Projection onto an $\ell_1$-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 $\ell_1$-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.)
- \( \rho \) is a given positive parameter
- the optimization variable is \( x \in \mathbb{R}^n \)

Motivations

- \( \ell_1 \)-norm constraint encourages sparsity in \( x \) for a sufficiently small \( \rho \)
- 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  

\[
\begin{align*}
\text{minimize} & \quad \|Ax - b\|_2^2 \\
\text{subject to} & \quad \|x\|_1 \leq \rho
\end{align*}
\]

with variable \( x \in \mathbb{R}^n \)

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

![Graphs showing \( \|x\|_1 \leq \rho \) and \( \|x\|_2 \leq \rho \)]
Sparse Autoregressive (AR) Models

a multivariate autoregressive process of order $p$

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

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

Sum of $\ell_2$-norm

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

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

makes some $c_k$'s zero vectors (for a sufficiently small $\rho$)

idea: to make a common sparsity in $A_k$'s

$$b_{ij} = \begin{bmatrix} (A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix}$$

$$\sum_{i \neq j} \|b_{ij}\| \leq \rho$$

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

projection

$B = \{\|b_{ij}\|\}$
Estimation problem

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

minimize $\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t - k)\|^2$

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

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

- summation over $(i, j)$ plays a role of $\ell_1$-type norm
- using the $\ell_2$ 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 \( \rho \) decreases, \( A_k \)'s contain more zeros
Numerical solutions

The estimation problem can be expressed by

\[
\begin{align*}
\text{minimize} \quad & f(x) \\
\text{subject to} \quad & x \in C
\end{align*}
\]

with variable \( x \in \mathbb{R}^n \) and \( C \) is a convex set (here \( \ell_1 \) ball)

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

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

\[
x^{(k+1)} = P_C(x^{(k)} - t^{(k)} \nabla f(x^{(k)}))
\]

- \( t^{(k)} \) is a step size, and \( \nabla f \) is the gradient of \( f \)
- \( P_C \) is a Euclidean projection onto \( C \), defined by

\[
P_C(y) = \arg\min_x \|x - y\| \quad \text{subject to} \quad x \in C.
\]
Outline

- Sparse identification
- Projection onto an $\ell_1$-norm ball
- Numerical examples
Euclidean projections

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

minimize $\|y - a\|^2_2$

subject to $\|y\|_p \leq 1$

\begin{align*}
\ell_1\text{-norm ball} & \quad \ell_2\text{-norm ball} & \quad \ell_\infty\text{-norm ball}
\end{align*}
projection onto $\ell_2$ ball

\[ y = \frac{a}{\|a\|_2} \]

projection onto $\ell_\infty$ ball

\[ y_k = \begin{cases} a_k, & |a_k| \leq 1 \\ \text{sign}(a_k), & |a_k| \geq 1 \end{cases} \]

projection onto $\ell_1$ ball

no closed-form solution
Projection onto the $\ell_1$-norm ball

**Primal problem**

minimize $\|y - a\|_2^2$

subject to $\|y\|_1 \leq 1$

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

**Dual problem**

maximize $g(\lambda) := \sum_k g_k(\lambda) - 2\lambda$

subject to $\lambda \geq 0$, $

where $g_k$ is given by

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

with variable $\lambda \in \mathbb{R}$
$g'_k$ is a piecewise linear function in $\lambda$

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

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

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

2$\|a\|_1 - 2$

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

$$
\begin{array}{c|c}
\lambda & g'(\lambda)/2 \\
\hline
|a_0| = 0 & \|a\|_1 - 1 \\
|a_1| & (1 - n)|a_1| + \sum_{k=2}^{n} |a_k| - 1 \\
|a_2| & (2 - n)|a_2| + \sum_{k=3}^{n} |a_k| - 1 \\
\vdots & \vdots \\
|a_{n-1}| & -|a_{n-1}| + |a_n| - 1 \\
|a_n| & -1 \\
\end{array}
$$

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

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

5. Using \( \lambda^* \) 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 \( \lambda^* \) is derived via duality
- it shows the location of zeros in \( y \)
Outline

- Sparse identification
- Projection onto an $\ell_1$-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

• $n$ ranges from 40 to 200 and $p = 3$ ($n^2p$ ranges from 4800 to 120000)

• using $\rho = 5$, compute a projection of $A_1, A_2, A_3$ onto the set

$$\sum_{i \neq j} \| [(A_1)_{ij} \ (A_2)_{ij} \ \cdots \ (A_p)_{ij}] \|_2 \leq \rho$$
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 $\ell_1$-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 $\ell_1$-type constraints

- solving large-scale sparse optimization problems requires cheap computation of a projection onto $\ell_1$-norm ball

- an efficient method to compute projections is derived via the dual problem
(Selected) References


