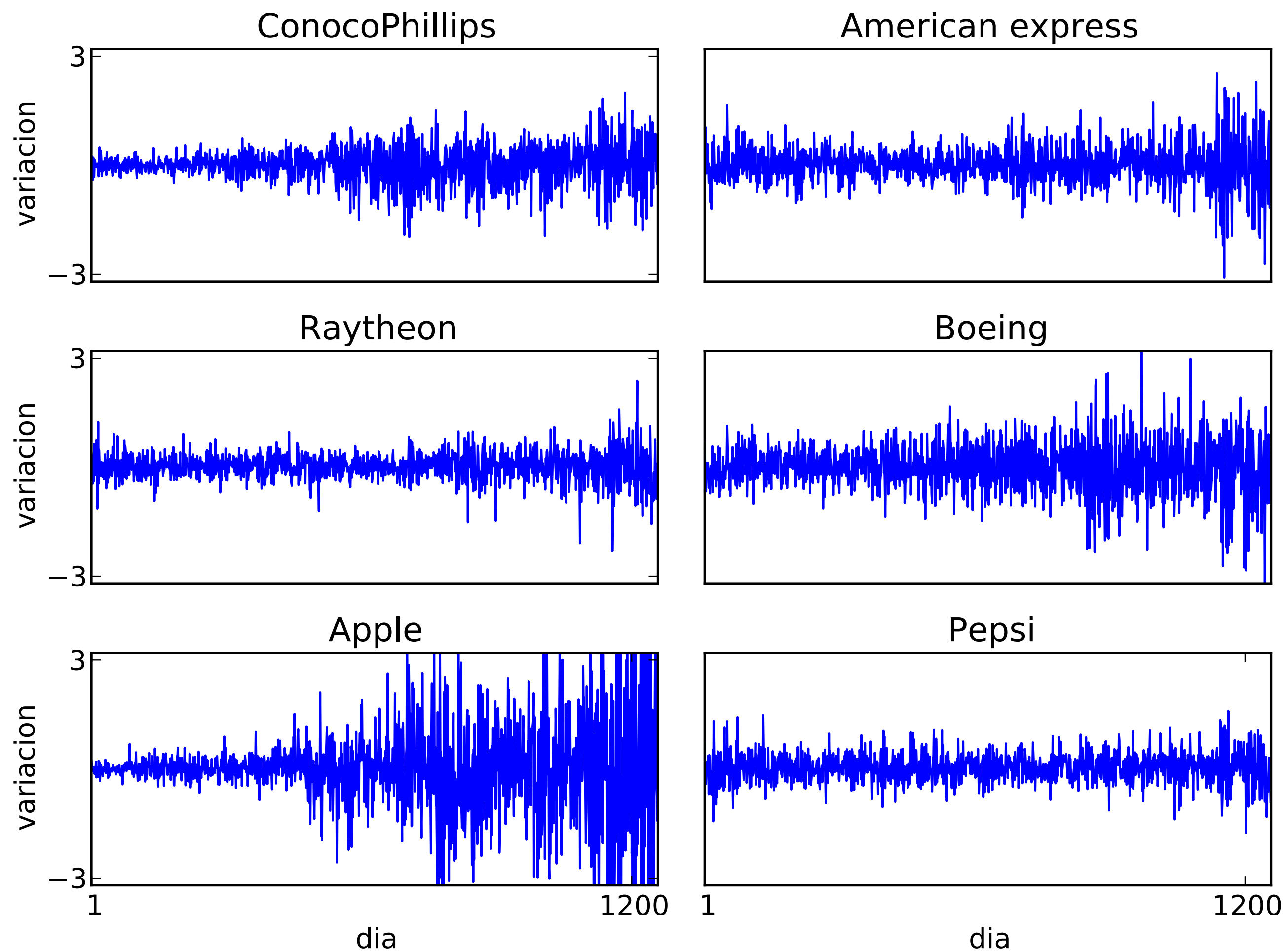
$$S := \frac{1}{n} \sum_{k=1}^{n} (y_k - \mu)(y_k - \mu)^T, \qquad ||X||_1 = \sum_{i,j} \operatorname{abs}(x_{ij})$$

Notar que $x_{ij} = 0$ implica que las variables x_i y x_j son condicionalmente independientes.

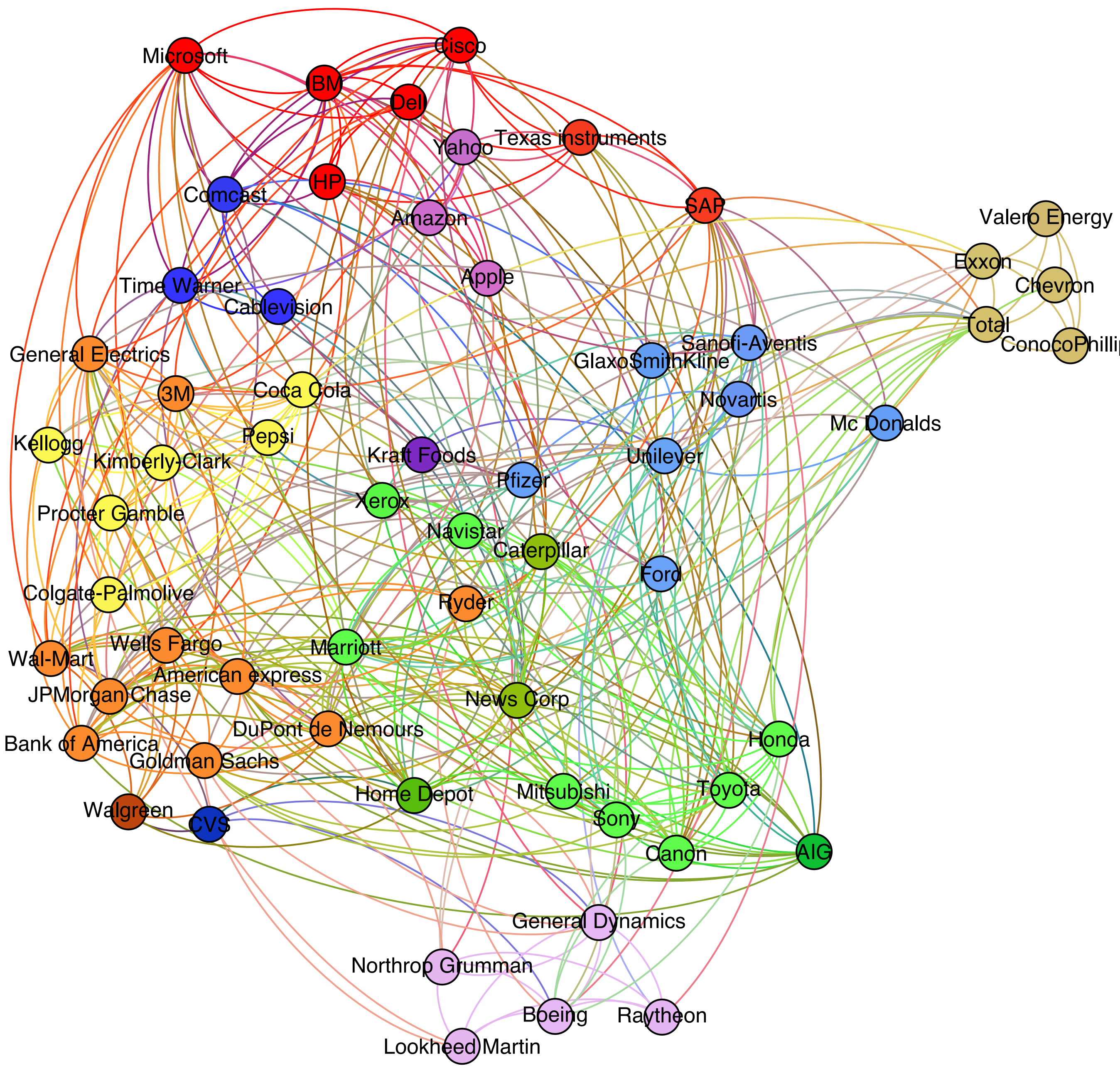
O. Banerjee et al., (2007). Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data

Ejemplo Graph Lasso: Correlación entre acciones

Dada la variación diaria de 61 acciones durante 5 años queremos encontrar posibles estructuras entre las acciones



Ejemplo Graph Lasso: Correlación entre acciones



http://scikit-learn.org/stable/auto_examples/applications/plot_stock_market.html