SPARSEVA
SPARSe Estimation based on a VAlidation criterion

Cristian R. Rojas and Håkan Hjalmarsson
ACCESS Linnaeus Center, School of Electrical Engineering
KTH - Royal Insititute of Technology, Stockholm

5th Swedish-Chinese Conference on Control
Lund May 30, 2011

Outline

Introduction

- Model structure/regressor selection one of the key issues in SI

Introduction

- Model structure/regressor selection one of the key issues in SI
- Forward selection
- Forward stepwise selection
- LARS
- Shrinkage
- Regularization
- Compressive sensing

The problem

Assumptions:

The problem

Assumptions:

- System:

  $Y_N = \Phi_N \theta^o + E_N$

  - $\theta^o \in \mathbb{R}^n$,
  - $E_N \sim \mathbf{N}(0, \sigma^2 I_N)$
  - $\Phi_N \in \mathbb{R}^{N \times n}$
  - $Y_N \in \mathbb{R}^N$.
  - $\theta^o = [\theta_1^{oT}\ \theta_2^{oT}]^T$, where $\theta_i^o \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_2^o = 0$.
  - $\lim_{N\to\infty} N^{-1} \Phi_N^T \Phi_N =: \Gamma > 0$

The problem

Assumptions:

- System:

  $Y_N = \Phi_N \theta^o + E_N$

  - $\theta^o \in \mathbb{R}^n$,
  - $E_N \sim \mathbf{N}(0, \sigma^2 I_N)$
  - $\Phi_N \in \mathbb{R}^{N \times n}$
  - $Y_N \in \mathbb{R}^N$.
  - $\theta^o = [\theta_1^{oT}\ \theta_2^{oT}]^T$, where $\theta_i^o \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_2^o = 0$.
  - $\lim_{N\to\infty} N^{-1} \Phi_N^T \Phi_N =: \Gamma > 0$

- Model: $Y_N = \Phi_N \theta + E_N$

The problem

Assumptions:

- System:

  $Y_N = \Phi_N \theta^o + E_N$
  - $\theta^o \in \mathbb{R}^n$,
  - $E_N \sim \mathbf{N}(0, \sigma^2 I_N)$
  - $\Phi_N \in \mathbb{R}^{N \times n}$
  - $Y_N \in \mathbb{R}^N$.
  - $\theta^o = [\theta_1^{oT}\ \theta_2^{oT}]^T$, where $\theta_i^o \in \mathbb{R}^{n_i}$ $(i = 1, 2)$ and $\theta_2^o = 0$.
  - $\lim_{N\to\infty} N^{-1} \Phi_N^T \Phi_N =: \Gamma > 0$

- Model: $Y_N = \Phi_N \theta + E_N$

Problem: Estimate zeros of θ & non-zero entries

Outline

ℓ_0 regularization

Idea: Estimate high dimensional model parameter vector θ, but enforce superfluous parameters to be zero:

ℓ_0 regularization

Idea: Estimate high dimensional model parameter vector θ, but enforce superfluous parameters to be zero:

$$\begin{aligned} \min_{\theta} \; & V_N(\theta) \\ \text{s.t. } & \|\theta\|_0 \leq c \end{aligned}$$

ℓ_0 regularization

Idea: Estimate high dimensional model parameter vector θ, but enforce superfluous parameters to be zero:

$$\begin{aligned} &\min_{\theta} \; V_N(\theta) \\ &\text{s.t. } \|\theta\|_0 \leq c \end{aligned}$$

Here:

- $V_N(\theta) := \frac{1}{N}(Y_N - \Phi_N\theta)^T(Y_N - \Phi_N\theta)$ (least squares criterion)

ℓ_0 regularization

Idea: Estimate high dimensional model parameter vector θ, but enforce superfluous parameters to be zero:

$$\begin{aligned} &\min_{\theta} \; V_N(\theta) \\ &\text{s.t. } \|\theta\|_0 \leq c \end{aligned}$$

Here:

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

ℓ_0 regularization

Idea: Estimate high dimensional model parameter vector θ, but enforce superfluous parameters to be zero:

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

Here:

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

Combinatorial problem (unfortunately)

Outline

Relaxation: ℓ_1 regularization and LASSO

Replace hopeless problem with relaxation:

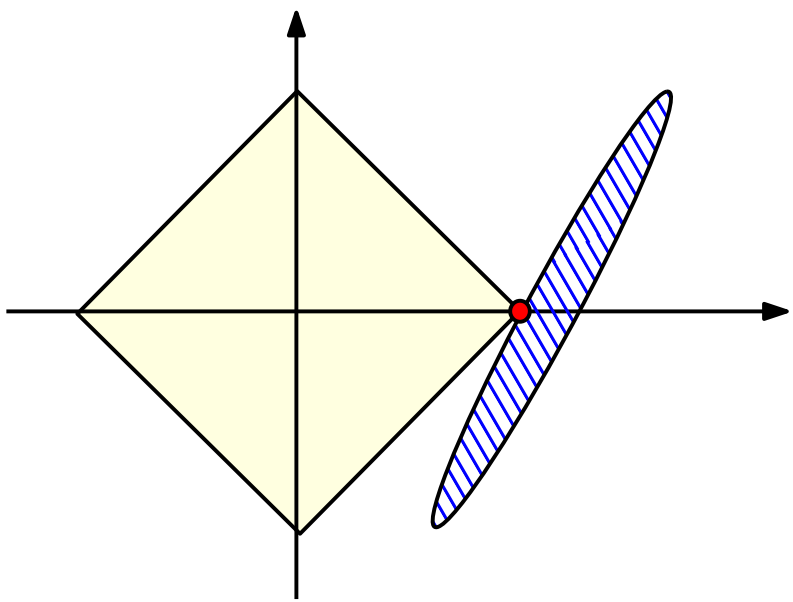
$$\begin{aligned} \min_{\theta} \; & V_N(\theta) \\ \text{s.t.} \; & \|\theta\|_1 \leq \lambda \end{aligned}$$

Relaxation: ℓ_1 regularization and LASSO

Replace hopeless problem with relaxation:

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

Illustration:

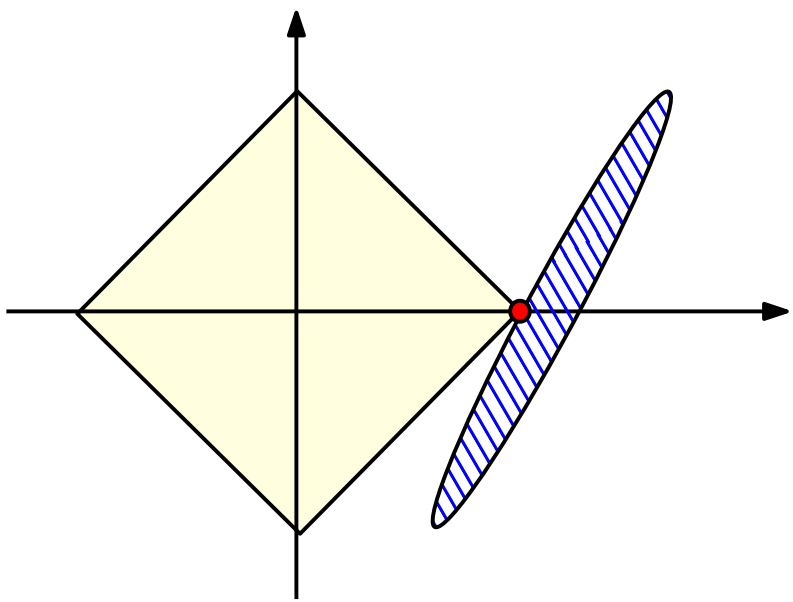


Relaxation: ℓ_1 regularization and LASSO

Replace hopeless problem with relaxation:

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

Illustration:



Problem: How determine λ? (No longer # of parameters)

Choosing regularization parameter λ

Standard approach: Cross-validation

Choosing regularization parameter λ

Standard approach: Cross-validation

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

Choosing regularization parameter λ

Standard approach: Cross-validation

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

Means repeated solution of original problem

Choosing regularization parameter λ

Alternative idea:

- Combine estimation and validation in one shot

Choosing regularization parameter λ

Alternative idea:

- Combine estimation and validation in one shot
- Introduce slack in cost, but allow for same cross-validation fit as ordinary least squares (all parameters estimated)

Choosing regularization parameter λ

Alternative idea:

- Combine estimation and validation in one shot
- Introduce slack in cost, but allow for same cross-validation fit as ordinary least squares (all parameters estimated)

AIC: $$V_N^{val}(\hat{\theta}_N^{LS}) = \left(1 + \frac{2\dim\theta}{N}\right) V_N^{est}(\hat{\theta}_N^{LS})$$

Choosing regularization parameter λ

Alternative idea:

- Combine estimation and validation in one shot
- Introduce slack in cost, but allow for same cross-validation fit as ordinary least squares (all parameters estimated)

AIC: $$V_N^{val}(\hat{\theta}_N^{LS}) = \left(1 + \frac{2\dim\theta}{N}\right) V_N^{est}(\hat{\theta}_N^{LS})$$

SPARSEVA (Version 1.0):

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

where $\varepsilon_N = 2\dim\theta/N$

Outline

SPARSe Estimation based on a VAlidation criterion

SPARSEVA:

SPARSe Estimation based on a VAlidation criterion

SPARSEVA:

i) First compute the ordinary least-squares estimate,

$$\hat{\theta}_N^{LS} = \left(\Phi_N^T \Phi_N\right)^{-1} \Phi_N^T Y_N$$

SPARSe Estimation based on a VAlidation criterion

SPARSEVA:

i) First compute the ordinary least-squares estimate,

$$\hat{\theta}_N^{LS} = \left(\Phi_N^T \Phi_N\right)^{-1} \Phi_N^T Y_N$$

ii) Obtain a sparse estimate $\hat{\theta}_N$ solving

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

SPARSe Estimation based on a VAlidation criterion

SPARSEVA:

i) First compute the ordinary least-squares estimate,

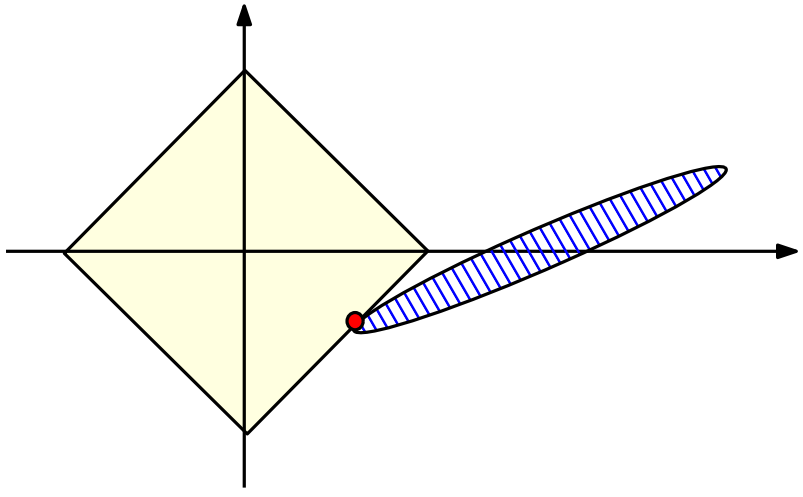
$$\hat{\theta}_N^{LS} = \left(\Phi_N^T \Phi_N\right)^{-1} \Phi_N^T Y_N$$

ii) Obtain a sparse estimate $\hat{\theta}_N$ solving

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

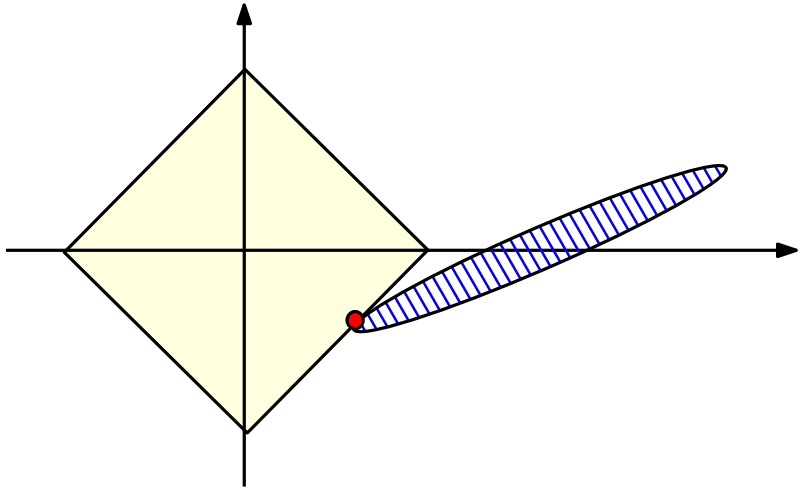
iii) Re-estimate the non-zero elements of $\hat{\theta}_N$ using ordinary least-squares.

A Shortcoming of ℓ_1 optimization



Sparse solution NOT obtained

A Shortcoming of ℓ_1 optimization



Sparse solution NOT obtained

- Shape of level curves of V_N depend on regressors Φ

A-SPARSEVA: Adaptive SPARSEVA

Solution:

$$\begin{array}{ll} \min\limits_{\theta} & \sum_k \left| \frac{\theta_k}{|\hat{\theta}_k^{LS}|^\gamma} \right| \\ \text{s.t.} & (1+\varepsilon_N) V_N(\hat{\theta}_N^{LS}) \geq V_N(\theta), \\ \gamma & > 0 \end{array}$$

Adaptive SPARSEVA

Inspired by *H. Zou. J. Am. Stat. Assoc. 2006.*

Outline

Consistency

Definition

$A_N = O_p(B_N)$ means that, given an $\varepsilon > 0$, there exists a constant $M(\varepsilon) > 0$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{|A_N| \leq M(\varepsilon)|B_N|\} \geq 1 - \varepsilon$.

Consistency

Definition

$A_N = O_p(B_N)$ means that, given an $\varepsilon > 0$, there exists a constant $M(\varepsilon) > 0$ and an $N_0(\varepsilon) \in \mathbb{N}$ such that for every $N \geq N_0(\varepsilon)$, $P\{|A_N| \leq M(\varepsilon)|B_N|\} \geq 1 - \varepsilon$.

Theorem (Consistency of (A-)SPARSEVA)

SPARSEVA and A-SPARSEVA are consistent in probability (i.e., $\hat{\theta}_N \xrightarrow{p} \theta^o$) if and only if $\varepsilon_N \to 0$. In particular, $\|\hat{\theta}_N - \theta^o\|_2 = O_p(N^{-1/2} + \sqrt{\varepsilon_N})$.

Sparsity

Definition

An estimator $\hat{\theta}_N$ is said to have the sparsity property if $\hat{\theta}_N = [(\hat{\theta}_N^1)^T \ (\hat{\theta}_N^2)^T]^T$, with $\hat{\theta}_N^i \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, where $P\{\hat{\theta}_N^2 = 0\} \to 1$ as $N \to \infty$.

Sparsity

Definition

An estimator $\hat{\theta}_N$ is said to have the sparsity property if $\hat{\theta}_N = [(\hat{\theta}_N^1)^T \ (\hat{\theta}_N^2)^T]^T$, with $\hat{\theta}_N^i \in \mathbb{R}^{n_i}$ $(i = 1, 2)$, where $P\{\hat{\theta}_N^2 = 0\} \to 1$ as $N \to \infty$.

Theorem (Sparseness of the adaptive SPARSEVA)

Assume that $\varepsilon_N \to 0$ *and* $\theta^o \neq 0$.
Then, A-SPARSEVA satisfies the sparseness property if $N\varepsilon_N \to \infty$-
If $N\varepsilon_N \to \infty$ *does not hold, A-SPARSEVA does not have the sparseness property.*

The Oracle Property

Definition (Oracle property)

An estimator $\hat{\theta}_N$ is said to have the oracle property if it has the same asymptotic distribution

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \in AsN(0, M^\dagger)$$

as the least-squares oracle, i.e. the least-squares estimator which knows which components of θ_o that are zero.

The Oracle Property

Definition (Oracle property)

An estimator $\hat{\theta}_N$ is said to have the oracle property if it has the same asymptotic distribution

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \in AsN(0, M^\dagger)$$

as the least-squares oracle, i.e. the least-squares estimator which knows which components of θ_o that are zero.

Theorem (The Oracle property)

Assume that $N\varepsilon_N \to \infty$.
Then θ^{A-RE} *has the orcale property.*

The Oracle Property

Definition (Oracle property)

An estimator $\hat{\theta}_N$ is said to have the oracle property if it has the same asymptotic distribution

$$\sqrt{N}(\hat{\theta}_N - \theta^o) \in AsN(0, M^\dagger)$$

as the least-squares oracle, i.e. the least-squares estimator which knows which components of θ_o that are zero.

Theorem (The Oracle property)

Assume that $N\varepsilon_N \to \infty$.
Then θ^{A-RE} *has the orcale property.*

- OBS: Convergence not uniform. See discussions by Leeb and Pötscher

Outline

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T .$$

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

Comparison of the following estimators:

- LS-oracle (the benchmark)

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

Comparison of the following estimators:

- LS-oracle (the benchmark)
- LASSO with regularization chosen using generalized cross-validation

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

Comparison of the following estimators:

- LS-oracle (the benchmark)
- LASSO with regularization chosen using generalized cross-validation
- SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

Comparison of the following estimators:

- LS-oracle (the benchmark)
- LASSO with regularization chosen using generalized cross-validation
- SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$
- SPARSEVA-BIC-RE: $\varepsilon_N = n \log N/N$

Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

Comparison of the following estimators:

- LS-oracle (the benchmark)
- LASSO with regularization chosen using generalized cross-validation
- SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$
- SPARSEVA-BIC-RE: $\varepsilon_N = n \log N/N$
- A-SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$

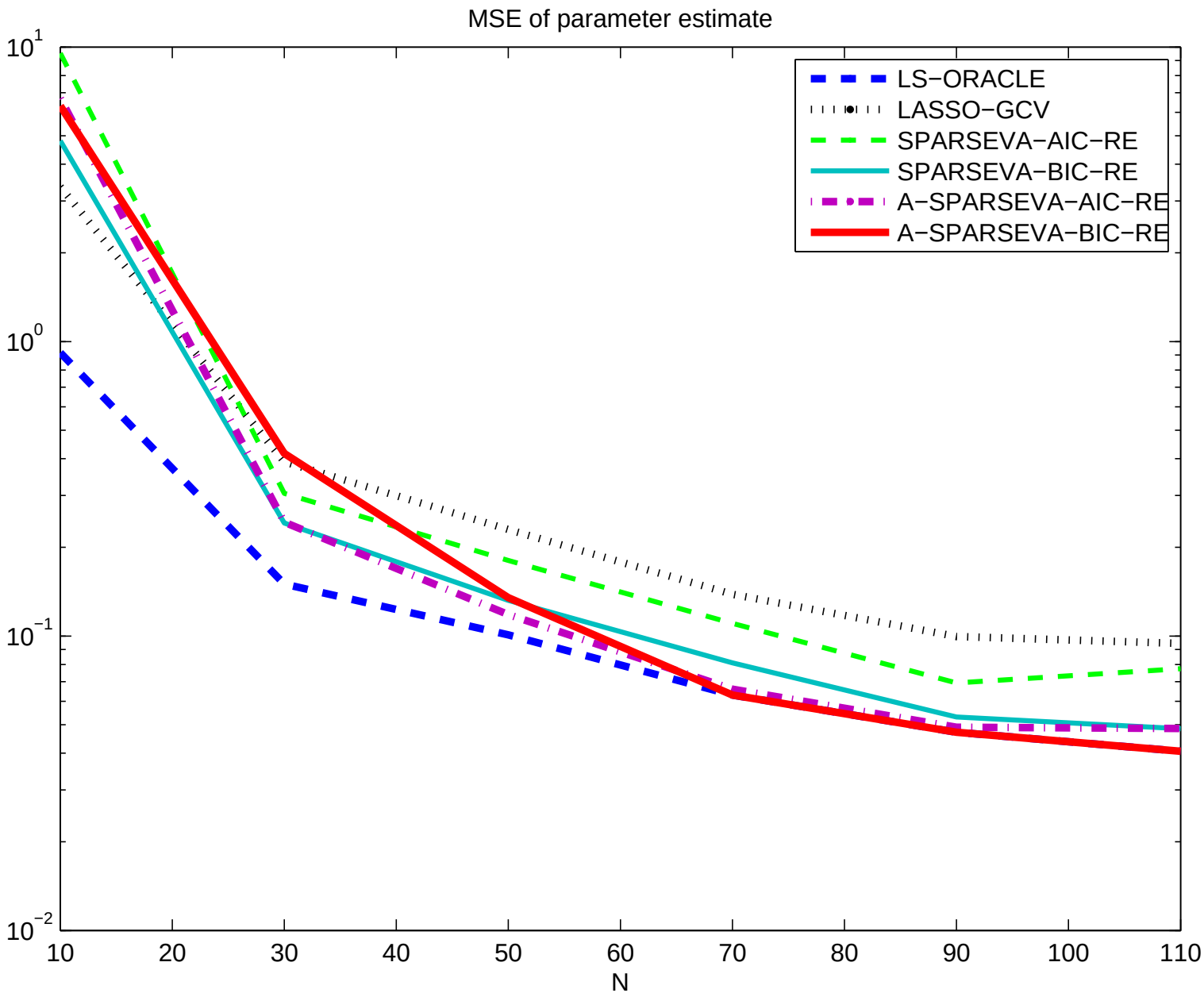
Example

$$\theta^o = \begin{bmatrix} 3 & 1.5 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}^T.$$

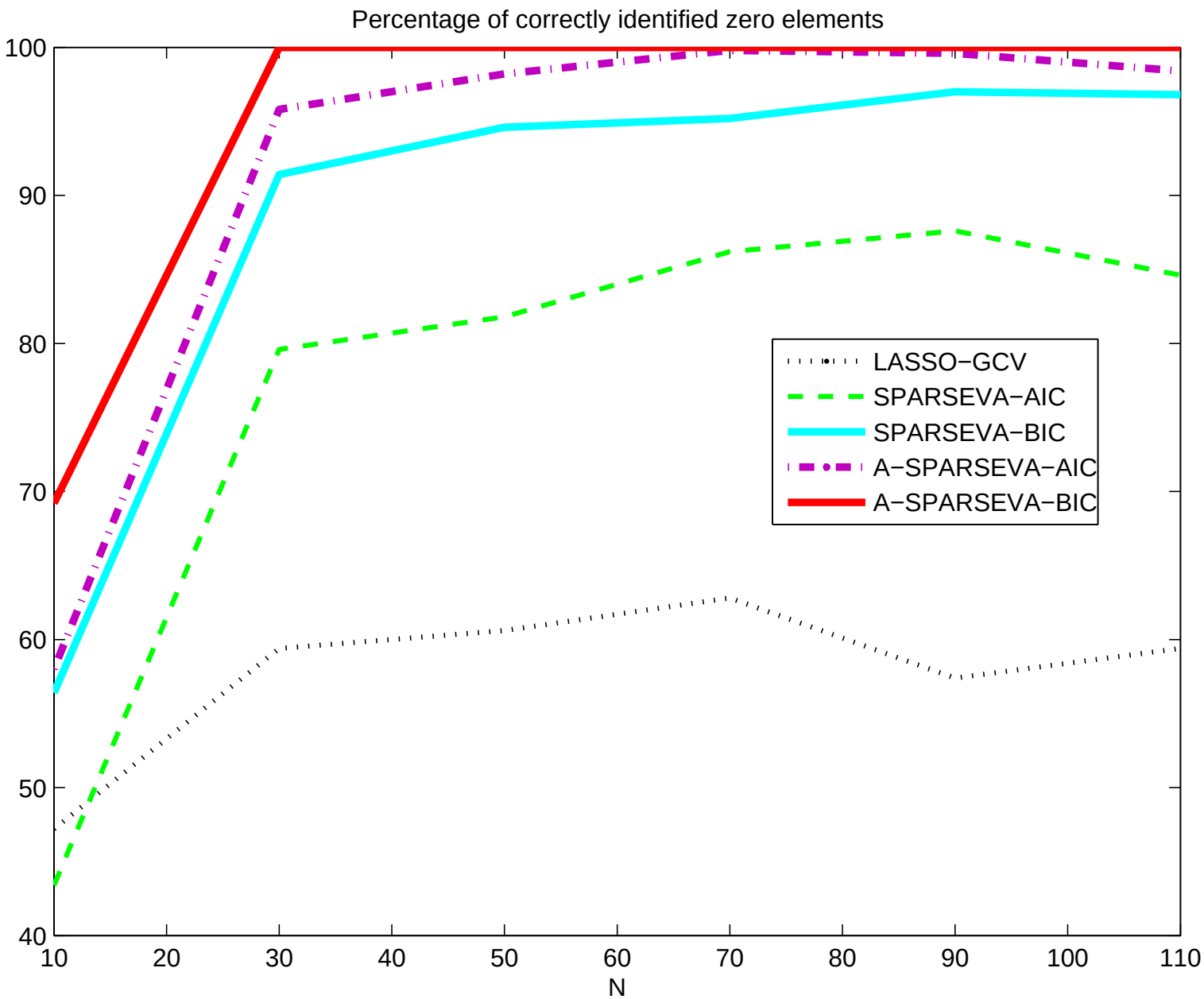
Comparison of the following estimators:

- LS-oracle (the benchmark)
- LASSO with regularization chosen using generalized cross-validation
- SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$
- SPARSEVA-BIC-RE: $\varepsilon_N = n \log N/N$
- A-SPARSEVA-AIC-RE: $\varepsilon_N = 2n/N$
- A-SPARSEVA-BIC-RE: $\varepsilon_N = n \log N/N$

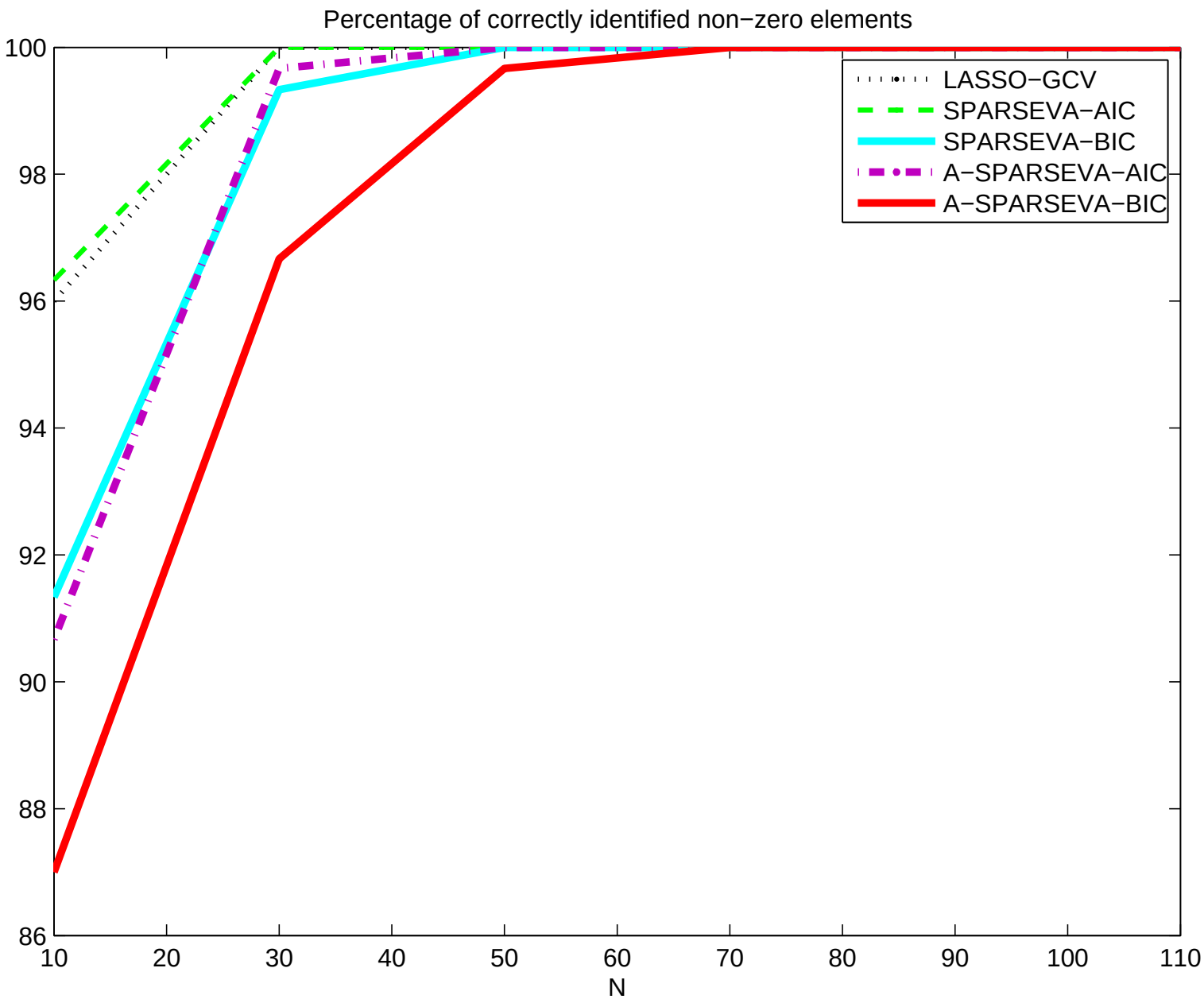
MSE as a function of the sample size N



Percentage of correctly identified zero elements.



Percentage of correctly identified non-zero elements.



Outline

Conclusions

- Automatic selection of regularization parameter in ℓ_1 regularization
- Mimicks AIC/BIC in one-shot
- Consistency
- Sparsity property
- Oracle property
- Idea can be applied to other norms as well