ℓ_1 sparse methods in system identification

Cristian R. Rojas and Bo Wahlberg

Automatic Control Lab and ACCESS Linnaeus Centre
KTH - Royal Institute of Technology, Stockholm, Sweden

ROYAL INSTITUTE OF TECHNOLOGY

Tutorial on "Sparse and low-rank representation methods in control, estimation and system identification"
ECC, July 17-19, 2013

Outline

- Problem statement
- Convex relaxation
- Applications
- How to tune the regularization parameter?
- When / why does it work?
- Some theory
- A model selection tradeoff
- An alternative: Bayesian methods
- Conclusions

Problem statement

Assumptions:

- System: $Y_N = \Phi_N \theta^o + E_N$
  - $E_N \sim (0, \sigma^2 I)$
  - n: # parameters
  - N: # samples
  - *Sparsity:* Most entries of θ^o are **zero**
- Model: $Y_N = \Phi_N \theta + E_N$

Problem: Estimate zeros of θ (detection - model selection)
& non-zero entries (estimation)

ROYAL INSTITUTE OF TECHNOLOGY

ℓ_0 regularization

Idea: Impose sparsity as constraint on $\#$ nonzero entries of θ:

$$\min_{\theta} V_N(\theta)$$
$$\text{s.t. } \|\theta\|_0 \leq c$$

Here:

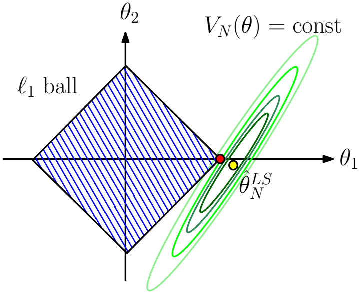
- $V_N(\theta) := \frac{1}{N}\|Y_N - \Phi_N\theta\|_2^2$ (least squares criterion)
- $\|\theta\|_0 := \#$ non-zero parameters

Problem: Combinatorial explosion (intractable if n is large)

Convex relaxation

Replace hopeless problem with relaxation!

$$\min_{\theta} V_N(\theta)$$
$$\text{s.t. } \|\theta\|_1 \leq \lambda \qquad \text{(LASSO)}$$



$$\|\theta\|_1 := |\theta_1| + \cdots + |\theta_n|$$

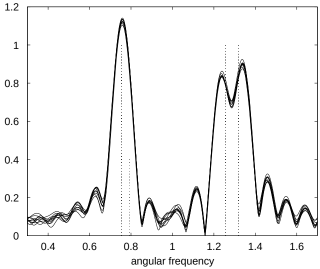
Still remaining: How to determine λ?

Applications

1. Spectral line estimation

$$y_t = \sum_{k=1}^{N} \alpha_k e^{j\omega_k t} + e_t,$$

$$\alpha_k \in \mathbb{C}$$



Idea: Grid the frequency range!

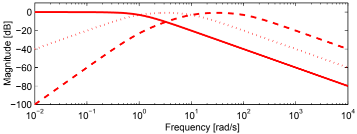
$$Y_N = \begin{bmatrix} y_{t_1} \\ \vdots \\ y_{t_N} \end{bmatrix} \quad \Phi_N = \begin{bmatrix} e^{j\omega_1 t_1} & \cdots & e^{j\omega_n t_N} \\ \vdots & \ddots & \vdots \\ e^{j\omega_1 t_N} & \cdots & e^{j\omega_n t_N} \end{bmatrix}$$

Impose sparsity constraint on $\theta = [\alpha_1 \ \cdots \ \alpha_n]^T$

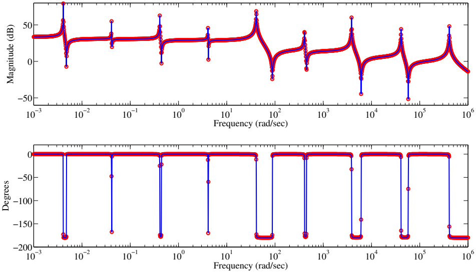
Applications (cont.)

2. Basis function selection / separable least squares

Same idea as for spectral line estimation, but using general basis functions: Laguerre, Kautz, ...



Basis functions



Fit

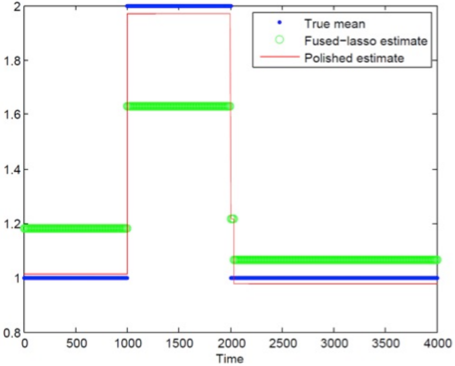
Ref: Welsh, Rojas, Hjalmarsson & Wahlberg, SYSID, 2012

Applications (cont.)

3. Change detection

$$y_t \sim \mathcal{N}(m_t, \sigma^2),$$

where $m_{t+1} = m_t$ *often*



Fused LASSO:

$$\min_{m_t} \frac{1}{2} \sum_{t=1}^{N} [y_t - m_t]^2 + \lambda \sum_{t=2}^{N} |m_t - m_{t-1}|$$

How to tune λ?

- **AIC / BIC:**

$$\min_{\lambda} \; V_N(\hat{\theta}_\lambda) + \mathsf{pen}(\mathsf{DF}(\hat{\theta}_\lambda))$$

where

$$\mathsf{DF}(\hat{\theta}_\lambda) = \|\hat{\theta}_\lambda\|_0$$

$$\mathsf{pen}(n) = 2n/N \text{ (AIC) or } = n\ln(N)/N \text{ (BIC)}$$

- **Cross-validation:**

$$\min_{\lambda} \; V_N^{val}(\hat{\theta}_\lambda)$$

How to tune λ? (cont.)

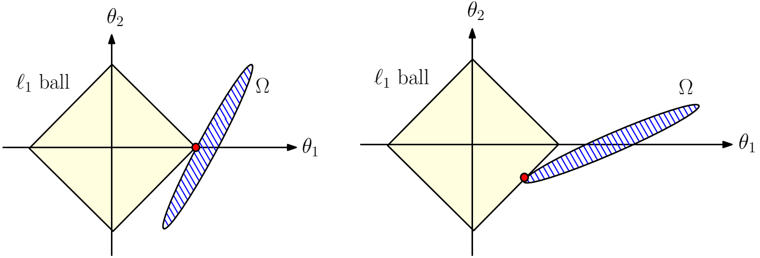
- **SPARSEVA:** (for $n < N$)

$$\begin{aligned} &\min_{\theta} \ \|\theta\|_1 \\ &\text{s.t.} \ \ V_N(\theta) \leq V_N(\hat{\theta}_N^{LS})(1+\varepsilon_N) \end{aligned}$$

  where $\varepsilon_N = 2n/N$ (AIC) or $= n\ln(N)/N$ (BIC)

- **Data independent choices:** E.g. $\lambda \propto N^c$ $(1/2 < c < 1)$

When / why does it work?



Sparse solution NOT obtained

- Shape of level curves of V_N depend on regressors Φ

When / why does it work? (cont.)

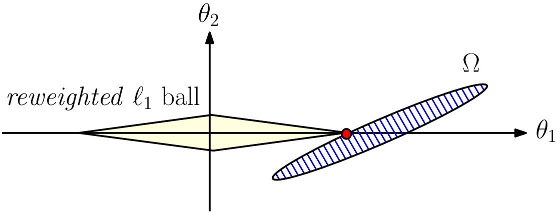
One solution: **Adaptive LASSO** (H. Zou, JASA, 2006)

$$\min_{\theta} V_N(\theta)$$
$$\text{s.t. } \sum_k \frac{|\theta_k|}{|\hat{\theta}_k^{LS}|} \leq \lambda$$

Interpretations:

(1) Resembles "$\|\theta\|_0 < \lambda$" !

(2) Reweighting of the ℓ_1 ball



Some theory

Definition (Consistence)

$\hat{\theta}_N$ is consistent in probability if $\hat{\theta}_N \xrightarrow{p} \theta^o$ as $N \to \infty$

Consistence is mostly useful if $n \ll N$. Otherwise, the following notion is relevant:

Definition (Persistence)

$\hat{\theta}_N$ is persistent if $E\{V_N(\hat{\theta}_N)\} - E\{V_N(\theta_N^*)\} \to 0$ as $N \to \infty$, where

$$\theta_N^* = \arg\min_{\theta} E\{V_N(\theta)\}$$

- LASSO-type estimators are typically consistent if AIC/BIC is used (for fixed n), and persistent when using CV

Some theory (cont.)

Definition (Model selection consistence / sparsistence)

$\hat{\theta}_N$ is model selection consistent if $\mathbf{P}\{\text{supp } \hat{\theta}_N = \text{supp } \theta^o\} \to 1$ as $N \to \infty$

- Adaptive LASSO with λ chosen via BIC is sparsistent for fixed n, while it is not with AIC
- For $n \to \infty$, (Adaptive-) LASSO is rarely sparsistent: at most one can enforce supp $\hat{\theta}_N \supseteq$ supp θ^o in probability

ROYAL INSTITUTE OF TECHNOLOGY

Some theory (cont.)

Definition (Oracle property)

$\hat{\theta}_N$ has the oracle property if

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \xrightarrow[N\to\infty]{d} \mathbf{N}(0, M^\dagger)$$

That is, $\hat{\theta}_N$ has the same asymptotic distribution as the least-squares oracle, which knows the sparsity pattern of θ^o

- If (Adaptive-) LASSO is sparsistent, one can achieve the oracle property by *polishing*: The non-zero entries of $\hat{\theta}_N$ are re-estimated using least squares

A model selection tradeoff

Definition (Minimax rate optimality)

$\hat{\theta}_N$ is minimax rate optimal if $E\{V_N(\hat{\theta}_N)\} - E\{V_N(\theta_N^*)\} \to 0$ at the fastest possible rate, uniformly in θ_0

- Minimax rate optimality $\Longrightarrow$ Optimal prediction ability
- Model selection consistence $\Longrightarrow$ Recovery of 'truth'

Can we have both?

NO! This is a fundamental limitation in estimation, independent of the estimator (Yang, 2005; Leeb & Ptscher, 2008)

ROYAL INSTITUTE OF TECHNOLOGY

An alternative: Bayesian methods

Idea: Assume that θ has a *prior* distribution

$$\theta_i \sim \mathcal{N}(0, \lambda_i), \quad \lambda_i \geq 0, \qquad i = 1, \ldots, n$$

- $\hat{\lambda}_i$ determined by maximizing $p(Y_N; \lambda_i)$
- $\hat{\theta}_i$ estimated as $E\{\theta_i | Y_N, \hat{\lambda}_i\}$

- $\lambda_i = 0 \implies \theta_i = 0!$ (the prior induces sparsity!)

- Seems to induce better sparse estimates than LASSO (i.e., more sparse for same amount of shrinkage), but relies on non-convex programming (local minima!)

Ref: Aravkin, Burke, Chiuso & Pillonetto, CDC, 2011

Conclusions

- ℓ_1 regularization as a means to impose sparsity
- Applications to model / regressor / basis function selection + estimation
- How to choose the regularization parameter?
- Theoretical properties and tradeoffs

Conclusions (cont.)

- Extensions to nonlinearly parameterized models and other kinds of sparsity (piecewise constant signals, graphical models, ...)
- Alternatives: (Empirical-) Bayesian approaches, iterative / greedy methods