An Efficient Optimization Approach for Subset Selection, with Application to Linear Regression and Auto-Regressive Time Series

ODS Genova, 6th September 2019

L. Di Gangi, **M. Lapucci**, F. Schoen, A. Sortino

DINFO, Università degli Studi di Firenze

The Two (Related) Problems

- Linear Regression
  - Given a dataset $X = [x_1, \ldots, x_P] \in \mathbb{R}^{N \times P}$
  - and a vector $Y \in \mathbb{R}^N$ of response variables
  - find a vector of parameters $\beta \in \mathbb{R}^P$ and a value c such that

$$y_i = \sum_{j=1}^{P} \beta_j x_{ij} + c + \varepsilon_i, \qquad i = 1, \ldots, N,$$

being $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

The Two (Related) Problems

- Linear Regression
  - Given a dataset $X = [x_1, \ldots, x_P] \in \mathbb{R}^{N \times P}$
  - and a vector $Y \in \mathbb{R}^N$ of response variables
  - find a vector of parameters $\beta \in \mathbb{R}^P$ and a value c such that

$$y_i = \sum_{j=1}^{P} \beta_j x_{ij} + c + \varepsilon_i, \qquad i = 1, \ldots, N,$$

  being $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.
- Autoregressive Time-Series Model Selection and Fitting
  - Given a time-series $\{X_t\}$
  - find an order P, a vector of parameters $\beta \in \mathbb{R}^P$ and a value c such that

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_P X_{t-P} + \varepsilon_t,$$

  being $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Best Subset Selection in Linear Regression

$$\min_{c \in \mathbb{R},\ \beta \in \mathbb{R}^P} \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2$$

$$\text{s.t.} \quad \|\beta\|_0 \text{ is reasonably small}$$

Best Subset Selection in Linear Regression

$$\min_{c \in \mathbb{R},\ \beta \in \mathbb{R}^P} \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2$$

$$\text{s.t.} \quad \|\beta\|_0 \text{ is reasonably small}$$

Ok, but what does **reasonably small** mean?

Goodness-Of-Fit Measures

- **AIC:** $-2\ell(\beta, c, \sigma^2) + 2(\|\beta\|_0 + 2)$
- **BIC:** $-2\ell(\beta, c, \sigma^2) + \log(N)(\|\beta\|_0 + 2)$
- **HQIC:** $-2\ell(\beta, c, \sigma^2) + 2(\|\beta\|_0 + 2)\log(\log N)$

$\ell(\beta, c, \sigma^2)$ is the log-likelihood of the linear regression model:

$$-2\ell(\beta, c, \sigma^2) = N\log(\sigma^2) + N\log(2\pi) + \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - c - \sum_{j=1}^{P}\beta_j x_{ij}\right)^2$$

Sparse Linear Regressions with GOF Measures

$$\min_{\beta,\, c,\, \sigma^2} N \log(\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2 + \alpha \|\beta\|_0$$

State-of-the-art-approaches: Step 1

$$\min_{\beta,\, c,\, \sigma^2} N \log(\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2 + \alpha \|\beta\|_0$$

- Substitute σ^2 with the maximum-likelihood estimator

$$\sigma^2 = \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2 / N$$

to obtain

$$\min_{\beta,\, c} \; N \log \left(\frac{\sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2}{N} \right) + \alpha \|\beta\|_0$$

State-of-the-art approaches: Step 2

$$\min_{\beta,\, c} \; N \log \left(\frac{\sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2}{N} \right) + \alpha \|\beta\|_0$$

- Enumerate exhaustively variable subsets and solve 2^P continuous least squares problems.
- Employ step-wise heuristics instead of exhaustive enumeration.
- Solve an equivalent MIQP model for

$$\min_{\substack{\beta,\, c \\ \|\beta\|_0 \leq k}} \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2$$

  for all $k = 1, \ldots, P$ and then compare GOF measures of the solutions.
- Solve an overall MINLP model.

Our Proposal: An Alternate Minimization Approach

$$\min_{\beta, c, \sigma^2} \quad N \log(\sigma^2) + \frac{1}{\sigma^2} R(\beta, c) + g(\beta)$$
$$\text{s.t.} \;\; \sigma^2 > 0, \quad c \in \mathbb{R}, \quad \beta \in \mathbb{R}^P$$

- $g(\beta) = \alpha \|\beta\|_0$
- $R(\beta, c) = \sum_{i=1}^{N} \left(y_i - c - \sum_{j=1}^{P} \beta_j x_{ij} \right)^2$

Our Proposal: An Alternate Minimization Approach

$$\min_{\beta, c, \sigma^2} \quad N \log(\sigma^2) + \frac{1}{\sigma^2} R(\beta, c) + g(\beta)$$

$$\text{s.t.} \;\; \sigma^2 > 0, \quad c \in \mathbb{R}, \quad \beta \in \mathbb{R}^P$$

Algorithm 1 Alternate Minimization (AM)

Input: $\beta^0, \; c^0, \; \sigma_0, \; k = 0$

let $g(\beta^{-1}) = \texttt{NaN}$
while $g(\beta^k) \neq g(\beta^{k-1})$ **do**
    set $\beta^{k+1}, \; c^{k+1} = \arg\min_{\beta, \, c} \frac{R(\beta, c)}{{\sigma_k}^2} + g(\beta)$
    set $\sigma^2_{k+1} = \arg\min_{\sigma^2 > 0} N \log(\sigma^2) + \frac{R(\beta^{k+1}, \, c^{k+1})}{\sigma^2}$
    set $k = k + 1$
end while
return $\beta^k, \; c^k, \; \sigma_k$

Handling the β-update Step

The (β, c)-update subproblem

$$\beta^{k+1},\ c^{k+1} = \arg\min_{\beta,\ c} \frac{R(\beta, c)}{{\sigma_k}^2} + \alpha\|\beta\|_0$$

can be reformulated, and thus solved, as a MIQP problem:

$$\begin{aligned} \min_{\beta, c, \delta}\ & R(\beta,\ c) + \alpha\sigma_k^2 \sum_{i=1}^{P} \delta_i \\ \text{s.t.}\ & \beta \in \mathbb{R}^P, \quad c \in \mathbb{R}, \quad \delta \in \{0, 1\}^P, \\ & \beta_i \neq 0 \Rightarrow \delta_i = 1 \quad \forall\, i = 1, \ldots, P. \end{aligned}$$

Convergence Properties of AM

Proposition

Let $\{\beta^k\}$ be the sequence of iterates produced by AM and let $g^k = g(\beta^k)$. The following properties hold:

- a *For each iteration k, either $g^k = g^{k-1}$ and the algorithm terminates, or $g^k \neq g^h$ for all $h < k$.*
- b *The algorithm terminates in at most $P + 1$ iterations, returning a solution $(\bar{\beta}, \bar{\sigma})$.*
- c *Let $\bar{k}$ be the index of the last iteration. If there exists $\beta^\star$ s.t. $f(\beta^\star) < f(\bar{\beta})$, then $g(\beta^\star) \notin \{g^1, \ldots, g^{\bar{k}}\}$.*
- d *If $\bar{k} = P + 1$, then the returned solution $\bar{\beta}$ is optimal.*
- e *Let the pair $\beta^\star$, $\sigma^\star = R(\beta^\star)/N$ be optimal for the considered problem. Then, the following bound holds:*

$$0 \leq f(\bar{\beta}) - f(\beta^\star) \leq -N \log(1 - \eta^2 \exp(\theta - 1)),$$

  where $\theta \in (0, 1)$ and $\eta = (g(\bar{\beta}) - g(\beta^\star))/N$.

AR Time-Series Model Selection and Fitting Problem

Fit $X_t \approx c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_P X_{t-P} + \varepsilon_t$, based on GOF measures. Problem similar to the Sparse Linear Regression case:

$$\min_{\varphi,\, c,\, \sigma^2} \; -2\ell(\varphi,\, c,\, \sigma^2) + \alpha\, g(\varphi)$$

AR Time-Series Model Selection and Fitting Problem

Fit $X_t \approx c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_P X_{t-P} + \varepsilon_t$, based on GOF measures. Problem similar to the Sparse Linear Regression case:

$$\min_{\varphi,\, c,\, \sigma^2} \; -2\ell(\varphi,\, c,\, \sigma^2) + \alpha\, g(\varphi)$$

BUT:

- The exact log-likelihood function is more complicated

$$-2\ell(\varphi, c, \sigma^2) = N\log(2\pi) + N\log(\sigma^2) - \log|V_p^{-1}| + \\ + \frac{1}{\sigma^2}(\bar{X}_p - \mu_p)^T V_p^{-1}(\bar{X}_p - \mu_p) + \frac{1}{\sigma^2}\sum_{t=p+1}^{N}\left(X_t - c - \sum_{i=1}^{p}\varphi_i X_{t-i}\right)^2$$

AR Time-Series Model Selection and Fitting Problem

Fit $X_t \approx c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_P X_{t-P} + \varepsilon_t$, based on GOF measures. Problem similar to the Sparse Linear Regression case:

$$\min_{\varphi,\, c,\, \sigma^2} \; -2\ell(\varphi,\, c,\, \sigma^2) + \alpha\, g(\varphi)$$

BUT:

- The exact log-likelihood function is more complicated

$$-2\ell(\varphi, c, \sigma^2) = N\log(2\pi) + N\log(\sigma^2) - \log|V_p^{-1}| + \\ + \frac{1}{\sigma^2}(\bar{X}_p - \mu_p)^T V_p^{-1}(\bar{X}_p - \mu_p) + \frac{1}{\sigma^2}\sum_{t=p+1}^{N}\left(X_t - c - \sum_{i=1}^{p}\varphi_i X_{t-i}\right)^2$$

- $g(\varphi) = \mathrm{ord}(\varphi) = \min\{j = 0, \ldots, P \mid \varphi_h = 0 \;\forall h > j\}$.

AR Time-Series Model Selection and Fitting Problem

Fit $X_t \approx c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_P X_{t-P} + \varepsilon_t$, based on GOF measures. Problem similar to the Sparse Linear Regression case:

$$\min_{\varphi,\, c,\, \sigma^2} \; -2\ell(\varphi,\, c,\, \sigma^2) + \alpha\, g(\varphi)$$

BUT:

- The exact log-likelihood function is more complicated

$$-2\ell(\varphi, c, \sigma^2) = N \log(2\pi) + N \log(\sigma^2) - \log |V_p^{-1}| + \\ + \frac{1}{\sigma^2}(\bar{X}_p - \mu_p)^T V_p^{-1} (\bar{X}_p - \mu_p) + \frac{1}{\sigma^2} \sum_{t=p+1}^{N} \left(X_t - c - \sum_{i=1}^{p} \varphi_i X_{t-i} \right)^2$$

- $g(\varphi) = \mathrm{ord}(\varphi) = \min\{j = 0, \ldots, P \mid \varphi_h = 0 \;\forall h > j\}$.
- φ has to satisfy stationarity constraints, i.e. the roots of

$$\pi_\varphi(z) = 1 - \varphi_1 z - \varphi_2 z^2 - \cdots - \varphi_p z^p = 0$$

  should lie outside the unit complex circle.

Adapting AM to the AR Case

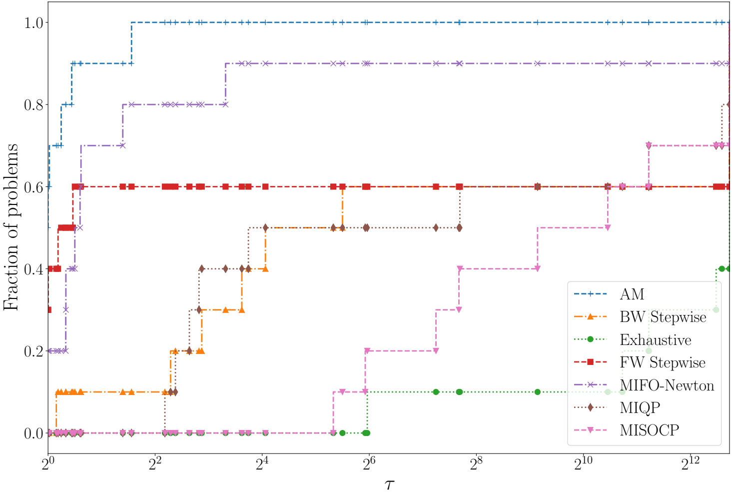
- Set $g(\varphi) = \mathrm{ord}(\varphi)$.
- Approximate the exact log-likelihood with the conditional log-likelihood (same form as LR).
- Perform a refinement step of the retrieved solution $\bar{\varphi}$ at the end of the process.
    - Fix $g(\varphi) = g(\bar{\varphi})$.
    - Start at $\bar{\varphi}$.
    - Local optimization of the exact log-likelihood.
- Insert closed form constraints to enforce stationarity of solutions of order 1 and 2.
- If at some step the new iterate is not stationary, reject it and add a constraint to the model to prohibit solutions with the same order of the rejected solution.

Numerical Experiments: Linear Regression

- 10 standard data sets;
- AIC/BIC/HQIC;
- a total of 30 problems;
- time limit of 10000 seconds for each run.

Method	# optimal	total time (sec)
AM	**29/30**	**31299**
BW Stepwise	18/30	38666
Exhaustive	17/30	267405
FW Stepwise	17/30	3530
MIFO-Newton	29/30	32759
MIQP	25/30	155295
MISOCP	22/30	172215

Performance profiles of runtime - AIC



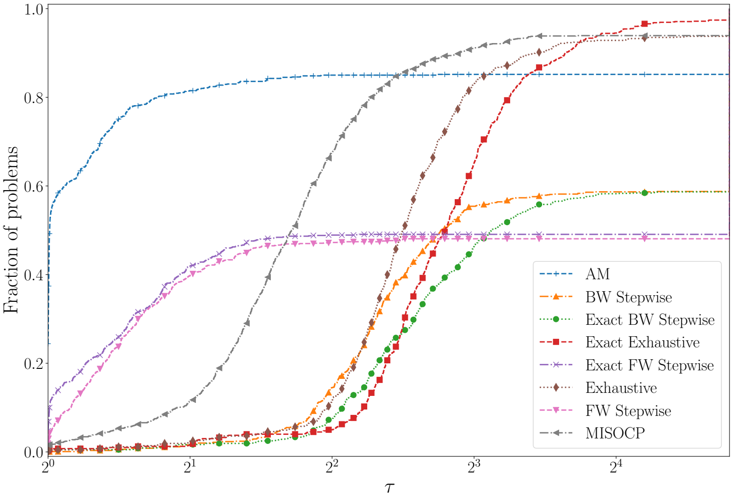
Numerical Experiments: Autoregression

- 10 real time series;
- AIC/BIC/HQIC;
- a total of 30 problems.

Method	# optimal	total time (sec)
AM	**27/30**	**75**
BW Stepwise	18/30	134
Exact BW Stepwise	15/30	113
Exact Exhaustive	30/30	273
Exact FW Stepwise	21/30	121
Exhaustive	29/30	279
FW Stepwise	21/30	130
MISOCP	29/30	2513

Numerical Experiments: Autoregression

Performance profiles of runtime - 1200 synthetic series



Wrap Up

- Simple AM method, exploiting MIQP solvers, for GOF-based sparse linear regression.
- Well understood convergence properties.
- Effective and efficient on LR problems, compared to the state-of-the-art.
- Suitable for automatic AR time-series model selection and fitting.

An Efficient Optimization Approach for Subset Selection, with Application to Linear Regression and Auto-Regressive Time Series

ODS Genova, 6th September 2019

L. Di Gangi, **M. Lapucci**, F. Schoen, A. Sortino

DINFO, Università degli Studi di Firenze