Some Fairy Tales from Sparseland

Christine De Mol

Université Libre de Bruxelles
Dept Math. and ECARES

“Fritto Misto” in onore di Mario Bertero
Genova, February 2, 2011

High-dimensional and complex data

- How to extract meaningful information (infer models) from a data-rich environment? e.g. in
  - Physics and Engineering: inverse imaging, computer vision, etc.
  - Bioinformatics: genomics and proteomics
  - Economics
- Different frameworks:
  - compressed sensing/sampling
  - regression/learning
  - inverse problems

Sampling or sensing problems

- Take discrete samples from a signal f and design a recovery scheme from the samples

- Classical example: Shannon's sampling theorem
a bandlimited signal $f(x)$ (with cutoff frequency ν_{max}) can be uniquely recovered from equidistant samples at the Nyquist rate:

$$f(x) = \sum_{k \in \mathbb{Z}} \frac{\sin[\frac{\pi}{\delta}(x - k\delta)]}{\frac{\pi}{\delta}(x - k\delta)} f(k\delta) \qquad (\delta = 1/2\nu_{max})$$

- Generalization: design a measurement (encoding) scheme (matrix or linear operator) Φ to "sense" an unknown signal f through Φf and devise an associated recovery (decoding) scheme allowing to compute f from Φf.

Linear regression problem

- “Input” (data) matrix: $X = \{x_{ij}\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$
- “Output” (response): y_i for each i (“supervised” setting)
- Assume linear dependence:

$$y_i = \sum_j x_{ij}\beta_j \qquad \text{or} \qquad y = X\beta$$

- Two distinct problems:
  Prediction (“generalization”): predict (forecast) the response y
  Identification (Variable Selection): find the regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$
  and/or identify the relevant predictors (essential for interpretation!)

Inverse problems

- Recover target ("object") f from indirect measurements, i.e. from an "image" $g = Af$ where A is a linear operator modelling the action of an instrument or imaging device (microscope, telescope, scanner, scattering medium, etc.)
- In continuous models, f is a function and A is an integral operator (acting in some Hilbert space of functions)

$$(Af)(x) = \int K(x, x')\, f(x')\, dx'$$

  where the kernel $K(x, x')$ is a (known) response function (deconvolution (deblurring) problems: $K(x, x') = K(x - x')$)
- In discrete models, A is a matrix, typically ill-conditioned (e.g. when arising from discretization of the previous continuous model).

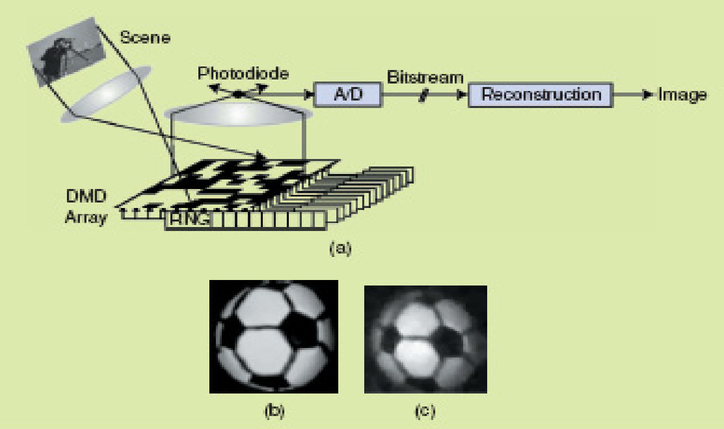
Compressed sensing or Compressive sampling

- To determine a signal *f* in a *p*-dimensional space, one needs to take in principle at least *p* linear measurements.
- Can we take profit of the sparsity of the signal (i.e. the fact that *f* has only *k* non-zero components, with $k << p$) to decrease the number of measurements?
  NB. The location of these components is unknown!
- This is the question addressed by the emerging field of compressed sensing, compressive sampling or else compressive sensing
  (Candès and Tao 2006; Candès Romberg and Tao 2006; Donoho 2006; etc. - see http://dsp.rice.edu/cs)

Compressed sensing or Compressive sampling

- The answer is yes, provided the measurement matrix Φ is close to an isometry on the class (not linear subspace!) of all k-sparse signals. This ensures that the recovery of f from $g = \Phi f$ will be possible (well conditioned)
- The price to pay for lowering the number of measurement values needed is the randomization of Φ
- Typical kind of result: take the matrix elements of Φ be i.i.d. random variables taken from a Gaussian distribution with mean zero and variance $1/p$; then k-sparse signals of length p can be recovered from only $m = ck \log(p/k) << p$ of these random measurements ("with overwhelming probability"!)
- Decoder (recovery scheme): minimize the L_1-norm of f: $\|f\|_1 = \sum_{j=1}^{p} |f_j|$ under the constraint $g = \Phi f$

Hardware prototype: the one-pixel camera (R. Baraniuk et al. @ Rice)



[FIG3] (a) Single-pixel, compressive sensing camera. (b) Conventional digital camera image of a soccer ball. (c) 64 × 64 black-and-white image $\hat{x}$ of the same ball ($N = 4,096$ pixels) recovered from $M = 1,600$ random measurements taken by the camera in (a). The images in (b) and (c) are not meant to be aligned.

(from R. Baraniuk, IEEE Signal Proc. Mag., July 2007)

Recent extensions

- Robustness in the presence of noise
- Object to sense f has a – or is well approximated by a – sparse expansion in a given basis or even on a coherent and redundant (overcomplete) dictionary (Candès et al., 2010)
- Decomposition of a large data matrix as $M = L_0 + S_0$ as a sum of a low-rank matrix L_0 and of a sparse matrix S_0: "Robust Principal Component Analysis" (Candès et al., 2009)
- Extension to the noisy case: "Stable Principal Component Pursuit" (Zhou et al., 2010)

Ordinary Least-Squares (OLS) Regression

- Noisy data: $y = X\beta + z$ (z = zero-mean Gaussian noise)
- Reformulate problem as a classical multivariate linear regression: minimize quadratic loss function

$$\Lambda(\beta) = \|y - X\beta\|_2^2 \qquad \left(\|y\|_2 = \sqrt{\sum_i |y_i|^2} = L_2\text{-norm}\right)$$

- Equivalently, solve variational (Euler) equation

$$X^T X\beta = X^T y$$

- If $X^T X$ is full-rank, minimizer is OLS solution

$$\beta_{ols} = (X^T X)^{-1} X^T y$$

Problems with OLS

- Not feasible if $X^T X$ is not full-rank i.e. has eigenvalue zero (in particular, whenever $p > n$)

  In many practical problems $p >> n$
  (large p, small n paradigm)
- Then the minimizer is not unique (system largely underdetermined), but you can restore uniqueness by selecting the "minimum-norm least-squares solution", orthogonal to the null-space of X
  (OK for prediction but not necessarily for identification!)
- Also $X^T X$ may have eigenvalues close to zero
  (happens when both p and n get large)
  $\longrightarrow X^T X$ has a large "condition number"
  (= ratio between largest and smallest e.v.)
  This is ill-conditioning, also referred to as the
  "curse of dimensionality"

A cure for the illness: Penalized regression

- To stabilize the solution (estimator), use extra constraints on the solution or, alternatively, add a penalty term to the least-squares loss
  $\rightarrow$ penalized least-squares

- This is a kind of "regularization"
  ($<$ inverse problem theory)

- Provides the necessary dimension reduction

- We will consider three examples: ridge, lasso and elastic-net regression

Ridge regression

(Hoerl and Kennard 1970 or Tikhonov's regularization)

- Penalize with L_2-norm of β:

$$\begin{aligned}\beta_{ridge} &= \operatorname{argmin}_\beta \left[\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right] \\ &= (X^T X + \lambda \, Id)^{-1} X^T y\end{aligned}$$

  ($\lambda > 0$ = "regularization parameter")

- Special case: orthonormal regressors ($X^T X = Id$)

$$\beta_{ridge} = \frac{1}{1+\lambda} X^T y$$

  (all coefficients are shrunk uniformly towards zero)

- Quadratic penalties provide solutions (estimators) which depend linearly on the response y but do not allow for variable selection (typically all coefficients are different from zero)

Lasso regression

name coined by Tibshirani 1996
but the idea is much older: Santosa and Symes 1986; Logan; Donoho, etc.

- Penalize with L_1-norm of β:

$$\beta_{lasso} = \mathrm{argmin}_\beta \left[\|y - X\beta\|_2^2 + \tau \|\beta\|_1 \right]$$

  where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$

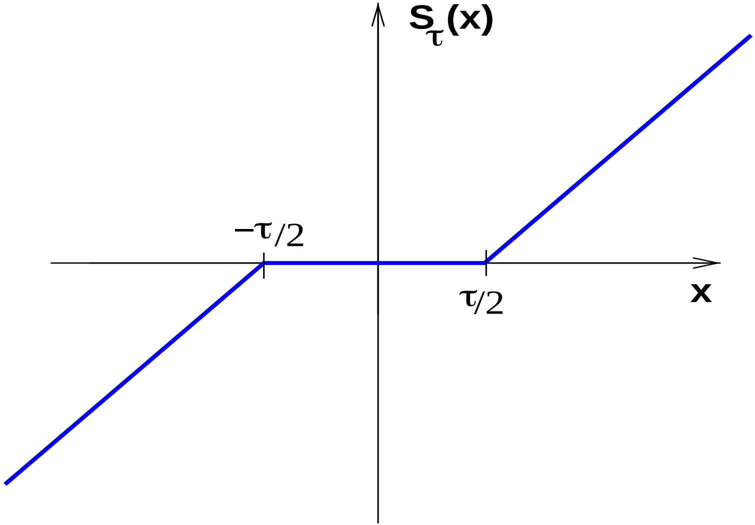
- Special case: orthonormal regressors ($X^T X = Id$)

$$[\beta_{lasso}]_j = S_\tau([X^T y]_j)$$

  S_τ is the soft-thresholder defined by

$$S_\tau(x) = \begin{cases} x + \tau/2 & \text{if} \quad x \le -\tau/2 \\ 0 & \text{if} \quad |x| < \tau/2 \\ x - \tau/2 & \text{if} \quad x \ge \tau/2 \end{cases}$$

Lasso regression: Soft-thresholding



Lasso regression

- Soft-thresholding is a nonlinear shrinkage: coefficients are shrunk differently depending on their magnitude.

  For orthonormal regressors, $[\beta_{lasso}]_j = 0$ if $|[X^T y]_j| < \tau/2$

- Enforces sparsity of β, i.e. the presence in this vector of many zero coefficients $\longrightarrow$

- Variable selection is performed!

Bayesian framework

- OLS can be viewed as maximum (log-)likelihood estimator for gaussian “noise”
  $\rightarrow$ penalized maximum likelihood
- Bayesian interpretation: MAP estimator and penalty interpreted as a prior distribution for the regression coefficients
- Ridge $\sim$ Gaussian prior
- Lasso $\sim$ Laplacian prior (double exponential)

Generalization

- Weighted L_α-penalties (weighted $\sim$ non i.i.d. priors)
  "bridge regression"

  (Frank and Friedman 1993; Fu 1998)

  Special cases: ridge ($\alpha = 2$) and lasso ($\alpha = 1$)

  NB. nonconvex for $\alpha < 1$

  Only $\alpha = 1$ allows for both sparsity and convexity

Lasso versus Model selection

- Limit case $\alpha = 0$: model selection with L_0-norm penalty

$$\|\beta\|_0 = \#\{\beta_j | \beta_j \neq 0\}$$

- $\alpha = 1$ is a good proxy for $\alpha = 0$

  Advantage: convex optimization instead of combinatorial algorithmic complexity!

- A lot of recent literature on the subject, e.g.

- "If the predictors are not highly correlated, then the lasso performs very well in prediction almost all the time" (probabilistic results) (Candès and Plan 2007)

Lasso regression: algorithmic aspects

- Quadratic programming (Tibshirani 1996; Chen, Donoho and Saunders 1998; Boyd and collaborators)

- Recursive strategy: LARS/Homotopy method (Efron, Hastie, Johnstone, Tibshirani 2004; Osborne, Presnell, Turlach 2000)

  Recursive way of solving the variational equations for $1, 2, ..., k$ active (non-zero) variables

  The regression coefficients are piecewise linear in τ
  $\rightarrow$ full path for the same computational cost

  Modification to take into account linear constraints (Brodie, Daubechies, De Mol, Giannone, Loris 2008)

Lasso regression: algorithmic aspects

- Iterative strategy: iterated soft-thresholding

$$\beta_{lasso}^{(l+1)} = \mathbf{S}_{\tau/C}\left(\beta_{lasso}^{(l)} + \frac{1}{C}[X^T y - X^T X \beta_{lasso}^{(l)}]\right)$$

  has been proved to converge to a minimizer of the lasso cost function with arbitrary initial guess $\beta_{lasso}^{(0)}$; provided $\|X^T X\| < C$ (compute norm e.g. by power method) ($\mathbf{S}_{\tau/C}$ performs soft-thresholding componentwise)

  (Daubechies, Defrise, De Mol 2004)

  NB. For $\tau = 0$: Landweber scheme converging to OLS (minimum-norm solution if $\beta_{lasso}^{(0)} = 0$)

- Many variations on this iterative scheme, and recent developments on accelerators see e.g. (Loris, Bertero, De Mol, Zanella and Zanni 2009)

Lasso regression: some applications

- Computer vision: selection of dictionary elements appropriate for a given classification task (e.g. face detection or face authentication) (Destrero, De Mol, Odone, Verri 2009)
- Assets for portfolio optimization in finance $\longrightarrow$ "Sparse and stable Markowitz portfolios" (Brodie, Daubechies, De Mol, Giannone, Loris 2009)
- Macroeconomic forecasting
  Standard paradigm for high-dimensional time series: Principal Component Regression
  Alternative: ridge or lasso regression
  (De Mol, Giannone, Reichlin 2008)

Nonparametric regression

- Nonlinear regression model : $y = f(X)$
  where the regression function f is assumed to have a sparse expansion on a given basis $\{\varphi_j\}$: $f = \sum_j \beta_j \varphi_j$
- Solve

$$\beta_{lasso} = \operatorname{argmin}_\beta \left[\|y - \sum_j \beta_j \varphi_j\|_2^2 + \tau \|\beta\|_1 \right]$$

- Vector β possibly infinite-dimensional (ℓ_1-penalty)
- cf. “basis pursuit denoising”
  (Chen, Donoho and Saunders 2001)

Inverse problems

(Daubechies, Defrise, De Mol 2004)

- Linear inverse problem $g = Af$, knowing the object has a sparse expansion on a given basis: $f = \sum_j \beta_j \varphi_j$
- Recover f by minimizing $\left[\|g - Af\|_2^2 + \tau\|\beta\|_1\right]$
- Infinite-dimensional framework where $\{\varphi_j\}$ = arbitrary orthonormal basis, as Fourier, wavelets, etc.
  (or even redundant "frame" or "dictionary")
- Typically, images (e.g. natural images) are sparse in the wavelet domain
- Proper "regularization method" for ill-posed inverse problems (as is Tikhonov for quadratic penalties)
- Strong convergence of iterated soft-thresholding
  (with soft-thresholding applied to the coefficient vector)

Extensions

- Mixed penalties/multiple components:

$$f = u + v + \dots$$

  where u is sparse (ℓ_1-penalty in some basis), v is smooth (ℓ_2 -penalty), etc.

  (Defrise and De Mol 2004; Daubechies and Teschke 2004; Anthoine 2005)

- Nonlinear inverse problems
  (through iterative soft-thresholding)

  (Teschke and Ramlau 2005)

Instability of Lasso for variable selection

- In learning theory (random design), the matrix X becomes also random
- In inverse problems, the imaging operator A may be subject to errors
- With random matrix, lasso regression does not provide a stable selection of variables if correlated $\longrightarrow$ possible remedy: “elastic net”

Elastic Net

- "Elastic net": combined penalties $L_1 + L_2$ to select sparse groups of correlated variables (Zou and Hastie 2005, for fixed-design regression, with n and p fixed).

$$\beta_{en} = \text{argmin}_\beta \left[\|y - X\beta\|_2^2 + \tau\|\beta\|_1 + \lambda\|\beta\|_2^2 \right]$$

  While the L_1-penalty enforces sparsity, the additional L_2-penalty takes care of possible correlations between the coefficients (enforces democracy in each group)

- NB. The groups are not known in advance ($\neq$ joint sparsity measures - mixed norms - group Lasso)

- Extension to learning (random design) and consistency results (De Mol, De Vito and Rosasco 2009)

Application to gene selection from microarray data

(De Mol, Mosci, Traskine and Verri 2009)

- Expression data for many genes and few examples (patients)
- Aim: prediction AND identification of the guilty genes
- Heavy correlations (small networks) $\to L_1 + L_2$ strategy
- Algorithm: damped iterated soft-thresholding

$$\beta_{en}^{(l+1)} = \frac{1}{1+\frac{\lambda}{C}} \mathbf{S}_{\tau/C}\left(\beta_{en}^{(l)} + \frac{1}{C}[X^T y - X^T X \beta_{en}^{(l)}]\right)$$

(contraction for $\lambda > 0$)