APES: Approximated Exhaustive Search for GLM

Kevin Y.X. Wang

School of Mathematics and Statistics
The University of Sydney
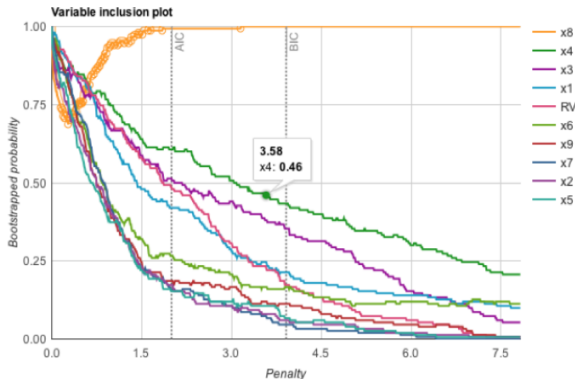
Peking University - Beijing
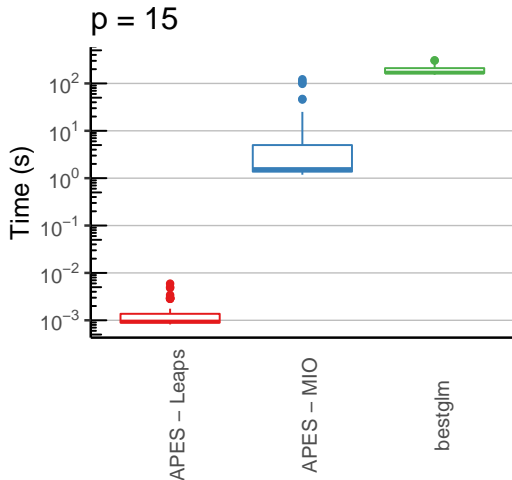29 December 2019

THE UNIVERSITY OF
SYDNEY

@KevinWang009

Acknowledgement

- This is joint work with Prof Samuel Müller, Dr Garth Tarr and Prof Jean Yang.
- `mplot` Tarr et al. (2018) is a package to assess model stability and variable selection for linear models and generalised linear models.

My talk in one slide

- Exhaustive variable selection for GLM is slow.
- APES improves the speed of exhaustive search using approximations.
- APES is now published at Australian and New Zealand Journal of Statistics: `https://doi.org/10.1111/anzs.12276`.

Background

Data and models

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$, with independent $y_i \in \{0, 1\}$, $i = 1, \ldots, n$.

- Design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$.

- Index the columns of $\boldsymbol{X}$ by $\{1, \ldots, p\}$.

- Let α denote any subset of p_α distinct elements from $\{1, \ldots, p\}$. Use $\mathcal{A}$ to denote the collection of all α, so $|\mathcal{A}| = 2^p$.

- $\boldsymbol{X}_\alpha$ denote the $n \times p_\alpha$ matrix with columns given by the columns of $\boldsymbol{X}$ whose indices appear in α.

Logistic regression

- We model the conditional response variable $Y_i|\boldsymbol{X}$ as $\mathrm{Bernoulli}(\pi_i)$, where $\pi_i = \mathbb{P}(Y_i = 1|\boldsymbol{X})$.

- We will use the **logistic function** as our link function, so $\boldsymbol{x}_i^\top \boldsymbol{\beta} = \mathrm{logit}(\pi_i) = \ln(\pi_i/(1 - \pi_i))$.

- Model fitting usually involves estimating $\boldsymbol{\beta}$ (or equivalently, $\boldsymbol{\pi}$).

Iterative Reweighted Least Square (IRLS)

1. Denote weights $w_i = \pi_i(1 - \pi_i)$ and other estimates at the t-th iteration with a superscript (t).

2. Construct

$$z_i^{(t)} = \underbrace{\text{logit}\left(\pi_i^{(t)}\right)}_{\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)}} + \frac{y_i - \pi_i^{(t)}}{\pi_i^{(t)}(1 - \pi_i^{(t)})}.$$

3. Update via

$$\widehat{\boldsymbol{\beta}}^{(t+1)} \leftarrow (\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)},$$

with $\boldsymbol{W}^{(t)} = \text{diag}\left(w_i^{(t)}\right)$.

4. At convergence, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z}$, which equals to the MLE of logistic regression model.

Challenges in exhaustive GLM variable selection

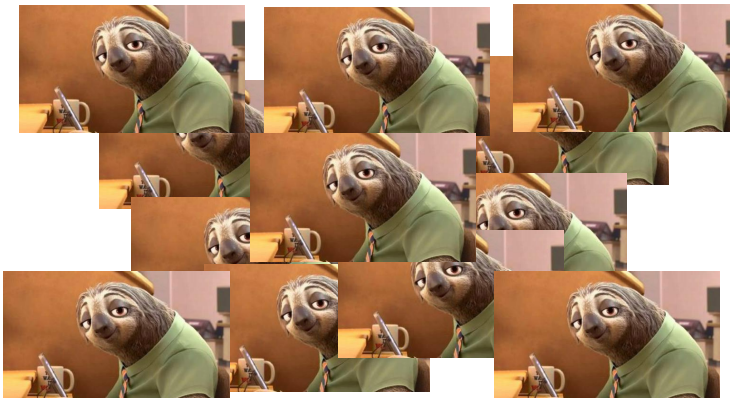For large p, exhaustive variable selection in GLM is difficult:

1. The computational cost of IRLS is $\mathcal{O}(np^2)$ per iteration.
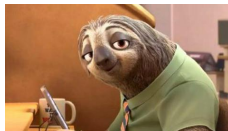
Challenges in exhaustive GLM variable selection

For large p, exhaustive variable selection in GLM is difficult:

1. The computational cost of IRLS is $\mathcal{O}(np^2)$ per iteration.
2. We need to explore all 2^p models.

Aim of APES

Can we perform linear exhaustive variable selection (which benefits from fast algorithms) and use the results to approximate exhaustive GLM variable selection?

GLM exhaustive ←—— Can this be done? —— LM exhaustive

1. Turning GLM to LM

Exhaustively computing modified MLE

- Hosmer et al. (1989) described an approximation to exhaustive variable selection for logistic regression without the need for numerical optimisation.

- Their method starts with the estimated probability $\widehat{\boldsymbol{\pi}}(\alpha_f)$, from the **full** logistic model.

- Then, for each model $\alpha \in \mathcal{A}$, we calculate:

$$\widehat{\boldsymbol{\beta}}(\alpha; \widehat{\boldsymbol{\pi}}(\alpha_f)) = (\boldsymbol{X}_\alpha^\top \boldsymbol{W}(\widehat{\boldsymbol{\pi}}(\alpha_f)) \boldsymbol{X}_\alpha)^{-1} \boldsymbol{X}_\alpha^\top \boldsymbol{W}(\widehat{\boldsymbol{\pi}}(\alpha_f)) \boldsymbol{z}(\widehat{\boldsymbol{\pi}}(\alpha_f)).$$

- This is **NOT** the MLE for α, which should be $\widehat{\boldsymbol{\beta}}(\alpha; \widehat{\boldsymbol{\pi}}(\alpha))$.

Variable selection using the modified estimator

- Given $\widehat{\boldsymbol{\beta}}(\alpha; \widehat{\boldsymbol{\pi}}(\alpha_f))$, we could **approximate** RSS or BIC for all $\alpha \in \mathcal{A}$.

- Upon selection of a small set of desired models, we can recompute the MLE and calculate other model fit statistics.

$$\text{GLM exhaustive} \xleftarrow{\text{Hosmer approx.}} \text{LM exhaustive}$$

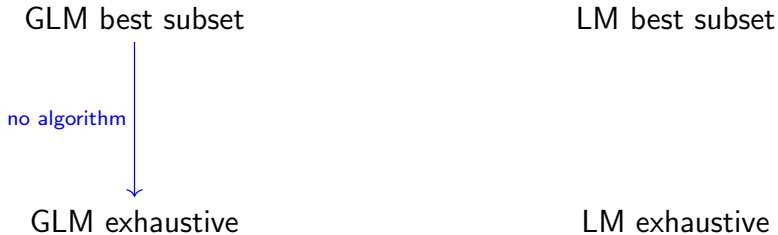# 2. Reducing computational time

Best subsets search

- Application of this approximation method is limited by the number of LMs we can explore.
- For $p \approx 50$, $\mathcal{A}$ is approximately 1 quadrillion in size, which is too large to explore exhaustively.
- A <u>best subset</u> algorithm limits our search to a subset of $\mathcal{A}$ but it guarantees to contain the global RSS-optimal model.
- **leaps** (Furnival and Wilson, 1974; Lumley and Miller, 2017) discard "branches" of models with insufficient fit.
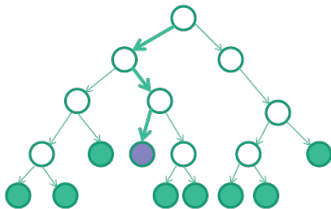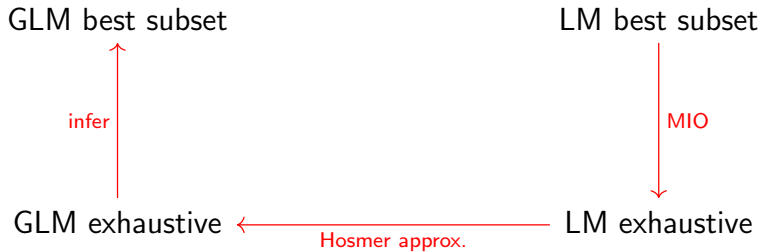
Mixed Integer Optimisation

- Bertsimas et al. (2016) showed that it is feasible to perform best subset search for linear models with p in the hundreds.
- The most attractive component: guaranteed **sub-optimality** if algorithm is terminated before full convergence. Thus allowing a upper bound for real time limit.
- Current implementation in R is bestsubset, (Hastie et al., 2017), which outputs the RSS-best linear model for each model size.
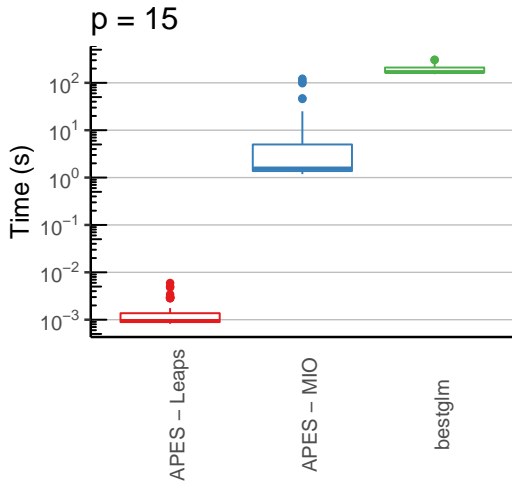
APES: Approximated Exhaustive Search

GLM best subset   LM best subset

no algorithm ↓

GLM exhaustive   LM exhaustive

APES: Approximated Exhaustive Search

GLM best subset   LM best subset

↑ infer   ↓ MIO

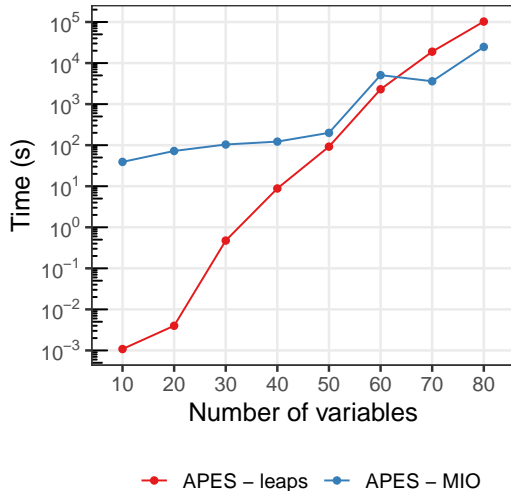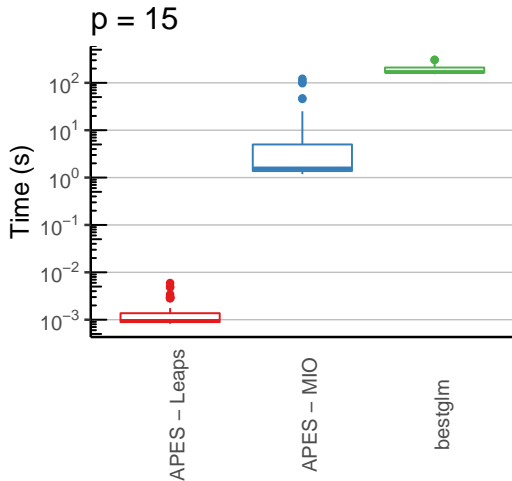GLM exhaustive ← Hosmer approx. LM exhaustive

How fast is APES compare to genuine exhaustive search?

Very.

How fast is APES compare to genuine exhaustive search?

Very.

Simulation

Simulation set-up

- $n = 500, p = 100$, number of non-zero coefficient is $k = 6$.
- Intercept term is set to 0, then we tried 3 different choices of $\boldsymbol{\beta}$:
    1. Equally spaced indices:
       $(\frac{1}{2}, 0, \ldots, 0, \frac{1}{2}, 0, \ldots, 0, \frac{1}{2}, 0, \ldots, 0, \frac{1}{2}, 0, \ldots, 0, \frac{1}{2}, 0, \ldots, \frac{1}{2})$.
    2. Block of indices, same magnitude/sign: $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$.
    3. Block of indices, different magnitude/sign: $(\frac{1}{3}, -1, 1, \frac{2}{3}, -\frac{2}{3}, -\frac{1}{3}, 0, \ldots, 0)$.
- Generating $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.5$ or -0.5. Then standardise.
- We repeated the simulation 100 times and compared APES against de-biased Lasso using various evaluation metrics.

Evaluation 1: saturated models

- For each model size, APES and Lasso outputs one model.
- In most cases, APES has less false exclusion of variables than Lasso.

Number of saturated models by each methods

Lasso ▬ APES
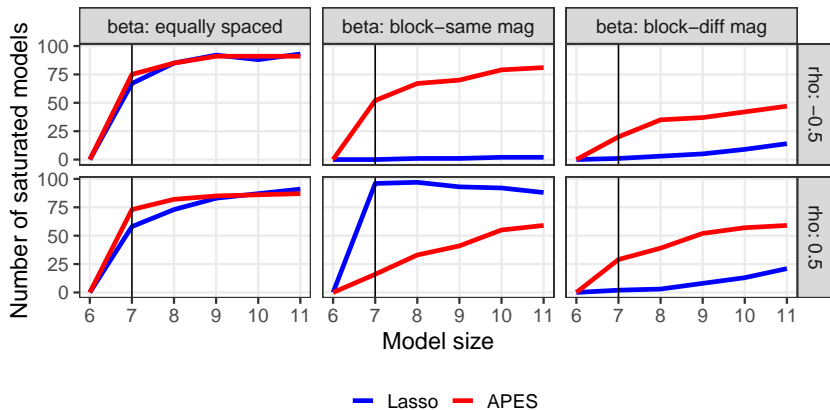
Evaluation 2: variable selection
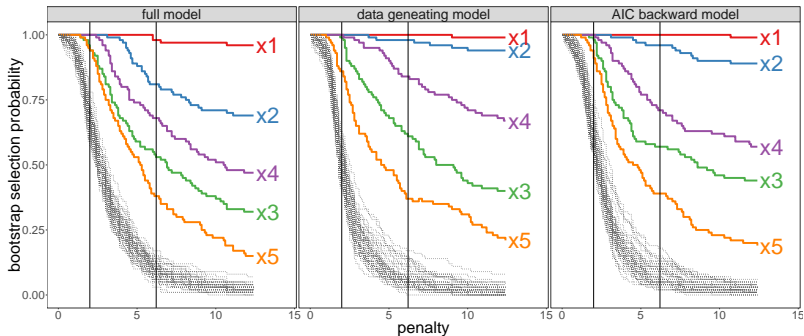
- BIC was used to select an optimal model.
- In most cases, APES has higher selection probability of active variables than Lasso.

Empirical selection probability over 100 simulation runs
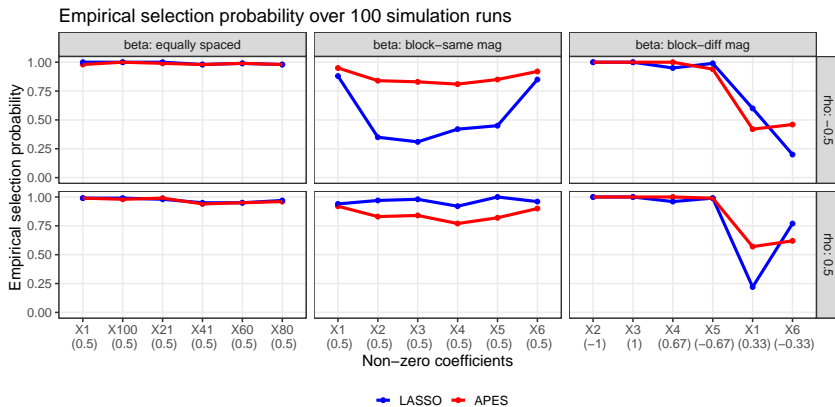
Some extensions

The choice of $\widehat{\boldsymbol{\pi}}(\alpha_f)$ can be relaxed in two different ways:

- The full model α_f is not necessarily the best for variable selection.

- We could replace the MLE by other estimators, e.g. the Lasso or robust quasi-likelihood estimator.

Software

- APES is available on GitHub: `https://github.com/kevinwang09/APES`.
- Example:
  `https://kevinwang09.github.io/APES/reference/plot_vi.html`.

Final remarks

1. APES is a **fast approximation** method for **exhaustive** variable selection in GLM.
2. APES pushes model dimensions into the hundreds and serves as a standard of comparison like a true exhaustive search.
3. APES is now published at ANZJS:
   `https://doi.org/10.1111/anzs.12276`.
   - https://github.com/kevinwang09/APES
   - https://github.com/garthtarr/mplot
   - Email: kevin.wang@sydney.edu.au.
   - Twitter: @KevinWang009

References

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016. ISSN 00905364. doi: 10.1214/15-AOS1388.

G. Furnival and R. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

T. Hastie, R. R. J. Tibshirani, and R. R. J. Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the Lasso following "Best subset selection from a modern optimization lens" by Bertsimas, King, and Mazumder (2016). *arXiv e-prints*, pages 1–52, 2017.

D. W. Hosmer, B. Jovanovic, and S. Lemeshow. Best subsets logistic regression. *Biometrics*, 45(4):1265–1270, 1989.

T. Lumley and A. Miller. leaps: regression subset selection, 2017. URL `https://cran.r-project.org/package=leaps`.

G. Tarr, S. Müller, and A. H. Welsh. mplot: An R package for graphical model stability and variable selection procedures. *Journal of Statistical Software*, 83(9):1–28, 2018.