An Efficient Optimization Approach for Best Subset Selection in Linear Regression

PhDINFO Seminar
23th January 2020

Alessio Sortino

The Two (Related) Problems

Linear regression problem

Given a dataset $X \in \mathbb{R}^{N \times P}$ and a vector $Y \in \mathbb{R}^N$ of response variables, we want to find a vector of parameters $\beta \in \mathbb{R}^P$ and a value $c \in \mathbb{R}$ such that

$$y_i = \sum_{j=1}^{P} \beta_j x_{ij} + c + \varepsilon_i$$

for each $i = 1, \dots, N$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

Autoregressive Time-Series Model Selection and Fitting

Given a time-series $\{X_t\}$ we want to find an order P, a vector of parameters $\beta \in \mathbb{R}^P$ and a value $c \in \mathbb{R}$ such that

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

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

Best Subset Selection in Linear Regression

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

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

Goodness-Of-Fit Measures

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

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

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

Sparse Linear Regressions with GOF Measures

The problem becomes

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

where $\alpha \in \mathbb{R}$ depends on the GOF

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

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

  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

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

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

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

1: let $g(\beta^{-1}) = \texttt{NaN}$
2: **while** $g(\beta^k) \neq g(\beta^{k-1})$ **do**
3: set $\beta^{k+1},\ c^{k+1} = \underset{\beta,\ c}{\arg\min}\ \frac{R(\beta,c)}{{\sigma_k}^2} + g(\beta)$
4: set $\sigma^2_{k+1} = \underset{\sigma^2>0}{\arg\min}\ N\log(\sigma^2) + \frac{R(\beta^{k+1},\ c^{k+1})}{\sigma^2}$
5: set $k = k + 1$
6: **end while**
7: **return** $\beta^k,\ c^k,\ \sigma_k$

Handling the β-update Step

The (β, c)-update subproblem

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

can be reformulated as a MIQP problem:

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

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 $g(\beta^\star) \notin \{g^1,\dots,g^{\bar k}\}$*

(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:*

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

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

AR Time-Series Model Selection and Fitting Problem

We want to 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. The problem is 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

We want to 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. The problem is 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

We want to 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. The problem is 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

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

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

AR Time-Series Model Selection and Fitting Problem

We want to 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. The problem is 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

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

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

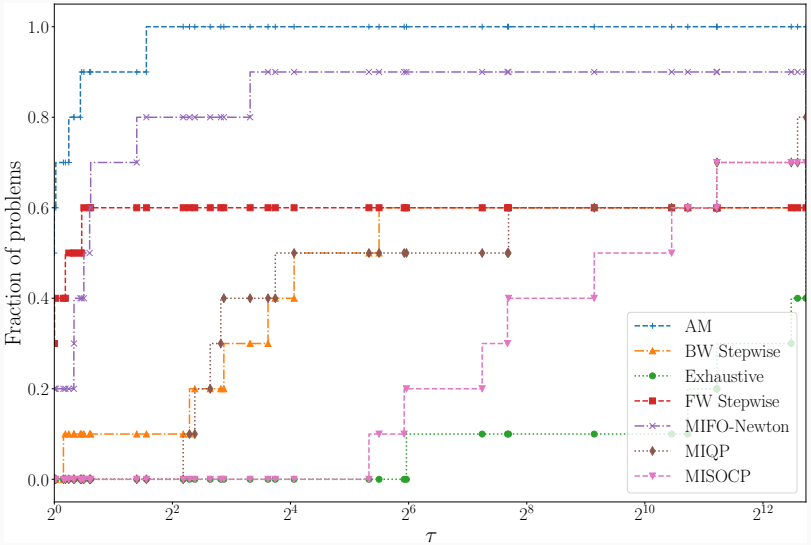
Numerical Experiments: Linear Regression

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

Method	# optimal	CPU 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

Numerical Experiments: Linear Regression

Performance profiles of runtime - AIC



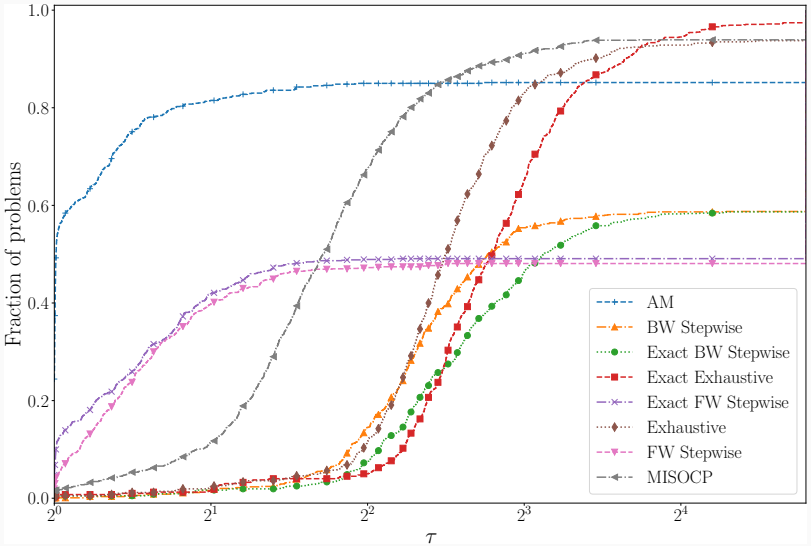
Numerical Experiments: Autoregression

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

Method	# optimal	CPU time (s)
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 on 1200 synthetic series



Extension #1: Sparse Logistic Regression

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w) + \lambda \|w\|_0$$

where:

- $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function,
- $\lambda \in \mathbb{R}_+$ is a penalty parameter,
- $\|\cdot\|_0$ is the ℓ_0 semi-norm of a vector.

Extension #1: Sparse Logistic Regression

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w) + \lambda \|w\|_0$$

where:

- $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function,
- $\lambda \in \mathbb{R}_+$ is a penalty parameter,
- $\|\cdot\|_0$ is the ℓ_0 semi-norm of a vector.

$\mathcal{NP}$-hard problem!

Extension #1: Sparse Logistic Regression

A large number of algorithm are present in the literature to solve this problem:

- Exhaustive approach
- Heuristics (forward and backward)
- Lasso (replacing ℓ_0 semi-norm with ℓ_1 norm)
- Penalty decomposition (duplicating the variable w)
- Concave approximation
- Piecewise linear approximation
- Greedy Sparse Simplex

Extension #1: Sparse Logistic Regression

- Greedy Sparse Simplex algorithm aims to solve problems of the form

$$\min_{w \in \mathbb{R}^n} \mathcal{L}(w)$$
$$\text{s.t.} \quad \|w\|_0 \leq s$$

- The idea we are working on is to generalize the GSS idea, in order to solve also our unconstrained problem

Extension #2: ARMA models

- Another interesting goal is to estimate the Moving Average (MA) part, i.e. a full ARMA model

$$X_t = c + \sum_{i=1}^{p} X_{t-i}\varphi_i + \sum_{j=1}^{q} \varepsilon_{t-j}\theta_j + \varepsilon_t$$

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

- For each pair (p, q), a model is fitted by maximum likelihood estimation. This is a nonconvex problem and therefore, in order to solve it, global optimization techniques are required.

Conclusions

- 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

Thank you for your attention

GOL

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

An Efficient Optimization Approach for Best Subset Selection in Linear Regression

PhDINFO Seminar
23th January 2020

Alessio Sortino