Projection onto an ℓ_1-norm Ball with Application to Identification of Sparse Autoregressive Models

Jitkomut Songsiri

Department of Electrical Engineering

Chulalongkorn University

Asean Symposium on Automatic Control

Mar 8-9, 2011

Outline

- **Sparse identification**
- Projection onto an ℓ_1-norm ball
- Numerical examples

Sparse identification

parameter estimation problems with sparsity-promoting regularization

$$\text{minimize} \;\; f(x) \;\; \text{subject to} \;\; \|x\|_1 \leq \rho$$

- f is a loss function (norm squared error, loglikelihood, etc.)
- ρ is a given positive parameter
- the optimization variable is $x \in \mathbf{R}^n$

Motivations

- ℓ_1-norm constraint encourages sparsity in x for a sufficiently small ρ
- many zeros in x correspond to a model with less number of parameters
- parsimonious models require fewer observations

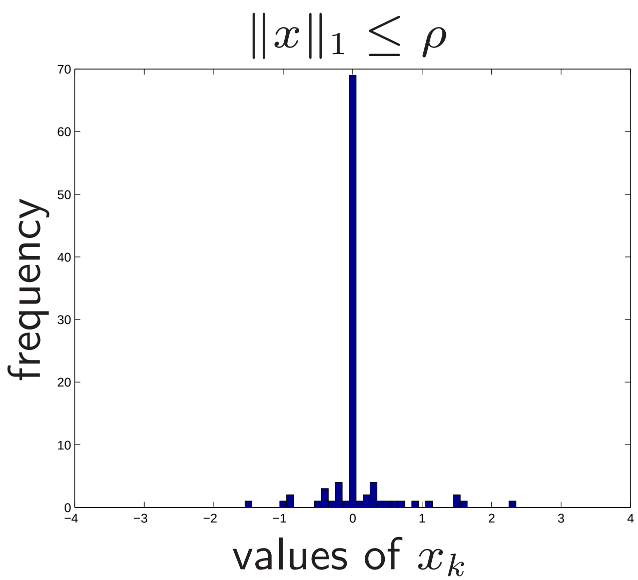
used in bioinformatics, digital communication, pattern recognition, ...

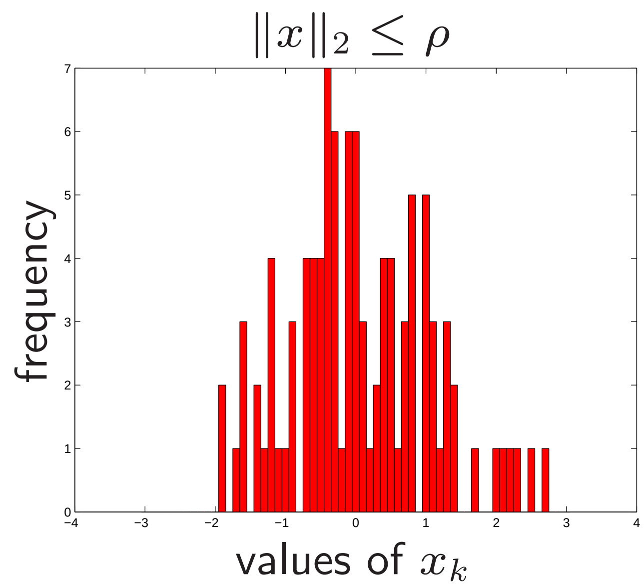
Example: Lasso problem (Tibshirani 1996)

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & \|x\|_1 \leq \rho \end{array}$$

with variable $x \in \mathbf{R}^n$

- a heuristic for regression selection to find a sparse solution
- find many applications on signal processing, image reconstruction, and compressed sensing, ...





Sparse Autoregressive (AR) Models

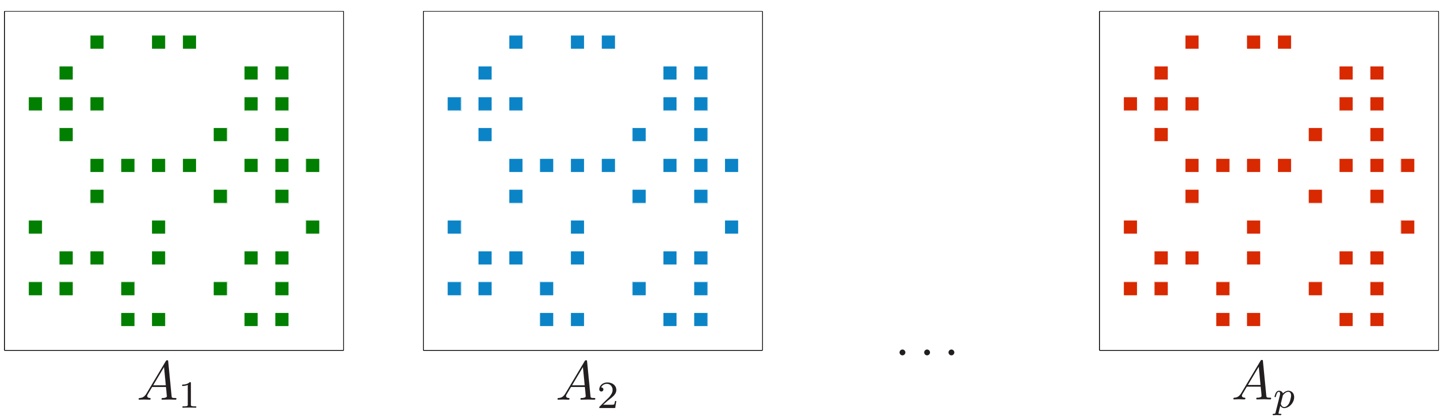
a multivariate autoregressive process of order p

$$y(t) = \sum_{k=1}^{p} A_k y(t-k) + \nu(t)$$

$y(t) \in \mathbf{R}^n$, $A_k \in \mathbf{R}^{n\times n}$, $k = 1, 2, \ldots, p$, $\nu(t)$ is noise

Problem: find A_k's that minimize the mean-squared error and

- A_k's contain many zeros
- common zero locations in $A_1, A_2, \ldots, A_p$



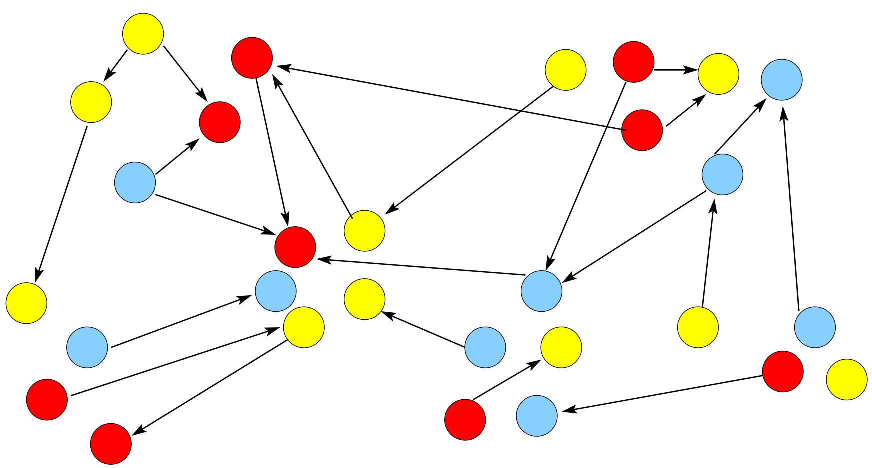
Statistical interpretation (Granger 1969)

sparsity in coefficients A_k

$$(A_k)_{ij} = 0, \quad \text{for } k = 1, 2, \ldots, p$$

is the characterization of **Granger causality** of AR models

- y_i is not *Granger-caused* by y_j
- knowing y_j does not help to improve the prediction of y_i



applications in neuroscience and system biology

(Salvador et al. 2005, Valdes-Sosa et al. 2005, Fujita et al. 2007, ...)

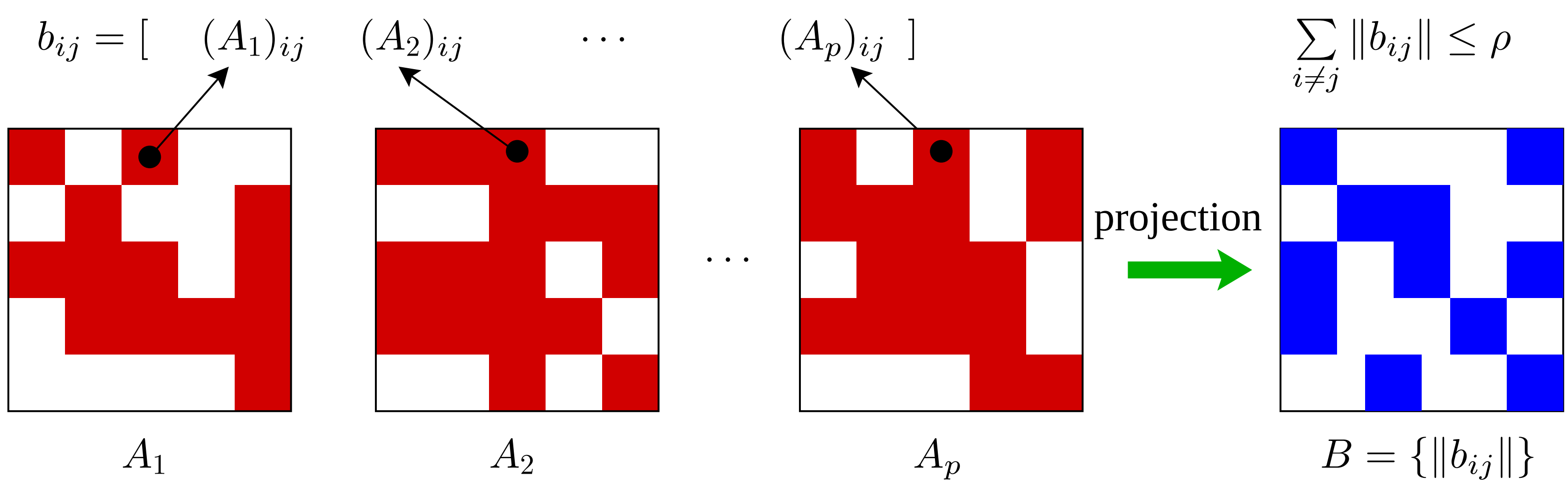
Sum of ℓ_2-norm

suppose c_k is a vector in $\mathbf{R}^n$, the constraint

$$\|c_1\| + \|c_2\| + \cdots + \|c_m\| \leq \rho$$

makes some c_k's *zero* vectors (for a sufficiently small ρ)

idea: to make a common sparsity in A_k's



$$\|b_{ij}\| = 0 \quad \Longleftrightarrow \quad (A_1)_{ij} = (A_2)_{ij} = \cdots = (A_p)_{ij} = 0$$

Estimation problem

given the measurements $y(1), y(2), \ldots, y(N)$

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{t=p+1}^{N} \|y(t) - \textstyle\sum_{k=1}^{p} A_k y(t-k)\|^2 \\ \text{subject to} & \displaystyle\sum_{i \neq j} \left\| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \right\|_2 \leq \rho \end{array}$$

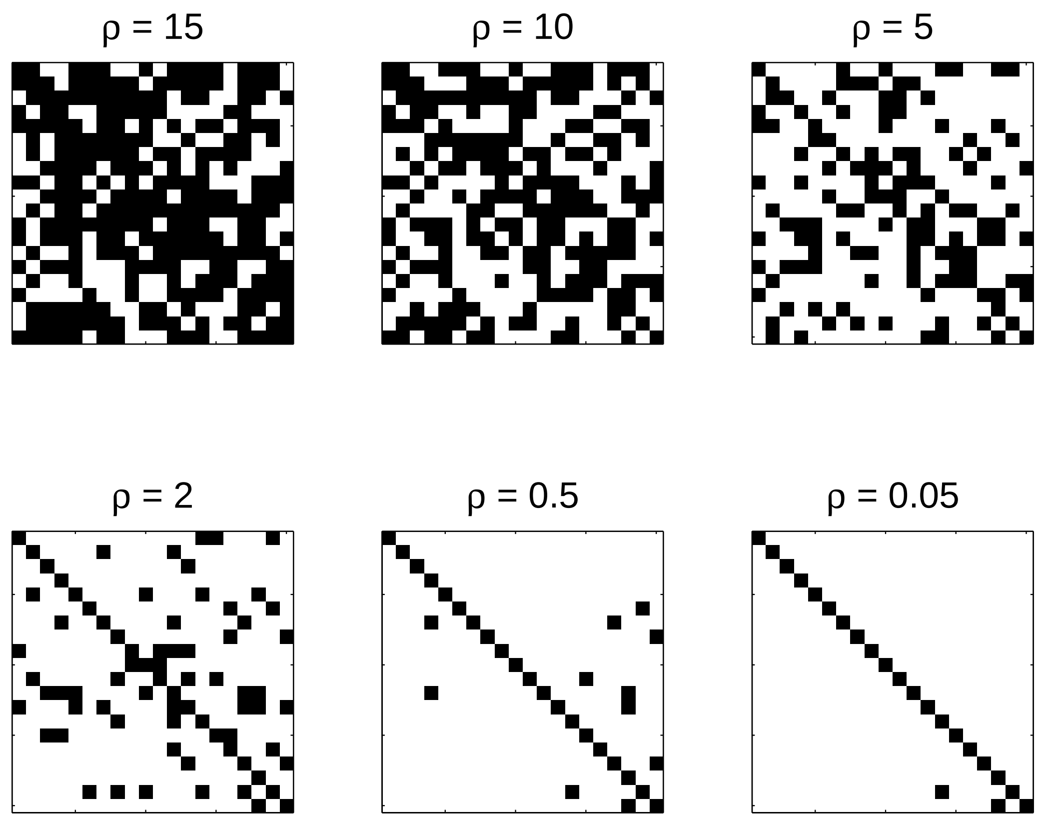
with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$

- summation over (i, j) plays a role of ℓ_1-type norm
- using the ℓ_2 norm of p-tuple of $(A_k)_{ij}$ yields a group sparsity

a heuristic convex approach to obtain sparse AR coefficients

Example: $n = 20$, $p = 3$

common zero patterns of a solution A_k, $k = 1, 2, \ldots, p$



as ρ decreases, A_k's contain more zeros

Numerical solutions

The estimation problem can be expressed by

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

with variable $x \in \mathbf{R}^n$ and $\mathcal{C}$ is a convex set (here ℓ_1 ball)

problem: how to solve this optimization problem in large scale ?

idea: use a projected gradient method which is based on the update

$$x^{(k+1)} = \mathcal{P}_{\mathcal{C}}(x^{(k)} - t^{(k)} \nabla f(x^{(k)}))$$

- $t^{(k)}$ is a step size, and ∇f is the gradient of f
- $\mathcal{P}_{\mathcal{C}}$ is a Euclidean projection onto $\mathcal{C}$, defined by

$$P_{\mathcal{C}}(y) = \text{argmin}_x \|x - y\| \;\text{ subject to }\; x \in \mathcal{C}.$$

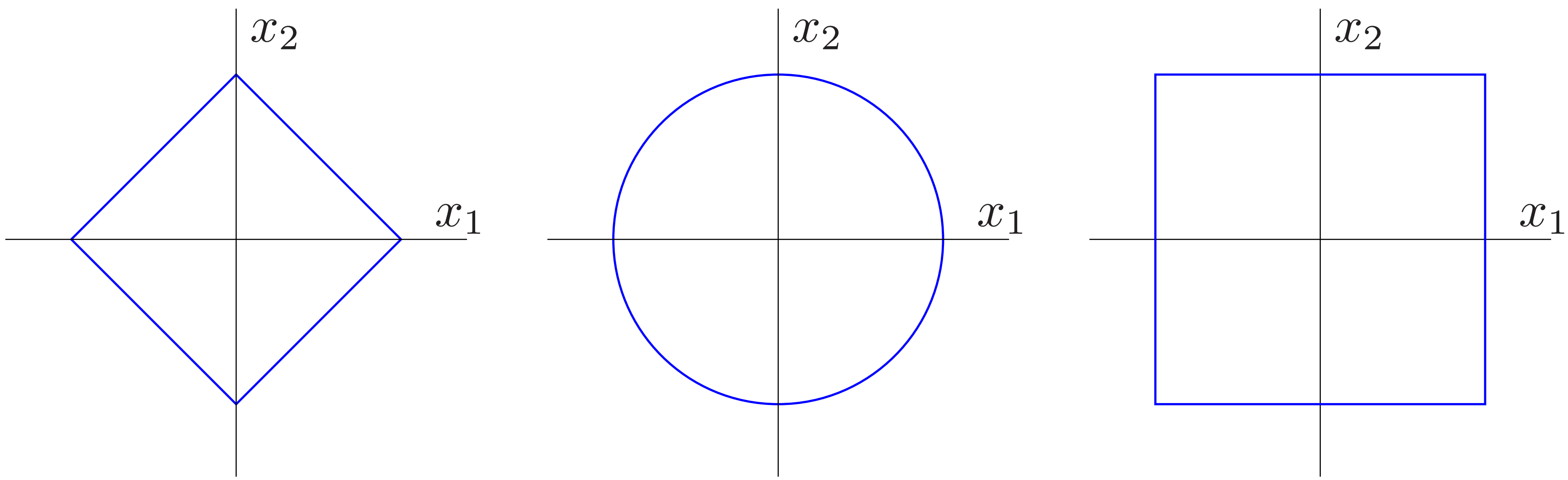
Outline

- Sparse identification
- **Projection onto an ℓ_1-norm ball**
- Numerical examples

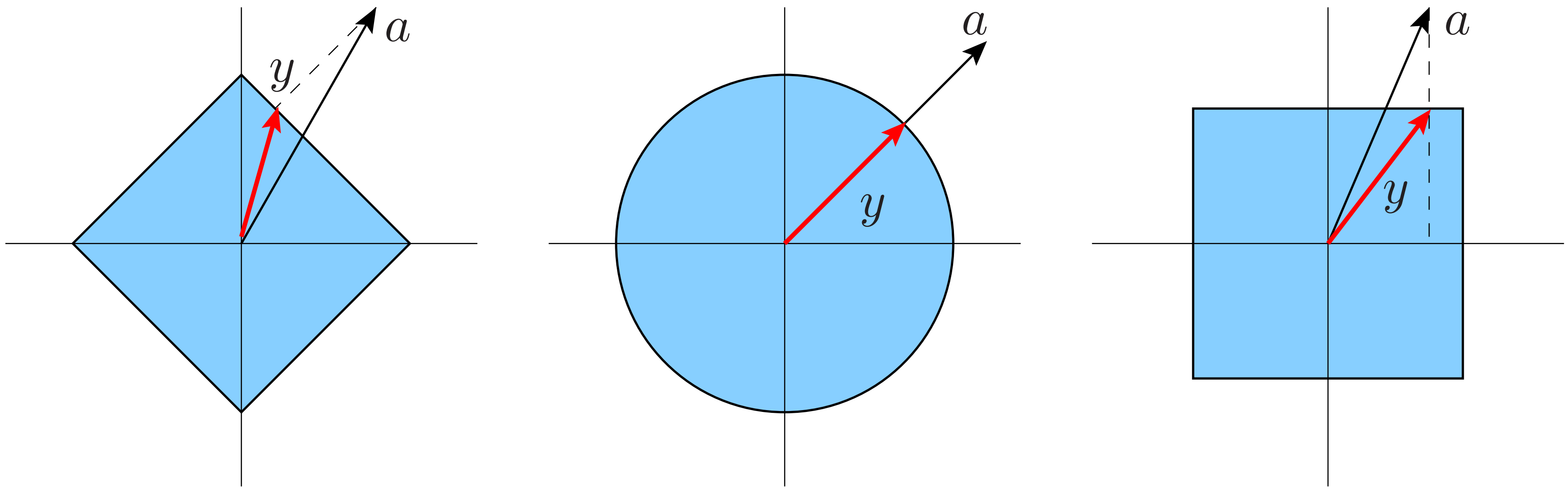
Euclidean projections

the Euclidean projection of a vector $a \in \mathbf{R}^n$ onto the unit ℓ_p-norm ball

$$\begin{array}{ll} \text{minimize} & \|y - a\|_2^2 \\ \text{subject to} & \|y\|_p \leq 1 \end{array}$$



ℓ_1-norm ball ℓ_2-norm ball ℓ_∞-norm ball



projection onto ℓ_2 ball

$$y = \frac{a}{\|a\|_2}$$

projection onto ℓ_∞ ball

$$y_k = \begin{cases} a_k, & |a_k| \leq 1 \\ \mathsf{sign}(a_k), & |a_k| \geq 1 \end{cases}$$

projection onto ℓ_1 ball

no closed-form solution

Projection onto the ℓ_1-norm ball

Primal problem

$$\begin{array}{ll} \text{minimize} & \|y - a\|_2^2 \\ \text{subject to} & \|y\|_1 \le 1 \end{array}$$

with variable $y \in \mathbf{R}^n$

Dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda) := \sum_k g_k(\lambda) - 2\lambda \\ \text{subject to} & \lambda \ge 0, \end{array}$$

where g_k is given by

$$g_k(\lambda) = \begin{cases} -(\lambda - |a_k|)^2 + a_k^2, & \lambda < |a_k| \\ a_k^2, & \lambda \ge |a_k| \end{cases}, \quad k = 1, 2, \ldots, n$$

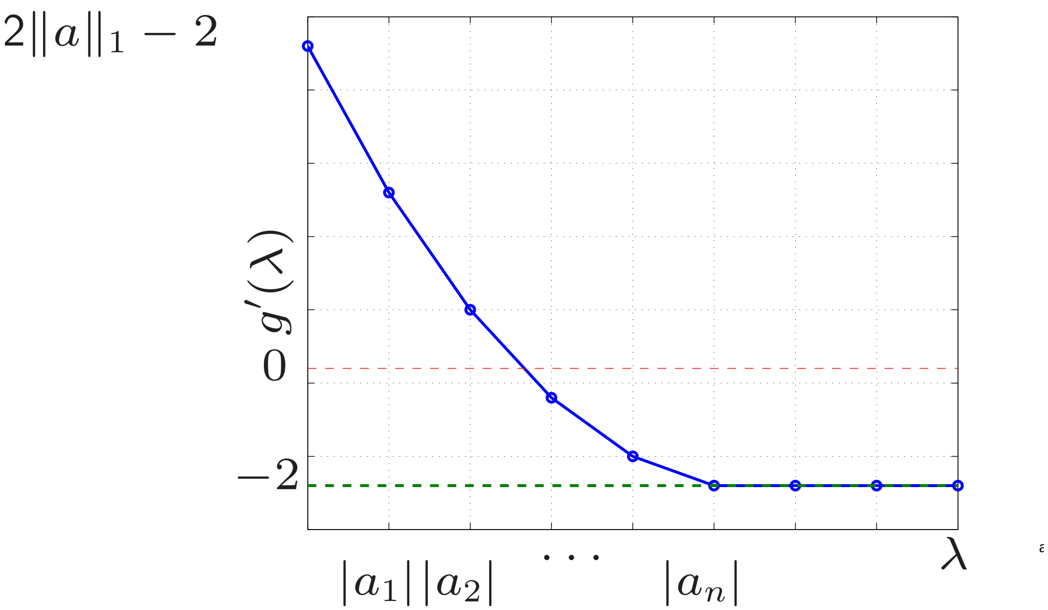
with variable $\lambda \in \mathbf{R}$

g_k' is a piecewise linear function in λ

$$g_k'(\lambda) = \begin{cases} 2(|a_k| - \lambda), & \lambda < |a_k| \\ 0, & \lambda \geq |a_k|. \end{cases}$$

if $\|a\|_1 > 1$, then the dual optimal point λ^* is given by the root of

$$g'(\lambda) = \sum_{k=1}^{n} \max(|a_k| - \lambda, 0) - 1 = 0$$



sort $|a_k|$ such that

$$|a_1| \leq |a_2| \leq \ldots \leq |a_n|$$

Algorithm

1. If $\|a\|_1 \leq 1$, then $\lambda^* = 0$.

2. Otherwise, define $a_0 = 0$ and sort $|a_k|$ in ascending order. Compute

λ	$g'(\lambda)/2$
$\lvert a_0\rvert = 0$	$\|a\|_1 - 1$
$\lvert a_1\rvert$	$(1-n)\lvert a_1\rvert + \sum_{k=2}^n \lvert a_k\rvert - 1$
$\lvert a_2\rvert$	$(2-n)\lvert a_2\rvert + \sum_{k=3}^n \lvert a_k\rvert - 1$
$\vdots$	$\vdots$
$\lvert a_{n-1}\rvert$	$-\lvert a_{n-1}\rvert + \lvert a_n\rvert - 1$
$\lvert a_n\rvert$	-1

3. Locate the interval where $g'(\lambda)$ changes its sign, *i.e.*, find k such that

$$g'(|a_k|) \geq 0 \quad \text{and} \quad g'(|a_{k+1}|) \leq 0$$

4. the point where $g'(\lambda) = 0$ is

$$\lambda^* = \frac{\left(\sum_{j=k+1}^{n} |a_j|\right) - 1}{(n-k)}$$

5. Using λ^* to compute the projection y^* from

$$y_k^* = \begin{cases} a_k + \lambda^*, & a_k \leq -\lambda^* \\ 0, & |a_k| < \lambda^* \\ a_k - \lambda^*, & a_k \geq \lambda^*, \end{cases}$$

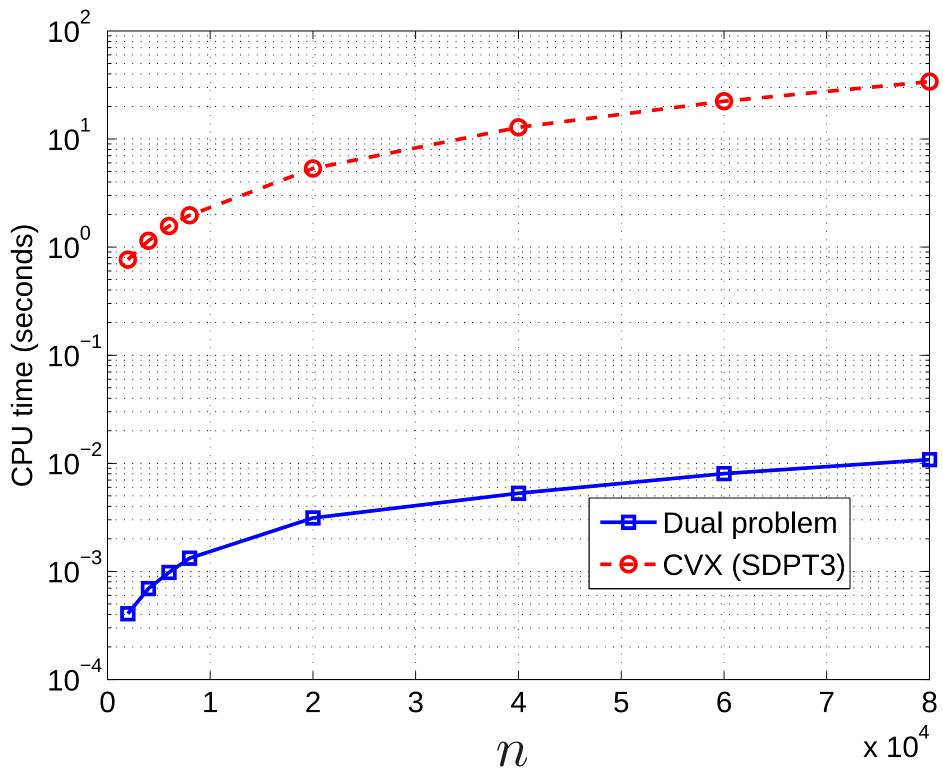
- the relation between y^* and λ^* is derived via duality
- it shows the location of zeros in y

Outline

- Sparse identification
- Projection onto an ℓ_1-norm ball
- **Numerical examples**

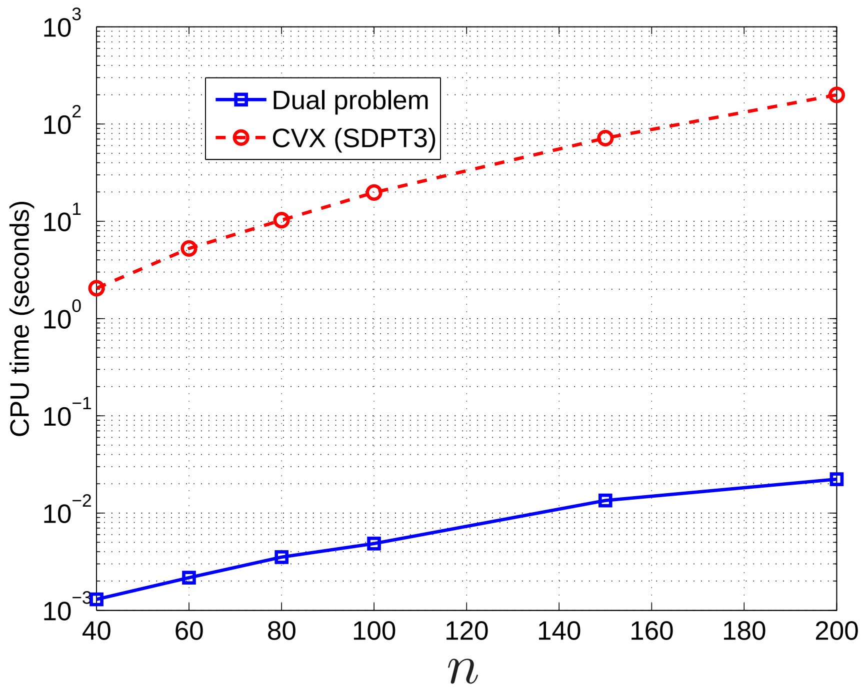
Numerical examples

projection problem with n ranges from 800 to 80000



- blue line - solve the dual problem by the proposed algorithm
- red line - solve the primal problem by an interior-point method

Projection of AR coefficients



blue line - proposed algorithm

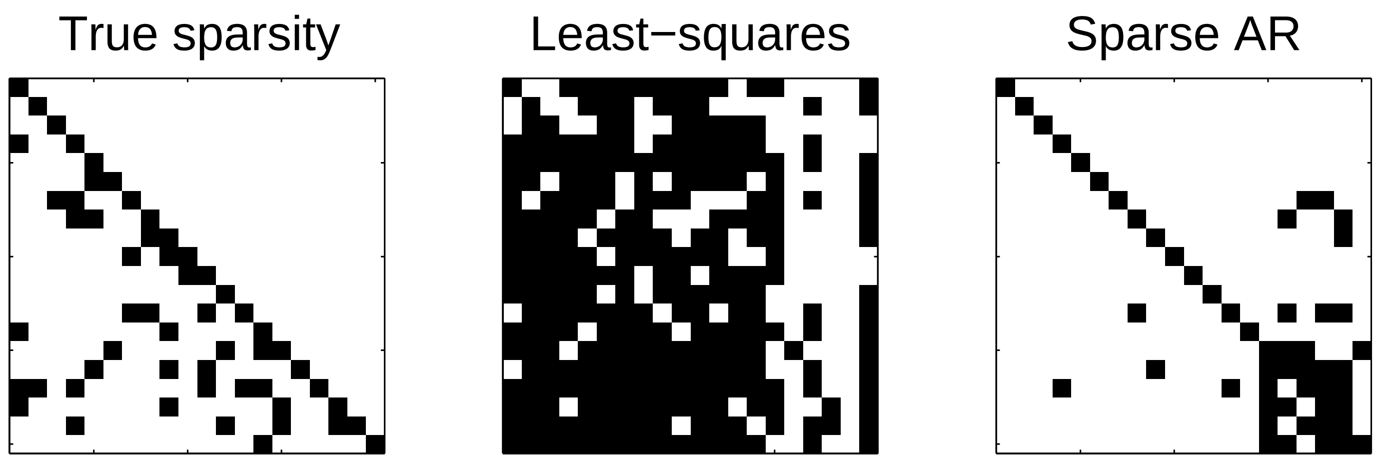
red line - IP method

- n ranges from 40 to 200 and $p = 3$ ($n^2 p$ ranges from 4800 to 120000)
- using $\rho = 5$, compute a projection of A_1, A_2, A_3 onto the set

$$\sum_{i \neq j} \| \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \|_2 \leq \rho$$

Sparse AR estimation

generate 500 time points from a sparse AR process with $n = 50$ and $p = 3$



- a few data and presence of noise make LS solution a bad estimate
- when a sparse solution is favor, adding ℓ_1-type contraints is an efficient convex approach to serve this purpose

Summary

- sparse identification is useful for learning structures in complex systems
- a heuristic approach to yield a sparse solution is to add ℓ_1-type constraints
- solving large-scale sparse optimization problems requires cheap computation of a projection onto ℓ_1-norm ball
- an efficient method to compute projections is derived via the dual problem

(Selected) References

- R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996
- M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006
- A. Fujita, P. Severino, J. Sato, and S. Miyano, "Granger causality in systems biology: modeling gene networks in time series microarray data using vector autoregressive models," *Advances in Bioinformatics and Computational Biology*, pp. 13–24, 2010
- M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007