The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

IAEA

2585-31

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

Sparsity for big data contd.

C. De Mol
ULB, Brussels
Belgium

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

Sparsity for Big Data
Part II - High-dimensional Regression

Christine De Mol

Université Libre de Bruxelles
Dept Math. and ECARES

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications
June 2014

Linear regression problem

- "Input" (data) matrix: $X = \{x_{ij}\}$
  for $i = 1, \ldots, n$
  and $j = 1, \ldots, p$

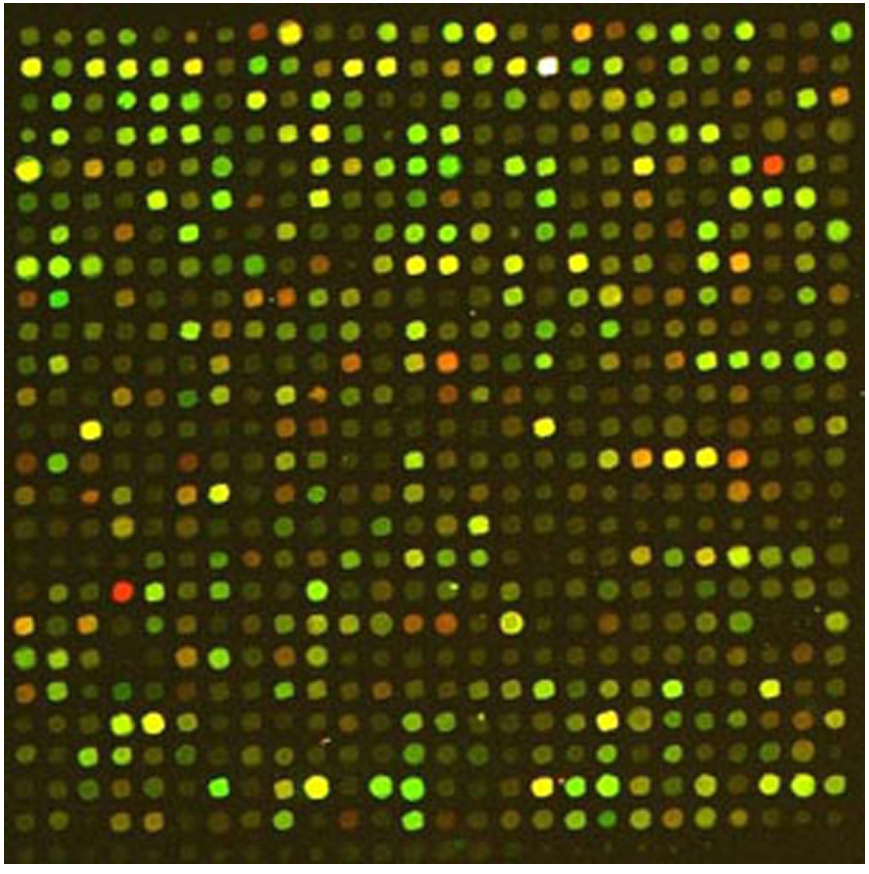
- "Output" (response): y_i for each i ("supervised" setting)

- Assume linear dependence: $y_i = \sum_j x_{ij}\beta_j$ or

$$y = X\beta$$

where $y = (y_1, y_2, \ldots, y_n)^T$

Example: Microarray data analysis



Example: Microarray data analysis

- Input: $X = \{x_{ij}\}$ = gene expression levels

  for p genes $j = 1, \ldots, p$

  and n patients or experiments $i = 1, \ldots, n$

- Output: y_i is a discrete label in classification problems (e.g. disease/healthy or type of illness)

  or a continuous index in true regression problems (e.g. survival time or gravity of illness)

Example: Forecasting time series

- Input: $X^T = \{x_{jt}\}$ = time series panel
  (e.g. macroeconomic or financial data)

  n = number of samples in time

  p = number of series $\times$ the number of lags used

- Output: y_t = series to be forecast on the basis of X

Two distinct problems

- Prediction ("generalization")
  predict (forecast) the response y

- Identification (Variable Selection)
  find the regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$
  i.e. identify the relevant predictors

  or SELECT them when many coefficients are zero
  i.e. when β is SPARSE

  Essential for interpretation!

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

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

- 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
  "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
- Increases bias to decrease variance
- 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:

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

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

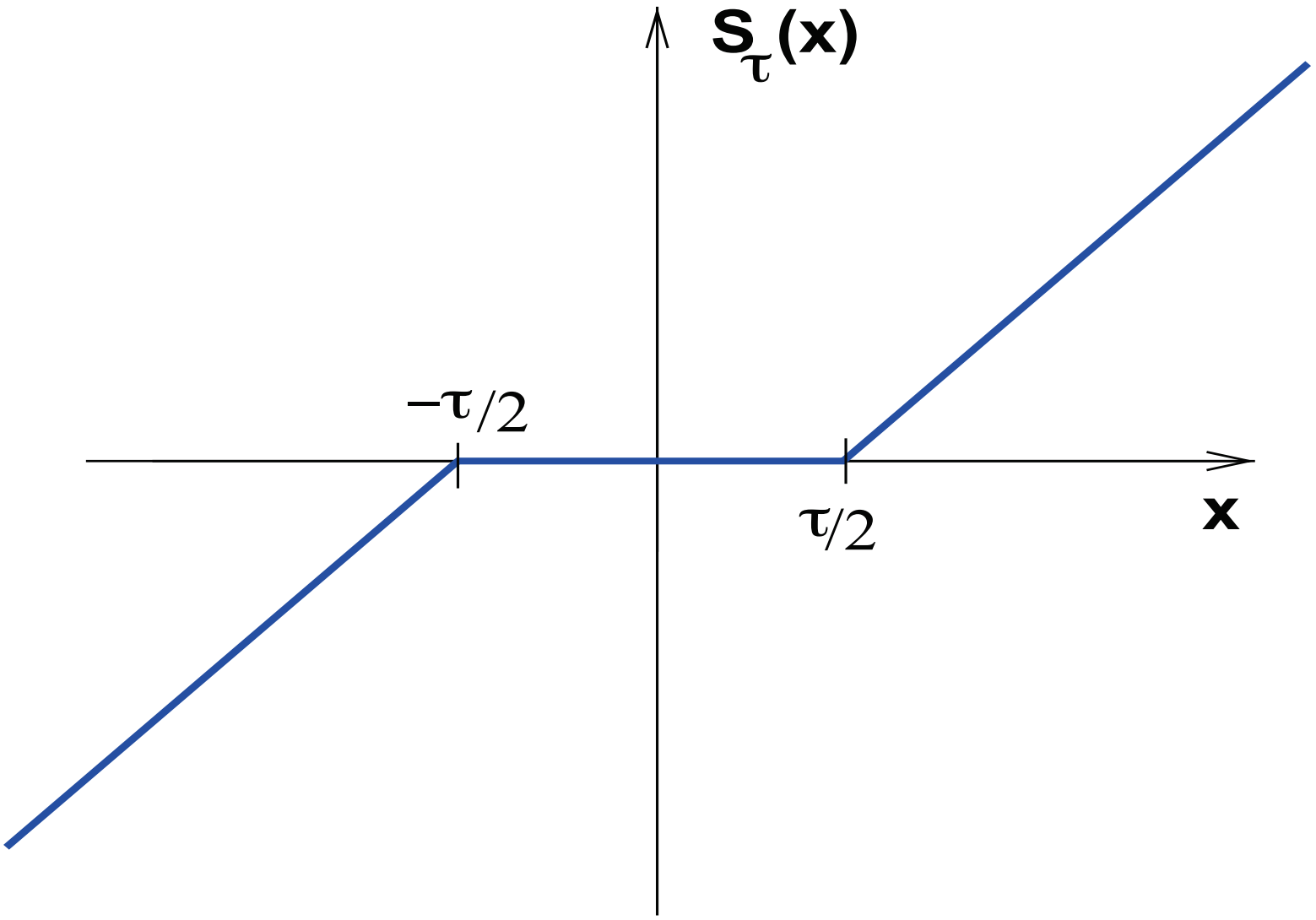
- Special case: orthonormal regressors ($X^T X = Id$)

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

S_τ is the soft-thresholder defined by

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

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 $|[X^T y]_j| < \tau/2$

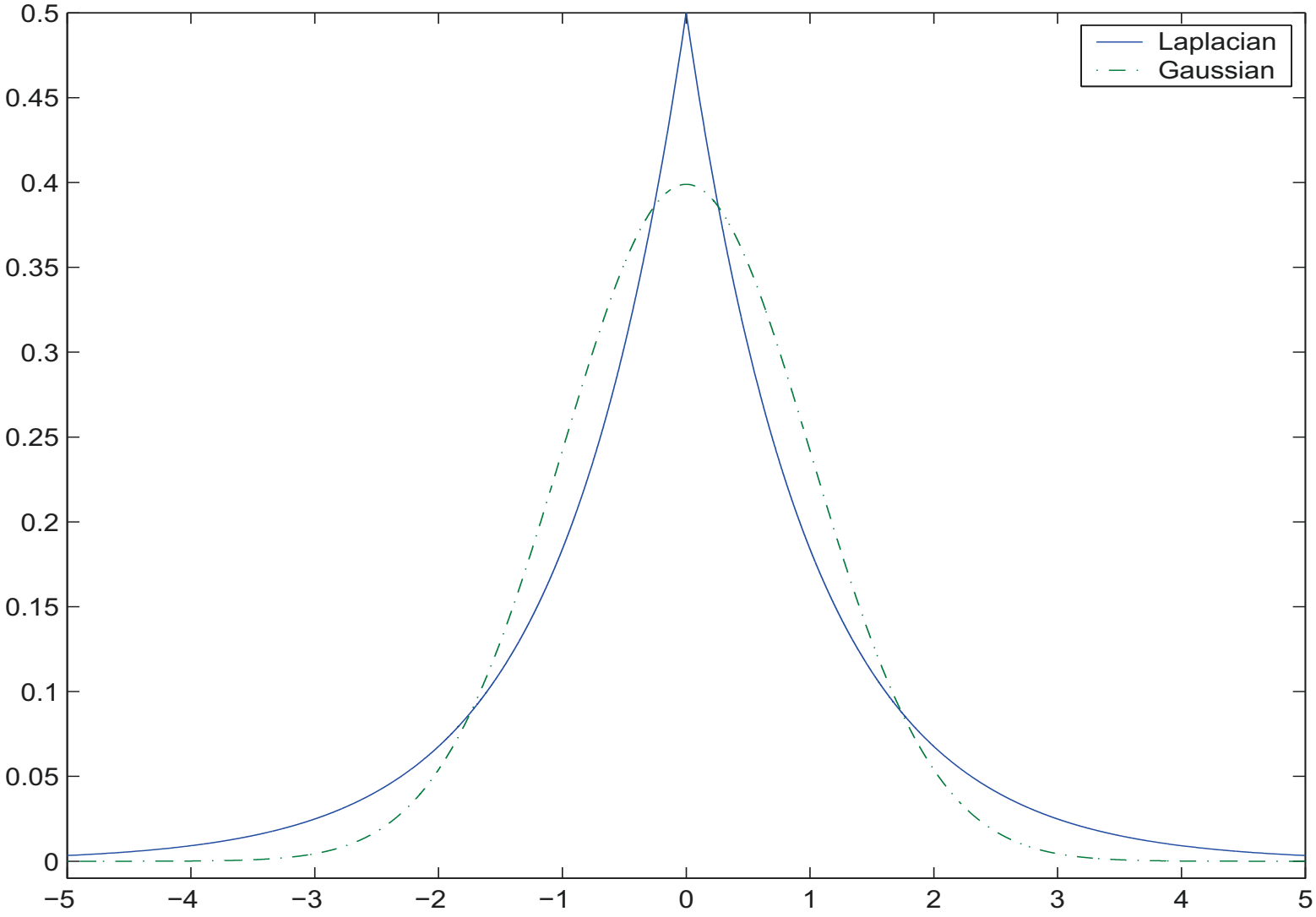
- Enforces sparsity of β, i.e. the presence in this vector of many zero coefficients $\longrightarrow$

- Variable selection is performed!

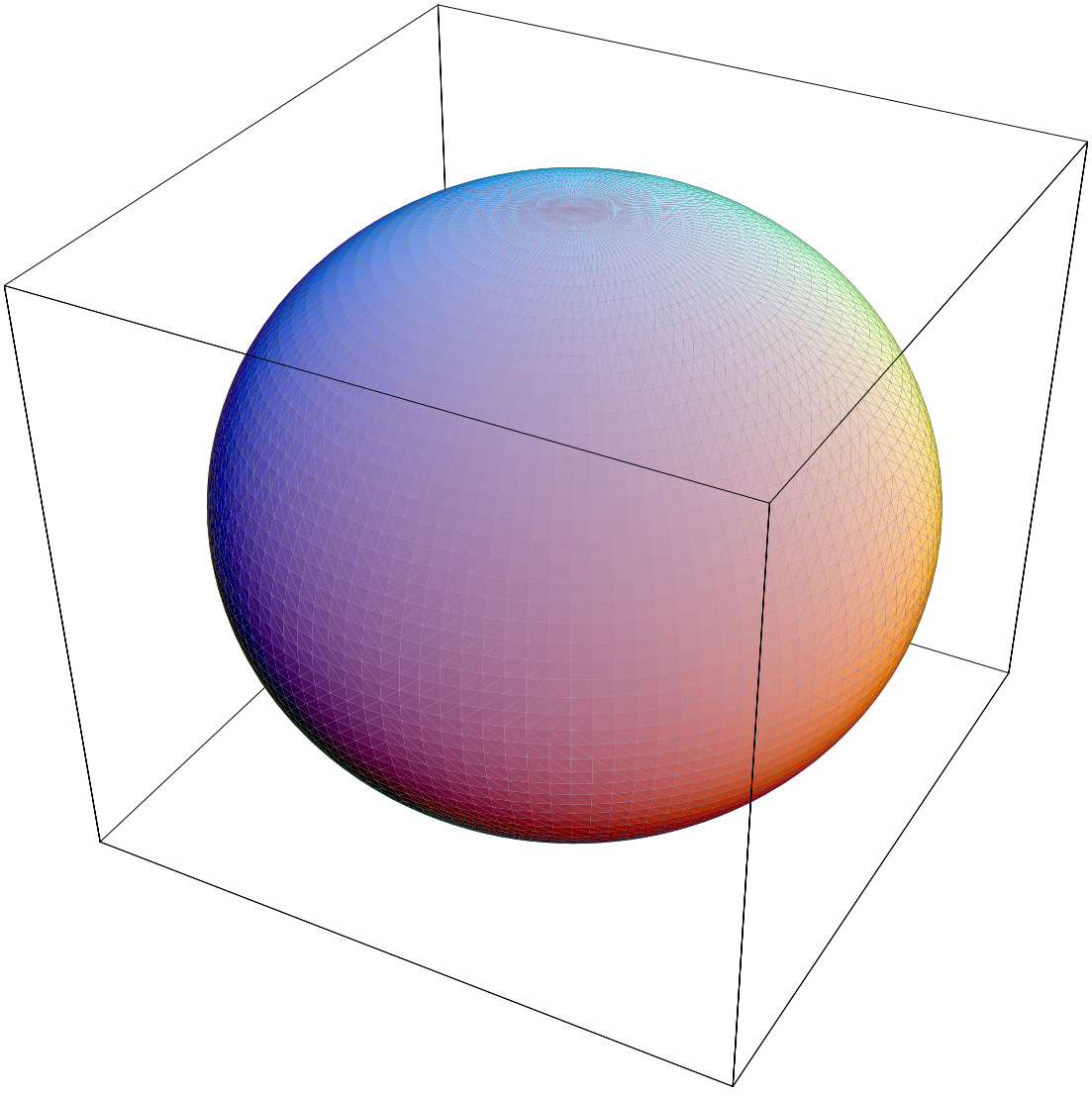
Bayesian framework

- 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)

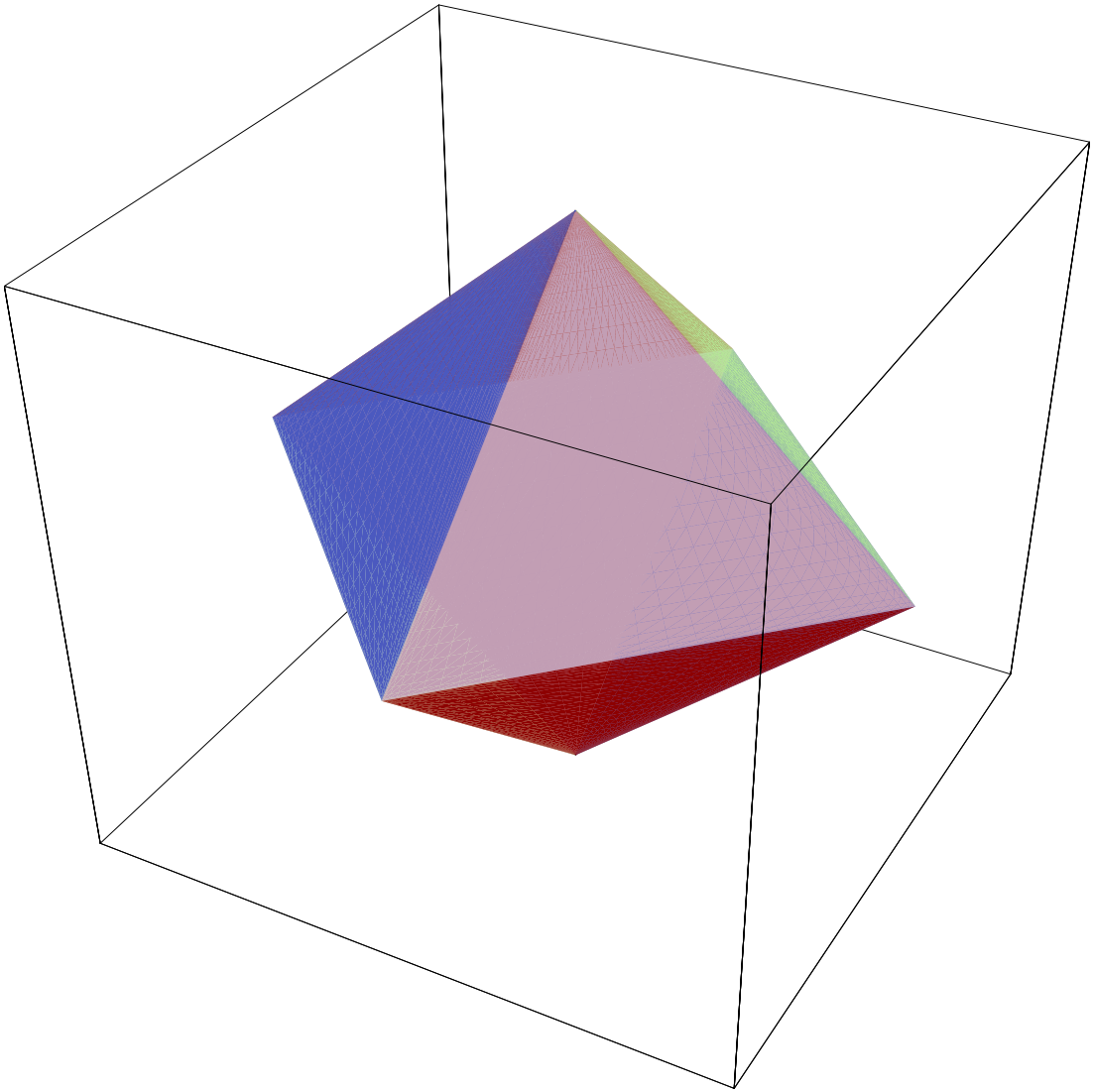
Gauss versus Laplace



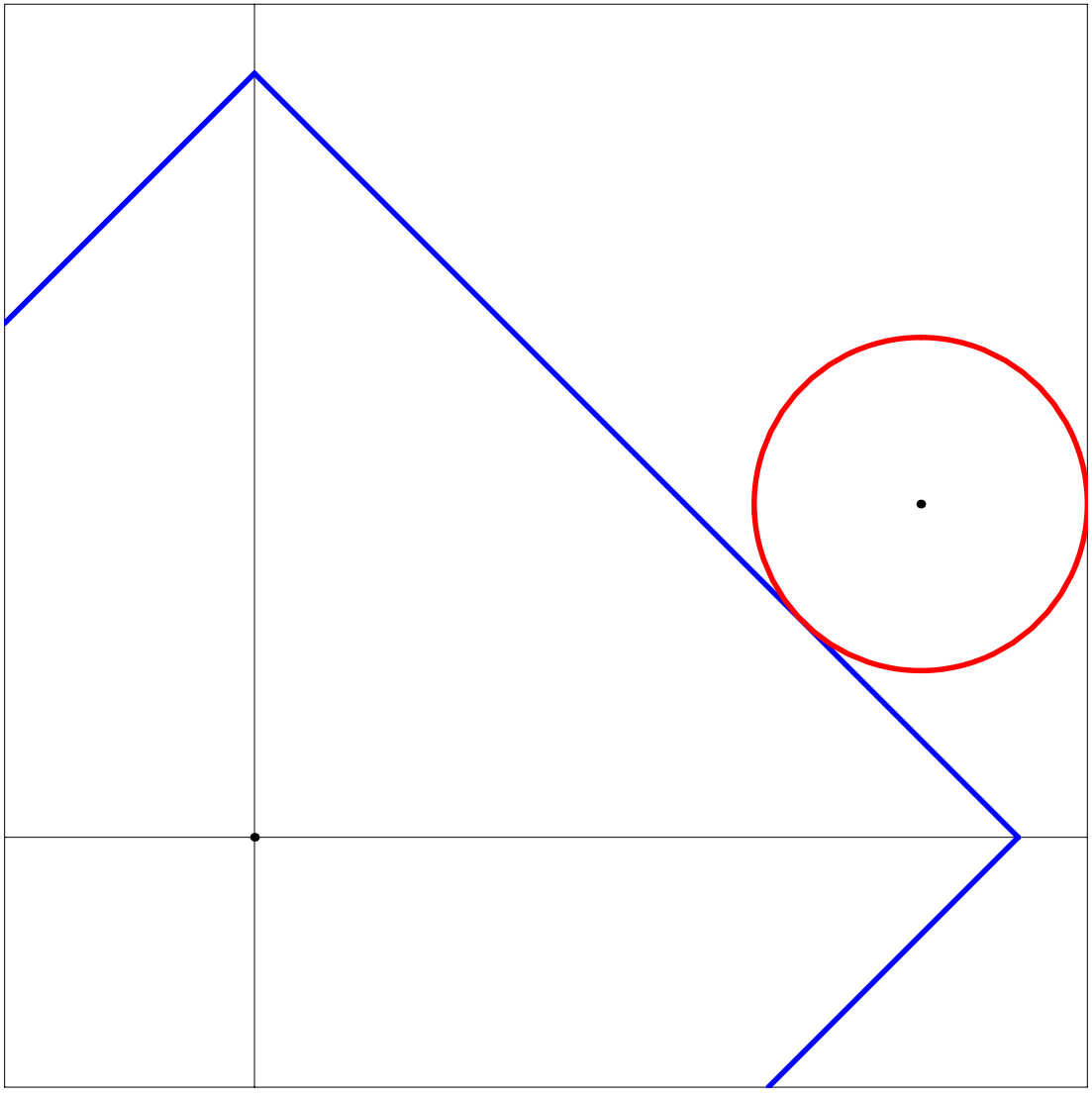
L_2 ball



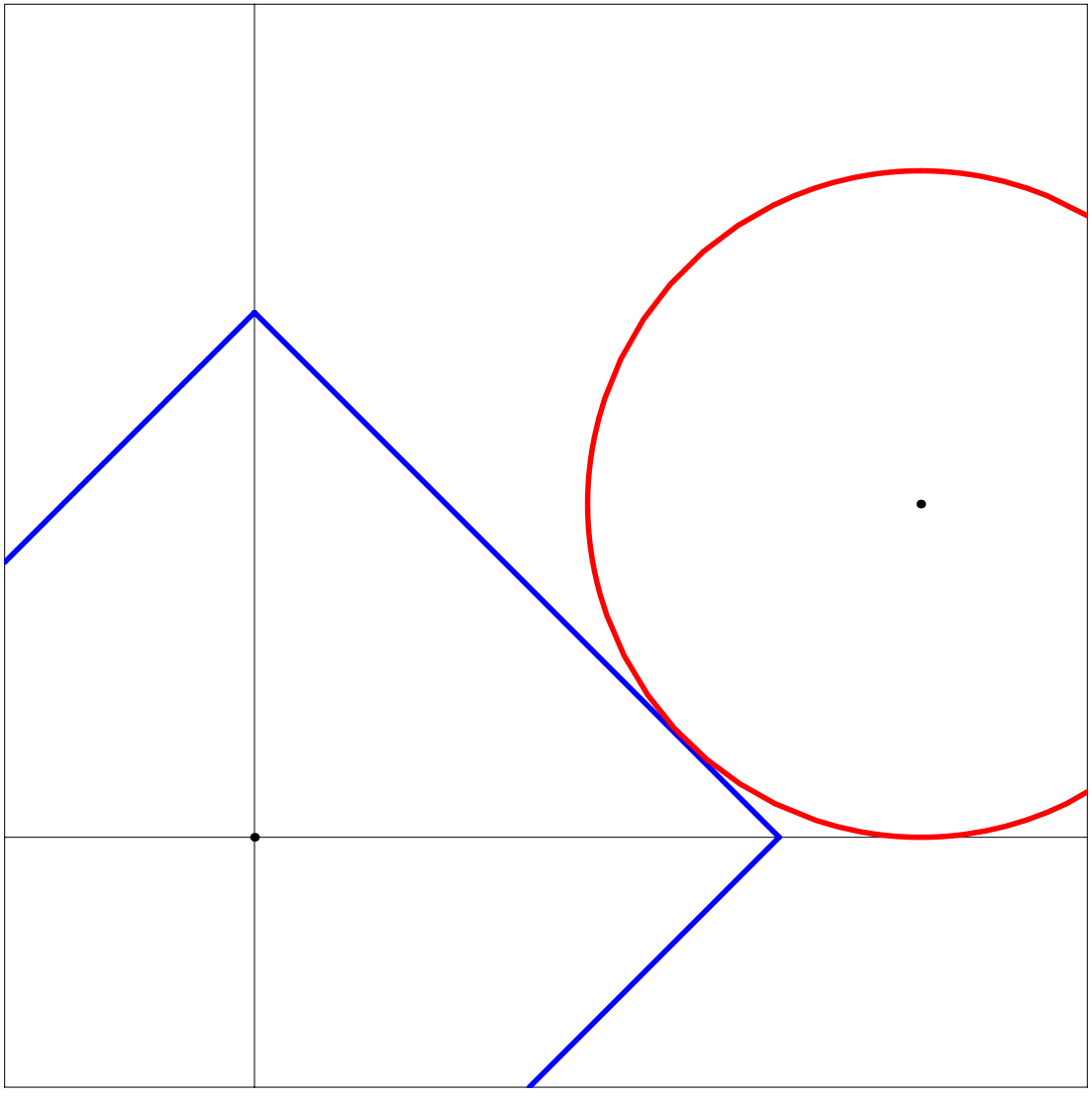
L_1 ball



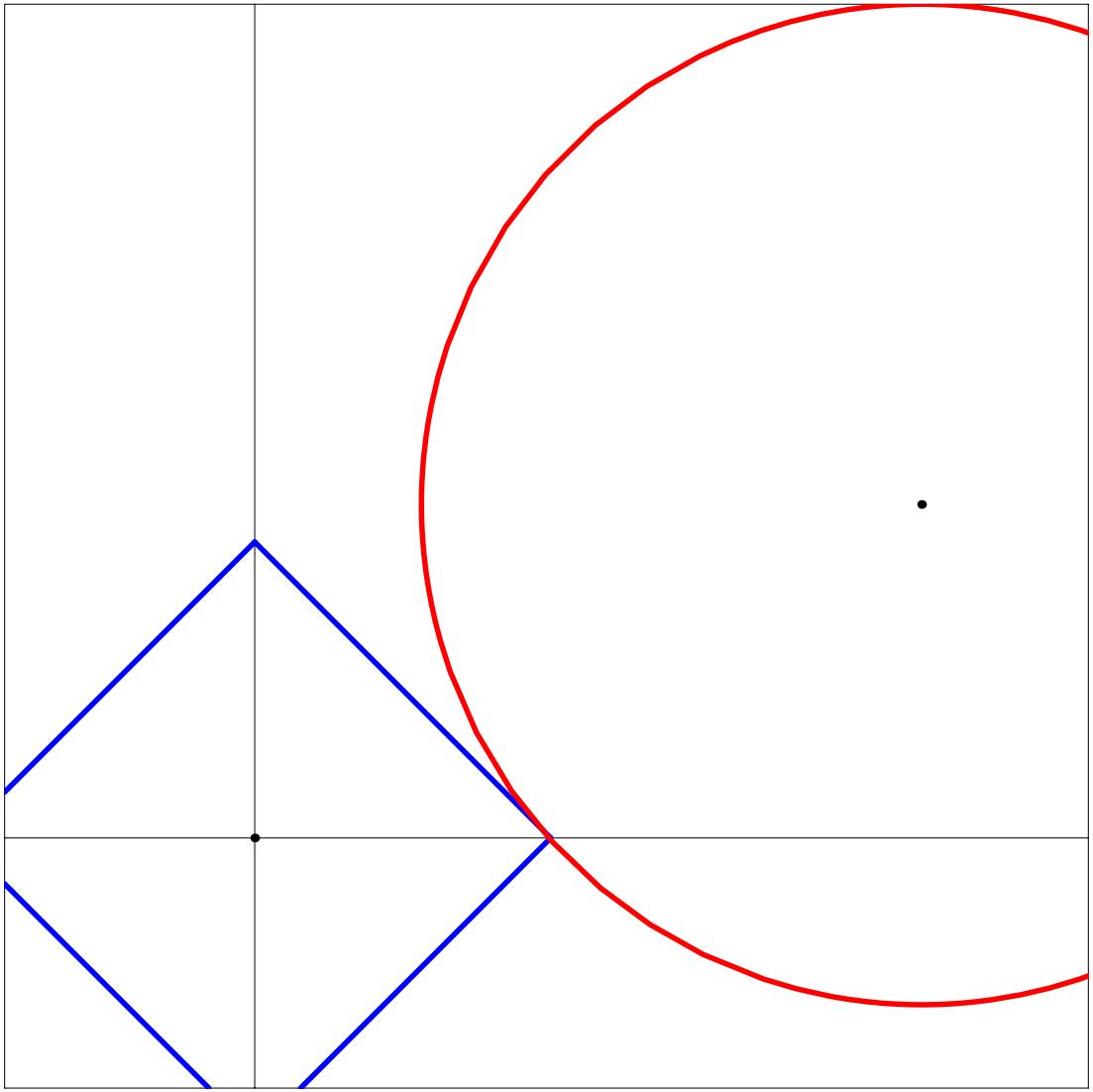
Lasso regression and sparsity



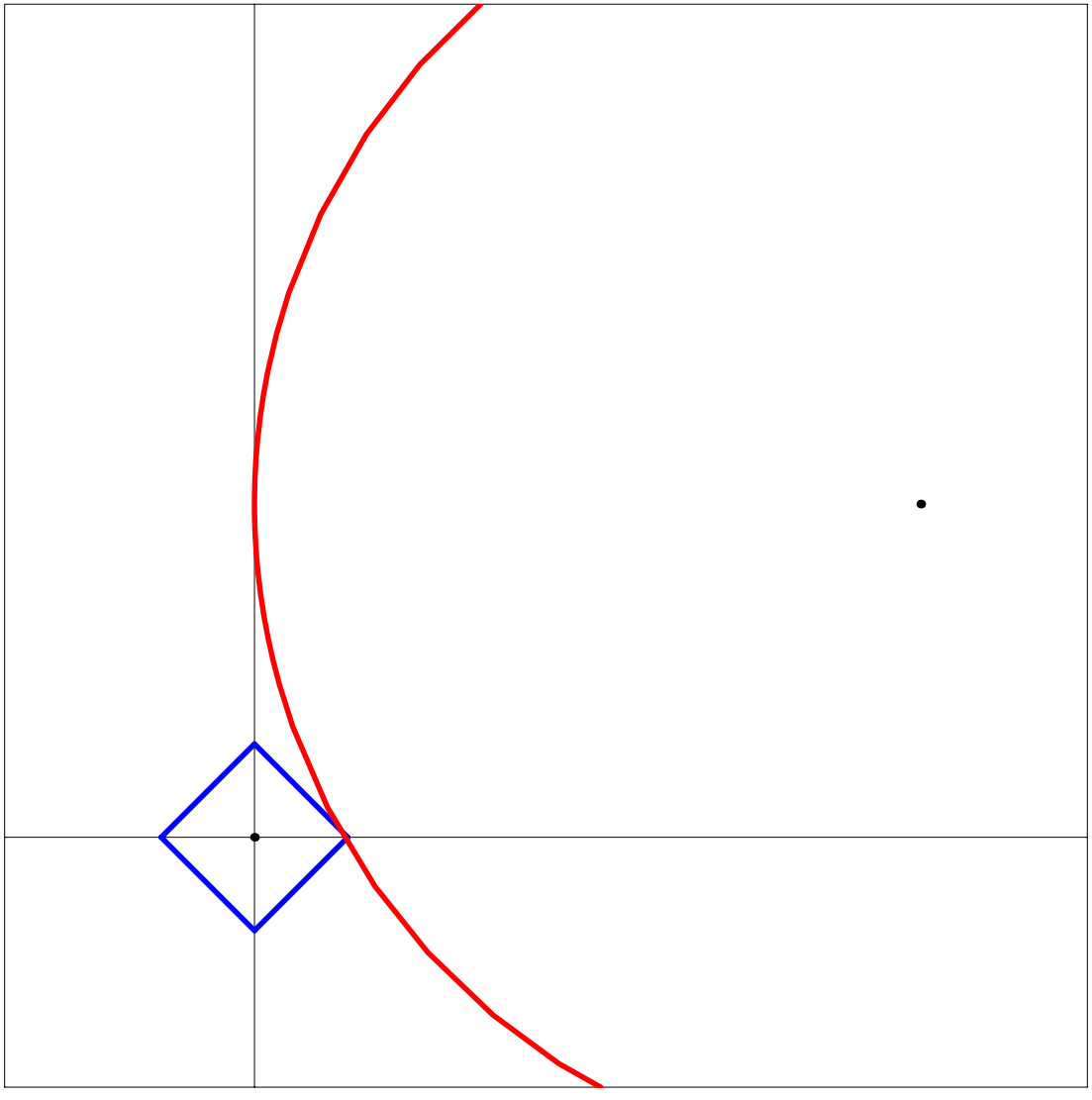
Lasso regression and sparsity



Lasso regression and sparsity



Lasso regression and sparsity



Generalization

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

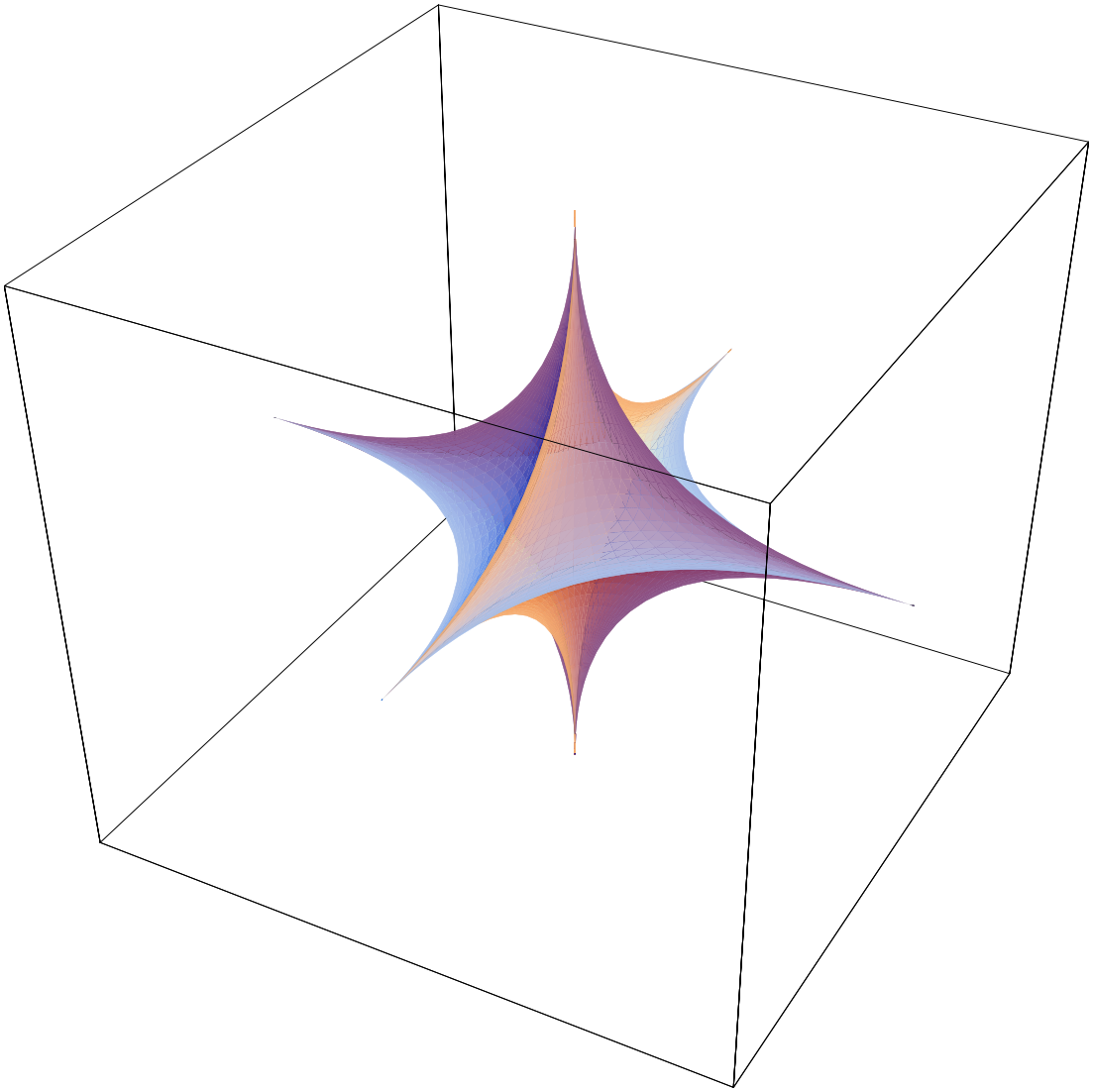
  (Frank and Friedman 1993; Fu 1998)

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

  NB. nonconvex for $\alpha < 1$

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

$L_{1/2}$ ball



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)

Macroeconomic forecasting

(De Mol, Giannone, Reichlin 2008)

- For high-dimensional time series, the standard paradigm is Principal Component Regression (Stock and Watson 2002 for static PC, Forni, Hallin, Lippi, Reichlin 2000 for dynamic PC)

$$\beta_{pcr} = \sum_{k=1}^{r} \frac{\langle X^T y, v_k \rangle}{\xi_k^2} v_k$$

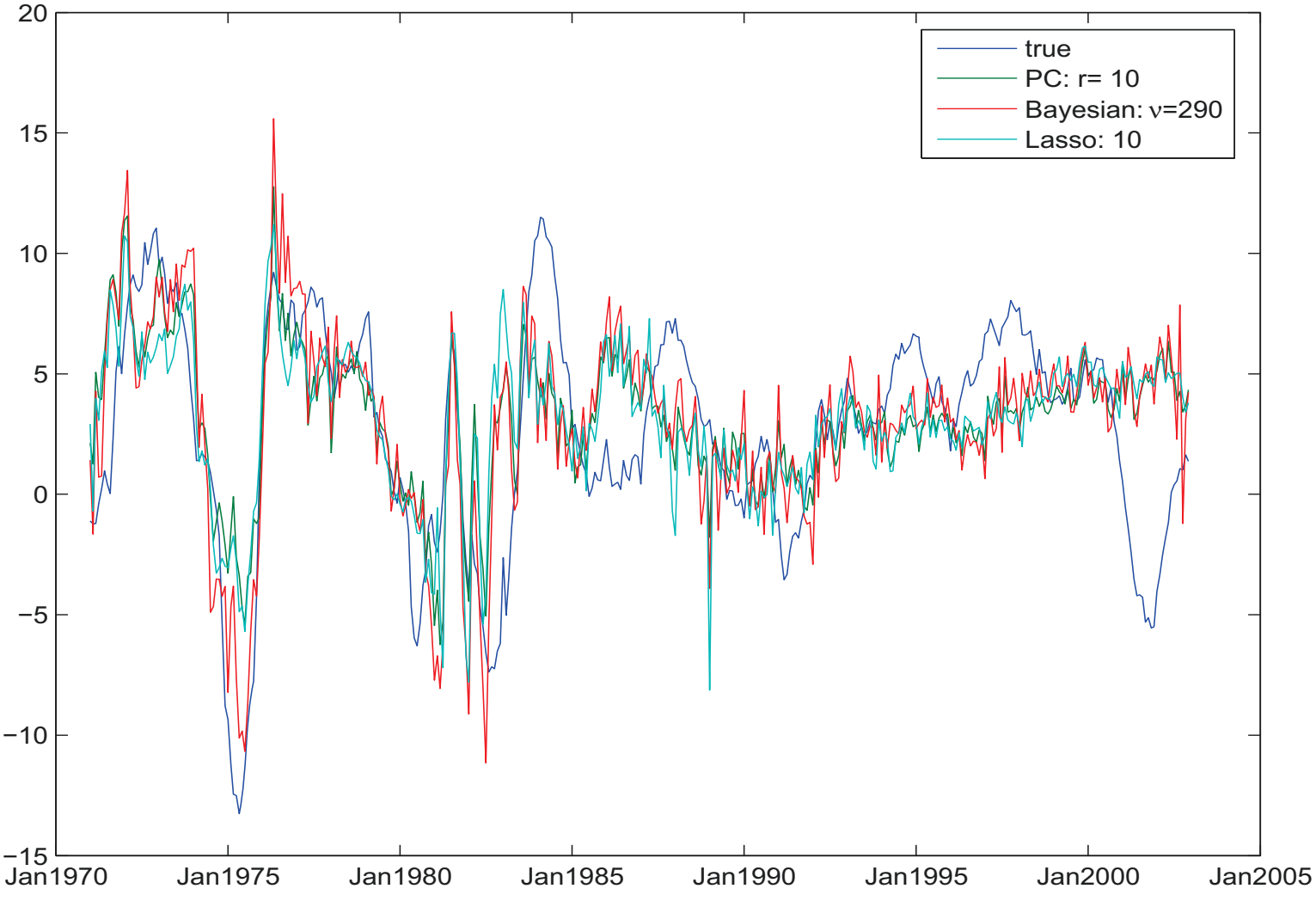
  where v_k are the eigenvectors of $X^T X$ with eigenvalues ξ_k^2. "Truncated SVD", at r before the rank ($\rightarrow$ dimension reduction)

- Alternative: Penalized regression (ridge, lasso, etc.)

Macroeconomic forecasting: empirical results

- Macroeconomic data-set of 131 monthly time series for the US economy from Jan59 to Dec03 (Stock and Watson 2005), transformed for stationarity and standardized
- Variable to forecast:
  1. Industrial Production: $y_{t+h} = (\log IP_{t+h} - \log IP_t) \times 100$
  2. Price inflation: $y_{t+h} = \pi_{t+h} - \pi_t$
- Simulated out-of-sample exercise:
  For each time $T = \mathrm{Jan}70, ..., \mathrm{Dec}01$, estimate β using the most recent 10 years of data (rolling scheme),
  with a forecast horizon of $h = 12$ months
  (No lags of the regressors included here; similar results when including lags)

Forecasting IP



Macroeconomic forecasting: theoretical results

- Consistency results in capturing the common part in an approximate factor structure, asymptotically as the number of series and time samples tend both to infinity

- Consistency is achieved along any path
  $\rightarrow$ suitable for large cross-section
  (even when the number of series is larger that the number of time samples)

(De Mol, Giannone, Reichlin 2008)

Portfolio optimization

- p securities with returns $r_{j\,t}$ at time t
- Expected returns : $\mu_j = \mathbf{E}[r_{j\,t}]$
  Covariance matrix $C = (C_{j\,k})$ of the returns:

$$C_{j\,k} = \mathbf{E}[(r_{j\,t} - \mu_j)(r_{k\,t} - \mu_k)]$$

- A portfolio is defined by a $p \times 1$ vector of weights β_j summing to one (unit of capital)
- Expected return of the portfolio: $\sum_j \beta_j\, \mu_j = \beta^T \mu$
  Variance of the portfolio: $\beta^T C \beta$

Sparse and stable Markowitz portfolios

(Brodie, Daubechies, De Mol, Giannone, Loris 2008)

- Markowitz portfolios: Find a portfolio β^* which has minimal variance for a given expected return ρ
- This is equivalent to the regression problem

$$\beta^* = \arg\min_{\beta} \mathbf{E}\left[|\rho - \sum_j \beta_j \, r_{j\,t}|^2\right]$$
$$\text{s. t.} \quad \sum_j \beta_j \, \mu_j = \rho \quad \text{and} \quad \sum_j \beta_j = 1$$

- For empirical implementation, replace expectations with sample averages and solve the following regression problem

Sparse and stable Markowitz portfolios

$$\begin{aligned} \beta^* \quad &= \quad \arg\min_{\beta} \left[\|\rho 1_n - X\beta\|_2^2 + \tau \|\beta\|_1 \right] \\ &\text{s. t.} \quad \sum_j \beta_j \, \hat{\mu}_j = \rho \quad \text{and} \quad \sum_j \beta_j = 1 \end{aligned}$$

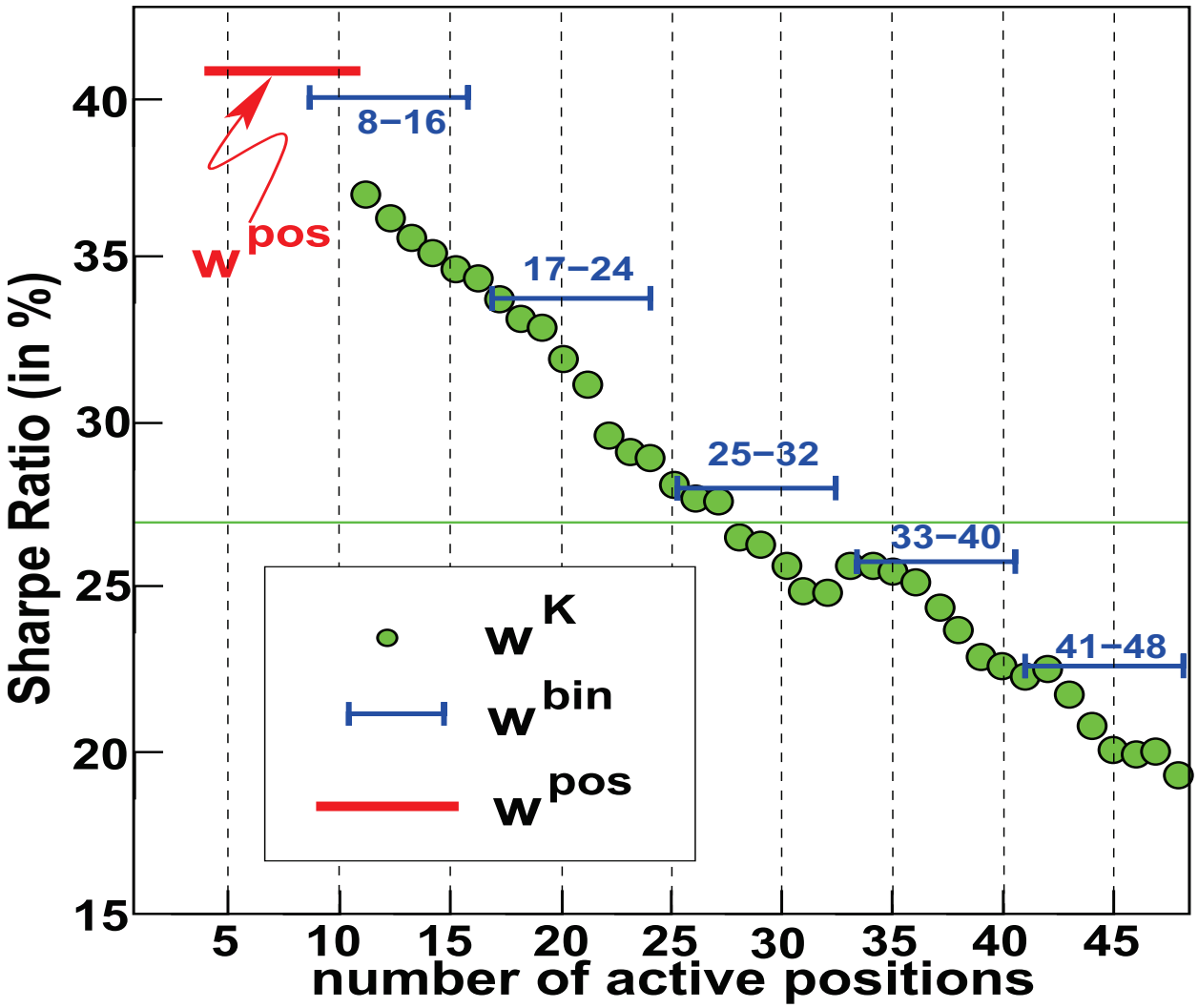
where 1_n is a $n \times 1$ vector of ones, $\hat{\mu}_j = \frac{1}{n} \sum_{t=1}^{n} r_{j\,t}$ and X is the $n \times p$ matrix of the sample returns

- The L_1-penalty ensures for sparsity and stability and accounts for transaction and monitoring costs ($\neq$ Markowitz)
- We devised a modification of LARS able to enforce the linear constraints; varying τ allows to tune the number of selected assets
- Special case: (sparse!) no-short portfolios (only positive weights)

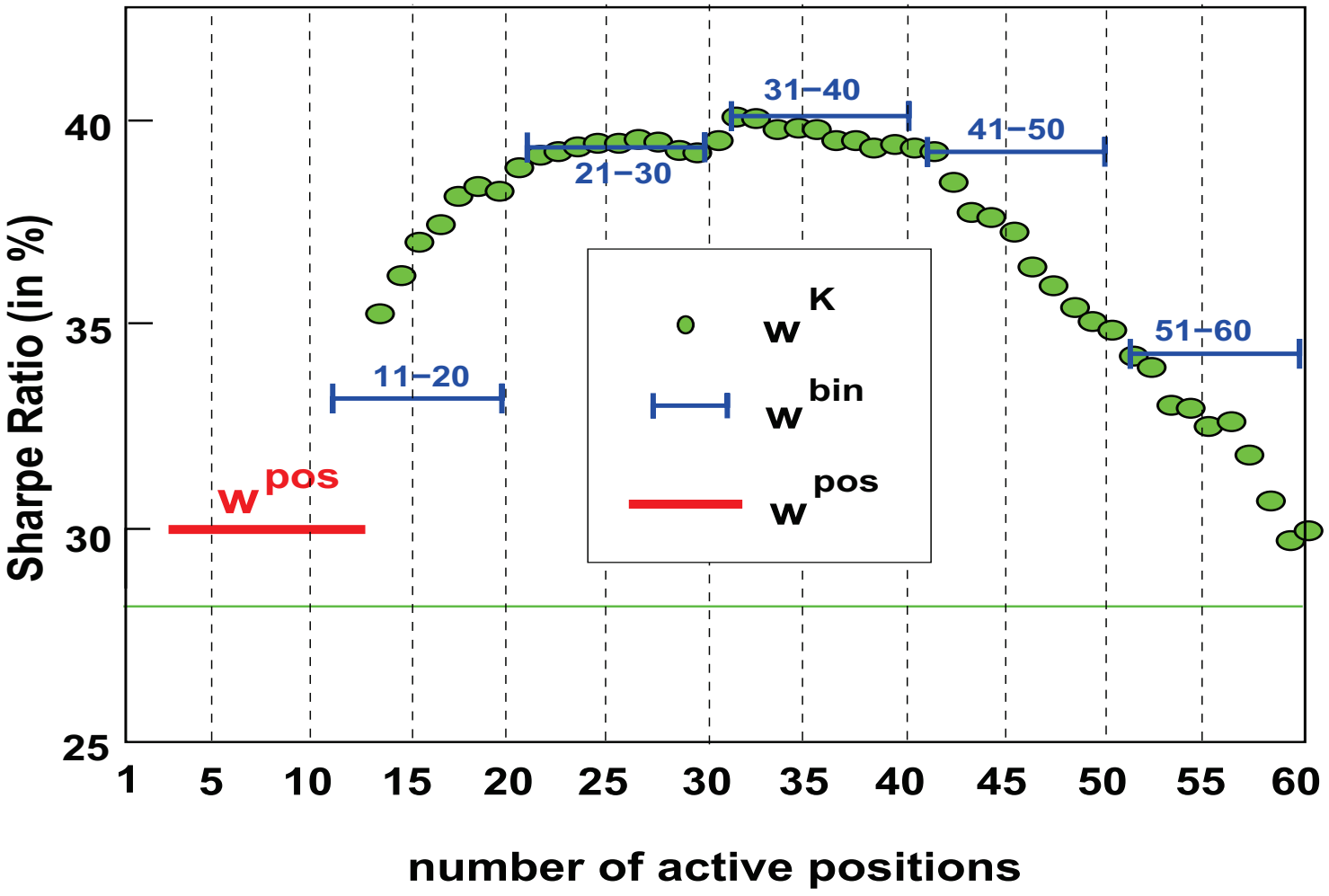
Empirical application

- We used as assets the Fama and French 48 industry portfolios (FF48) and 100 portfolios formed on size and book-to-market (FF100)
- We constructed our portfolios in June of each year from 1976 to 2006 using 5 years of historical (monthly) returns and a target return equal to the historical return of the equally-weighted portfolio
- Performance is evaluated by out-of-sample monthly mean return m, standard deviation σ and Sharpe ratio $S = m/\sigma$
- Benchmark (tough!) is the equal-weight portfolio, known to outperform many constructions (DeMiguel, Garlappi and Uppal 2007)

Empirical results FF48



Empirical results FF100



Optimal Forecast Combination

(Conflitti, De Mol, Giannone, 2012)

- Increase forecast accuracy by linearly combining individual forecasts (provided by different forecasters or models), using positive weights summing to one
- Minimizing the variance leads to an equivalent of Markowitz no-short portfolios
- Extension to the combination of density forecasts with a Kullback-Leibler based cost function
- Empirical application to the combination of ECB Survey Forecasts (ECB uses equal-weight combinations)

Feature Selection in Computer Vision

(Destrero, De Mol, Odone and Verri 2009)

- Huge dictionary $\{\varphi_j\}$ of 64000 “rectangle features” used in each image patch (19 x 19 pixels), i.e. scaled and translated versions of

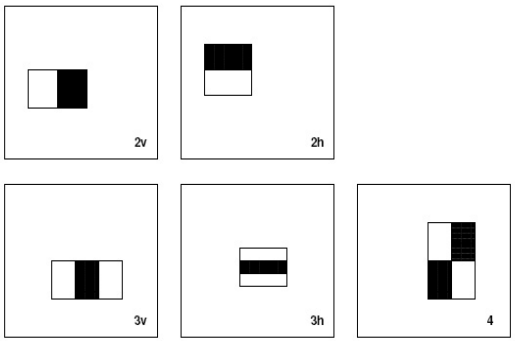


Figure 1: The support of the 5 types of rectangular features. The value of each feature is obtained by subtracting the sum of pixels in white areas from the sum of pixels in dark areas.

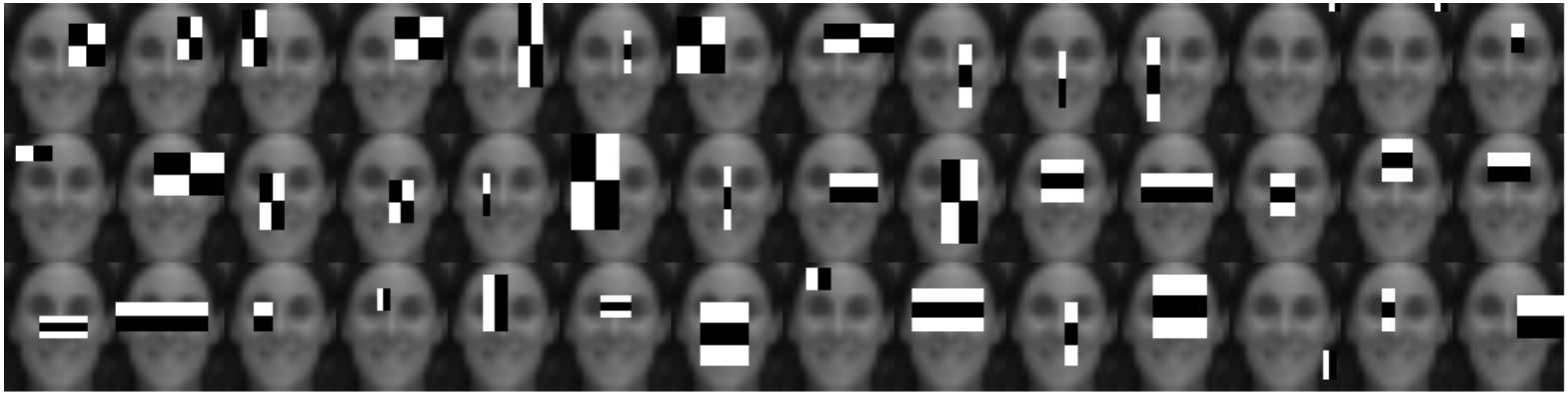
- Each row of the matrix X is filled with the scalar products with these rectangle features

Object (Face) detection

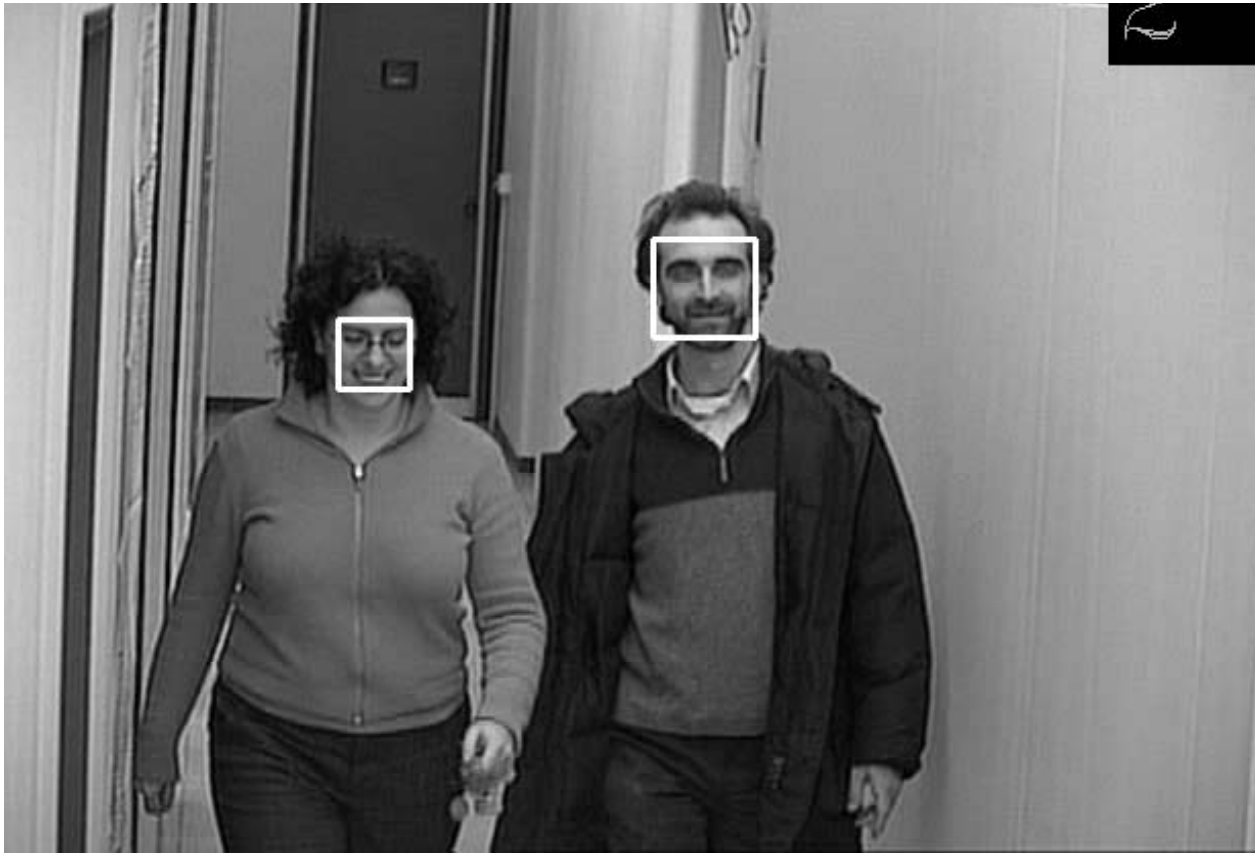
- Training set of 2000 positive and 2000 negative examples (binary classification response $y = [-1, +1]$)
- Cascade of lasso-type regression by randomized blocks (followed by a correlation analysis to eliminate redundancy) allows to select 42 relevant features
- Using a cascade of SVM filters on these features, we challenge a state-of-the-art Adaboost scheme (Viola and Jones 2004)
- Extension to face authentication

Object (Face) detection

The 42 selected features



Object (Face) detection



Object (Face) detection

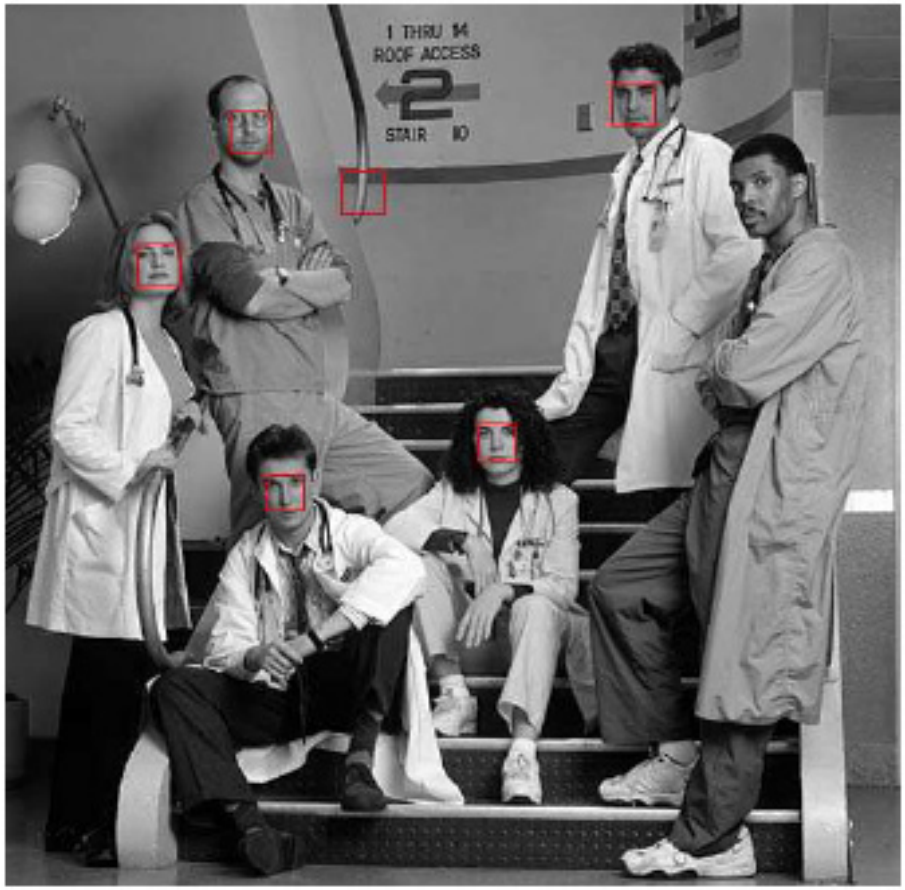


Figure 18: Face detection results on example images from the CMU dataset

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)

Instability of Lasso for variable selection

- In learning theory (random design), the matrix X becomes also random
- With a random matrix, lasso regression does not provide a stable selection of variables when they are 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).

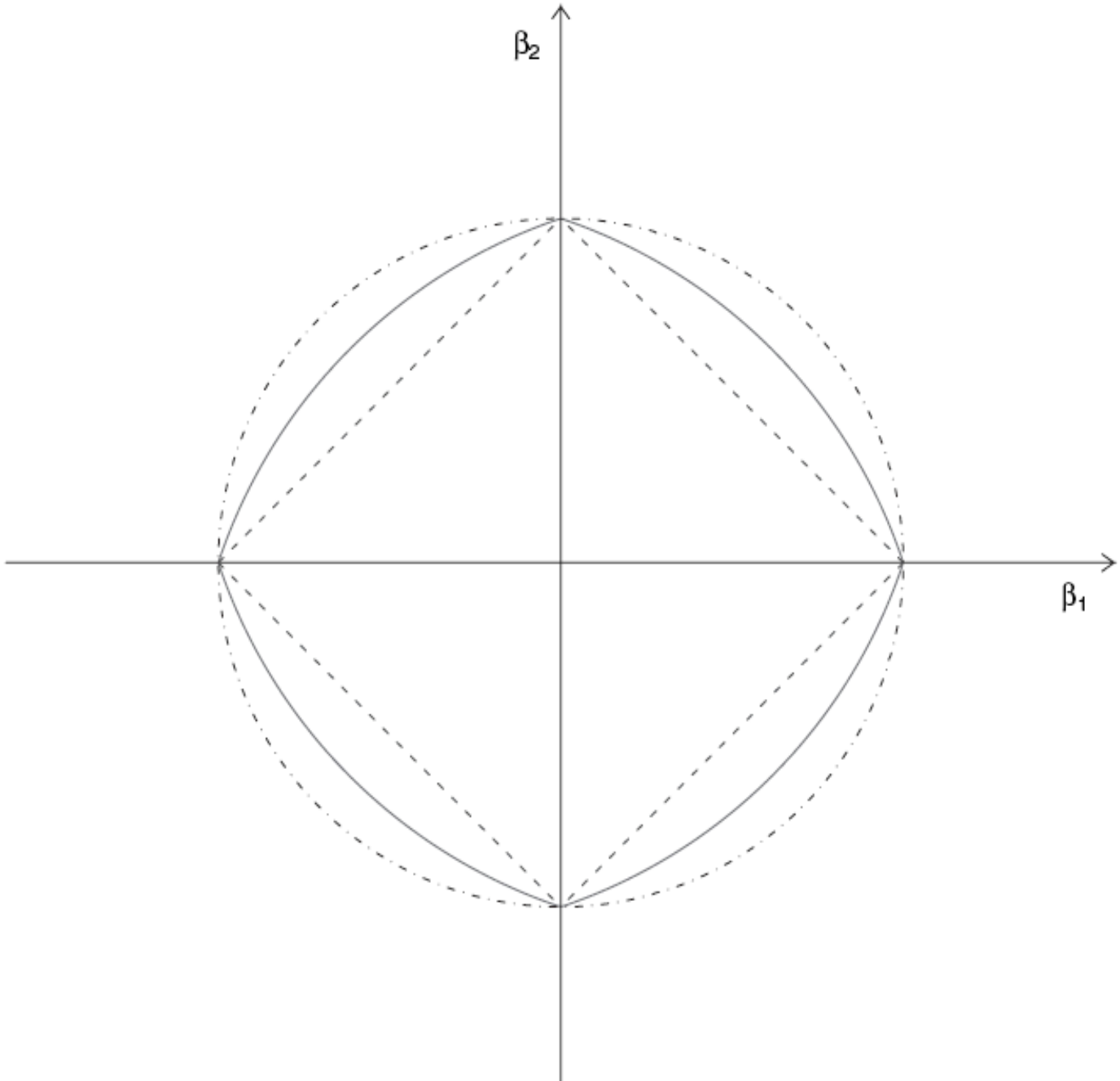
$$\beta_{en} = \text{argmin}_\beta \left[\|y - X\beta\|_2^2 + \tau\|\beta\|_1 + \lambda\|\beta\|_2^2 \right]$$

  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)

$L_1 + L_2$ ball (NB. Corners; $\neq L_\alpha$ ball for $\alpha > 1$)



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)
  $\rightarrow L_1 + L_2$ 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 $\lambda > 0$)

Full papers: Regularization and Algorithms

- An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint

  Ingrid Daubechies, Michel Defrise and Christine De Mol

  Comm. on Pure and Applied Math. 57 (2004): 1413-57
  http://arxiv.org/abs/math/0307152

- Accelerating gradient projection methods for l1-constrained signal recovery by steplength selection rules

  Ignace Loris, Mario Bertero, Christine De Mol, Riccardo Zanella and Luca Zanni

  Applied Computational and Harmonic Analysis 27 (2009) : pp. 247-254;
  http://arxiv.org/abs/0902.4424

Full papers: Economics

- Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components?

  Christine De Mol, Domenico Giannone and Lucrezia Reichlin

  Journal of Econometrics 146 (2008) : pp. 318-328
  ECB Working paper 700

- Sparse and stable Markowitz portfolios

  Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone and Ignace Loris

  Proc. Natl Acad. Sci. USA 106 (2009): pp. 12267-12272; http://arxiv.org/abs/0708.0046

Full papers: Computer Vision

- A sparsity-enforcing method for learning face features

  Augusto Destrero, Christine De Mol, Francesca Odone and Alessandro Verri

  IEEE Transactions on Image Processing 18 (2009) : pp. 188-201

- A Regularized Framework for Feature Selection in Face Detection and Authentication

  Augusto Destrero, Christine De Mol, Francesca Odone and Alessandro Verri

  International Journal of Computer Vision 83 (2009): pp. 164-177

Full papers: Elastic net

- A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data

  Christine De Mol, Sofia Mosci, Magali Traskine and Alessandro Verri

  Journal of Computational Biology 16 (2009) : pp. 677-690; http://arxiv.org/abs/0809.1777

- Elastic-Net Regularization in Learning Theory

  Christine De Mol, Ernesto De Vito and Lorenzo Rosasco

  Journal of Complexity 25 (2009) : pp. 201-230; http://arxiv.org/abs/0807.3423

Appendix:
A Simple Iterative Algorithm for Lasso Regression

Contraction Mapping Principle

- Definitions:
  - (i) T is non-expansive if $\|T(\beta) - T(\alpha)\| \leq \|\beta - \alpha\|$
  - (ii) T is a contraction if $\|T(\beta) - T(\alpha)\| \leq \rho\|\beta - \alpha\|$ with $\rho < 1$
  - (iii) β is a fixed point of T if $\beta = T(\beta)$

(T is a matrix or an operator in a Hilbert or Banach space – possibly nonlinear)

- Properties (proof: exercise)
  - (i) Any contraction is continuous
  - (ii) The fixed point of a contraction is necessarily unique
  - (iii) The product of a contraction and of a non-expansive operator is a contraction

Contraction Mapping Principle

- Contraction Mapping Principle

If T is a contraction, the sequence of iterates $\beta^{(l+1)} = T(\beta^{(l)})$, $l = 0, 1, 2, \ldots$, for any starting point $\beta^{(0)}$, is convergent (in norm) and its limit is the unique fixed point of T.

Proof: We have

$$\|\beta^{(l+1)} - \beta^{(l)}\| \leq \rho \|\beta^{(l)} - \beta^{(l-1)}\| \leq \cdots \leq \rho^l \|\beta^{(1)} - \beta^{(0)}\|$$

and by the triangle inequality, for any $m > 1$,

$$\begin{aligned} \|\beta^{(l+m)} - \beta^{(l)}\| &\leq \|\beta^{(l+m)} - \beta^{(l+m-1)}\| + \cdots + \|\beta^{(l+1)} - \beta^{(l)}\| \\ &\leq (\rho^{l+m-1} + \rho^{l+m-2} + \cdots + \rho^l)\|\beta^{(1)} - \beta^{(0)}\| \\ &\leq \frac{\rho^l}{1-\rho}\|\beta^{(1)} - \beta^{(0)}\| \end{aligned}$$

Contraction Mapping Principle

Since $\rho < 1$, this implies $\|\beta^{(l+m)} - \beta^{(l)}\| \to 0$ for $l \to \infty$, i.e. $\beta^{(l)}$ is a Cauchy sequence. In a complete space, this sequence has a limit f, which is the fixed point of T.
Indeed,

$$\begin{aligned} \|T(\beta) - \beta\| &\leq \|T(\beta) - \beta^{(l+1)}\| + \|\beta^{(l+1)} - \beta\| \\ &= \|T(\beta) - T(\beta^{(l)})\| + \|\beta^{(l+1)} - \beta\| \\ &\leq \rho\|\beta - \beta^{(l)}\| + \|\beta^{(l+1)} - \beta\| \end{aligned}$$

Since these two terms tend to zero for $l \to \infty$, we have $T(\beta) = \beta$, i.e. β is a fixed point.
The fixed point being unique, all sequences of iterates must converge to this fixed point, whatever the starting point.

The iterative Landweber scheme

- The ridge regression solution ($\lambda > 0$)

$$\beta_{ridge} = \arg\min_\beta \Phi(\beta) \quad \text{where} \quad \Phi(\beta) = \|X\beta - y\|^2 + \lambda\|\beta\|_2^2$$

can be computed e.g. via matrix inversion.

- Alternatively, the Euler equation $(X^T X + \lambda Id)\beta = X^T y$ or else

$$(1+\lambda)\,\beta = \beta + X^T y - X^T X\beta$$

suggests the following successive approximation scheme (which can be useful for large matrices)

- *Damped Landweber iteration*

$$\beta^{(0)} \text{ arbitrary;} \qquad \beta^{(l+1)} = T\,\beta^{(l)} \quad l = 0, 1, \ldots$$

with iteration map $T = (1+\lambda)^{-1} L$ where $L\beta \equiv \beta + X^T(y - X\beta)$.

The iterative Landweber scheme

• Renormalize X so that $\|X\| < 1$, i.e. $\|X^T X\| = \sigma_0^2 < 1$
(a numerical estimation of this largest eigenvalue can be obtained through the so-called 'power method').

• Then L is non-expansive: $\forall \beta, \alpha$

$$\|L\beta - L\alpha\| \leq \|\beta - \alpha\|$$

Proof: $\|L\beta - L\alpha\| = \|(Id - X^T X)(\beta - \alpha)\| \leq \|\beta - \alpha\|$

• Hence T is a contraction (for $\lambda > 0$)
$\rightarrow$ iteration converges to the unique fixed point of T
= unique minimizer of the strictly convex cost function $\Phi(\beta)$
(since it satisfies the Euler equation)

• Advantage: extra positivity or other constraints can be easily enforced (at each iteration)

The iterative Landweber scheme

• For $\lambda = 0$, the unregularized Landweber iteration is still a contraction provided that the matrix $X^T X$ has full rank. Indeed, in such a case

$$\|L\beta - L\alpha\| = \|(Id - X^T X)(\beta - \alpha)\| \leq \rho\|\beta - \alpha\|$$

with $\rho = \|Id - X^T X\| = sup_k(1 - \sigma_k^2) < 1$ (since all singular values lie between zero and one).

• The iteration converges to the generalized solution $\beta^\dagger$, i.e. to the least-squares (OLS) solution of minimal norm (if not unique).

• The contractive property of L holds no longer true in the presence of a non trivial null-space or in general for a compact operator in an infinite-dim. Hilbert space, but the convergence result still holds.

The iterative Landweber scheme

The proof is more involved (exercise): use SVD and rewrite the Landweber iteration as a linear filtering method, with spectral filters $w_k^l = 1 - (1 - \sigma_k^2)^l$.

• For ill-posed or ill-conditioned problems, the unregularized Landweber iteration can be regularized by early stopping (the number of iteration playing the role of the regularization parameter). This prevents the instabilities due the the smallest singular values and the noise to appear.

Lasso regression

• If the object is known a priori to be sparse (many zero entries in the vector β), replace L^2-penalty (Ridge sol. not sparse for noisy data) by a penalty on the L^1-norm of β:

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

i.e. increase penalty on $|\beta_j| < 1$ and decrease penalty on larger components to favor the restoration of objects with few but large components.

• $\rightarrow$ solve penalized least-squares problem ($\tau > 0$)

$$\beta_{ridge} = \arg\min_\beta \Phi(\beta) \quad \text{where} \quad \Phi(\beta) = \|X\beta - y\|^2 + \tau\|\beta\|_1$$

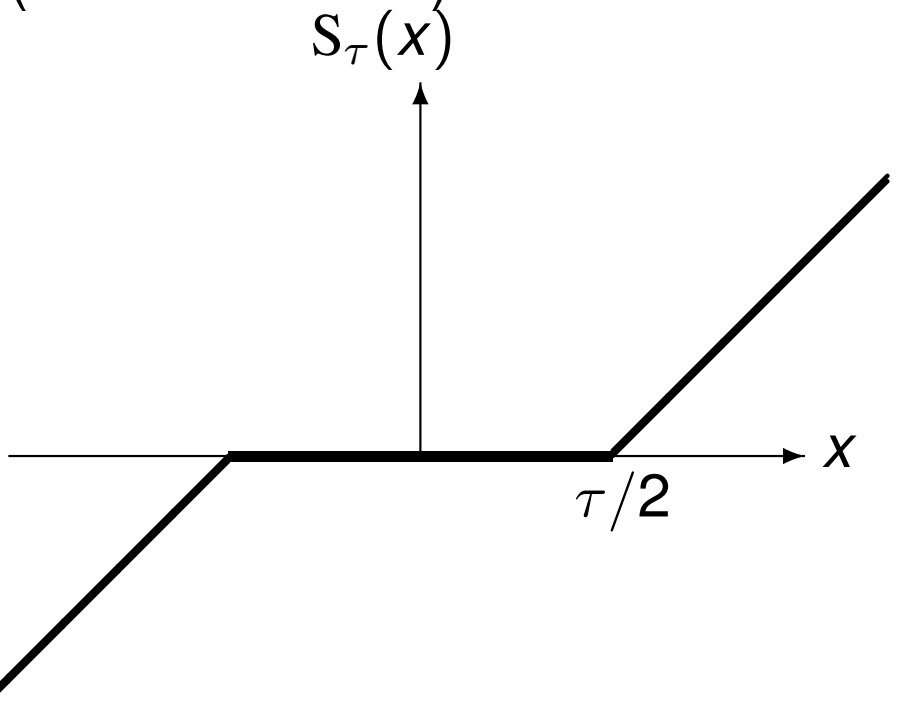
• Notice that $\Phi(\beta)$ is still convex – but not strictly if X has a non trivial null-space.

Denoising by Soft-Thresholding

- Lemma: The minimizer x^* of $(x-y)^2 + \tau|x|$ is given by $x^* = S_\tau(y)$, where

$$S_\tau(y) = \begin{cases} y - (\tau/2)\ \mathrm{sign}(y) & \text{if} \quad |y| \geq \tau/2 \\ 0 & \text{if} \quad |y| < \tau/2\ . \end{cases}$$

(S_τ: soft-thresholding = nonlinear shrinkage)
(Proof: exercise)



Denoising by Soft-Thresholding

- For $X = Id$ (and $n = p$), the minimizer β^* is

$$\beta^* = \mathbf{S}_\tau y$$

$$(\mathbf{S}_\tau y)_i = \begin{cases} y_i - (\tau/2)\ \mathrm{sign}(y_i) & \text{if} \quad |y_i| \geq \tau/2 \\ 0 & \text{if} \quad |y_i| < \tau/2 \ . \end{cases}$$

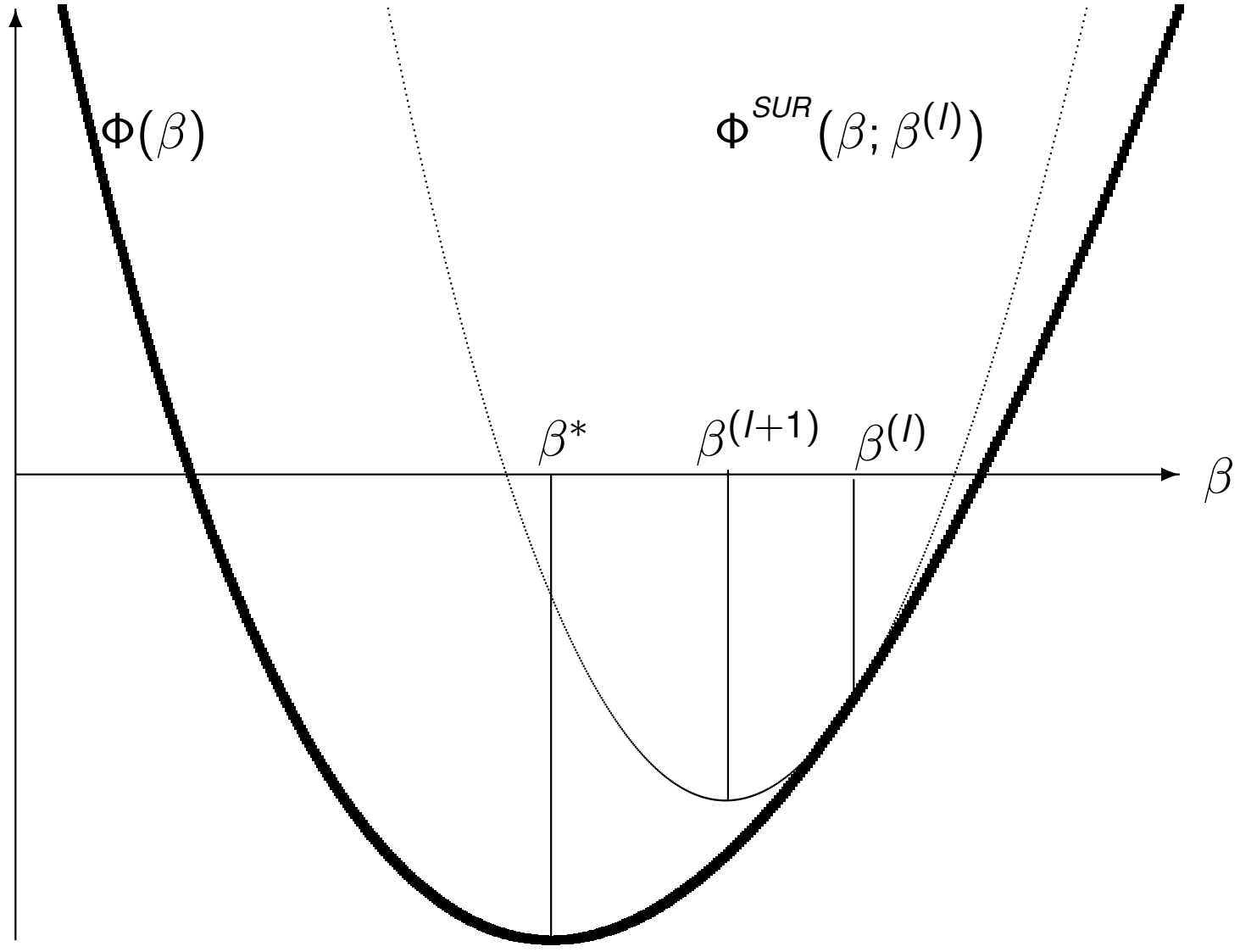
= componentwise *soft-thresholded* data vector.

- When implemented on wavelet coefficients, this is a simple *denoising* scheme as proposed by Donoho and Johnstone.

- When $X \neq Id$, it couples all object components
$\rightarrow$ complicated quadratic programming optimization problem.

- Alternatively, De Mol and Defrise (2002) proposed to use optimization transfer (De Pierro, Lange, etc.)

Optimization via Surrogate Cost Functions



Optimization via Surrogate Cost Functions

- Define

$$\Phi^{\text{SUR}}(\beta;\gamma) = \|X\beta - y\|^2 \quad - \quad \|X\beta - X\gamma\|^2 \\ + \quad \|\beta - \gamma\|^2 + \tau\|\beta\|_1$$

- Properties:
  - (i) $\Phi^{\text{SUR}}(\beta;\gamma)$ strictly convex $\forall\gamma$, since $\|X\| < 1$
  - (ii) $\Phi^{\text{SUR}}(\beta;\gamma) \geq \Phi(\beta)$
  - (iii) $\Phi^{\text{SUR}}(\beta;\beta) = \Phi(\beta)$
- Minimizer β^* of $\Phi(\beta)$ is approached through iterative scheme ($l = 0, 1, \ldots;$ $\beta^{(0)}$ arbitrary):

$$\beta^{(l+1)} = \arg\min_\beta \Phi^{\text{SUR}}(\beta;\beta^{(l)})$$

- This ensures a monotonic decrease of the cost function at each iteration since

$$\Phi(\beta^{(l+1)}) \leq \Phi^{\text{SUR}}(\beta^{(l+1)};\beta^{(l)}) \leq \Phi^{\text{SUR}}(\beta^{(l)};\beta^{(l)}) = \Phi(\beta^{(l)})$$

ISTA: Iterative Soft-Thresholding Algorithm

• At each iteration, the minimization problem is decoupled for each pixel value $\rightarrow$ solved explicitly (exercise):

$$\beta^{(l+1)} = T\ \beta^{(l)} \quad \text{with} \quad T = \mathbf{S}_\tau L$$

i.e. Landweber scheme $L\beta \equiv \beta + X^T(y - X\beta)$ with soft-thresholding at each iteration.

• When X has a zero null-space (i.e. when $X\beta = 0 \Rightarrow \beta = 0$), L is a contraction.

Then, since $\mathbf{S}_\tau$ is non-expansive, T is also a contraction.

$\rightarrow$ convergence of the iteration to the unique fixed point of T.

This fixed point is the unique minimizer of $\Phi(\beta)$ (i.e. satisfies the Euler equation; exercise).

• NB. Many faster algorithms have been proposed for sparse recovery in the recent literature

Generalizations

• Convergence of the previous scheme to a minimizer of $\Phi(\beta)$ holds under the following more general conditions (Daubechies, Defrise & De Mol, Comm. Pure Appl. Math, 2004)

(i) X has a non-zero null-space

(ii) strong convergence holds in infinite-dim. setting ($\|y\|$ = Hilbert L^2-norm; X bounded operator in L^2)

(iii) Penalty is $\|\beta\|_\alpha^\alpha = \sum_{j=1}^p w_j |\beta_j|^\alpha$ (weighted L^α-norm, $1 \leq \alpha \leq 2$) on the sequence of coefficients of β on a given o.n. basis in L^2
Examples: wavelet basis (Besov-norm penalty); Fourier basis ; basis in pixel space

• Moreover, we have a true regularization method for the (infinite-dim.) ill-posed problem (reducing to Tikhonov's regularization for $\alpha = 2$ and $\lambda = \tau$).