MATH1332: Minor Thesis Presentation

Identifying Optimal Set of Pairwise Interaction Terms by SP-FSR Algorithm and Empirical Comparison with Other Methods

Yong Kai Wong

30 May 2018

Background

- A motivating example: do gene interactions help predict cancer type?
- How to determine interactions in high dimensions *optimally*?
- Focus on pairwise interaction terms (no quadratic terms)
- Given p features with 2-way pairwise interaction terms, the number of possible features is

$$p + \binom{p}{2} = \frac{p(p+1)}{2}$$

- SP-FSR algorithm [Aksakalli and Malekipirbazari, 2016] shows excellent performance in feature selection. Can we utilise it to identify interaction terms?

Organisation of the presentation

1. Basic terminology
2. Earlier and current methods to identify pairwise interaction terms
3. SP-FSR algorithm for interaction identification
4. Experimental setup and results
5. Discussion
6. Conclusion

Notations and terminology

- Y: response feature
- X_j: explanatory feature j for $j = 1, 2, ...p$
- Model formulation:

$$g(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{i<j} \beta_{i:j} X_j X_i$$

- A precise definition [Lim and Hastie, 2018, p.1]:

"When a function $f(x_1, x_2)$ cannot be expressed as $h_1(x_1) + h_2(x_2)$ for some functions h_1 and h_2, we say that there is an interaction in f between x_1 and x_2."

- Introduction of "hierarchy"

Terminology: hierarchy

Lim and Hastie [2018] define:

Hierarchy	Description
Strong	Interactions are only among pairs of nonzero main effects
Weak	Each interaction has only one of its main effects present
Anti-hierarchical	Interactions are only among pairs of main effects that are not present
Pure interaction	No main effects present; only interactions

Terminology: an example of hierarchy

Consider three explanatory features: $\mathbf{X} = \{X_1, X_2, X_3\}$

1. Strong hierarchy: $\{X_1, X_2, X_1X_2\}$
2. Weak hierarchy: $\{X_1, X_1X_2, X_1X_3\}$
3. Anti-hierarchical: $\{X_2, X_1X_3\}$
4. Pure interaction: $\{X_2X_3, X_1X_3, X_1X_2\}$

In practice, how do we know? How can we detect?

Main methods to identify pairwise interaction terms

1. Statistical hypothesis test [Cox, 1984]
2. Regularisation, e.g. LASSO
3. Wrapper Feature Selection [Kohavi and John, 1997]

Statistical hypothesis test: an example

Considers two models:

1. $g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. $g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2$

Run hypothesis on test on $\beta_{1:2} = 0$.

Regularisation

- Let $l(y_i; \beta)$ be the **negative** log-likelihood contribution
- Elastic-net [Friedman et al., 2010]

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=i}^{n} l(y_i; \beta) + \lambda[(1-\alpha) \parallel \beta \parallel_2^2 /2 + \alpha \parallel \beta \parallel_1]$$

- `glinternet` [Lim and Hastie, 2018]: group-based LASSO ($\alpha = 1$) by imposing additional constraints on $\beta_{k:j}$ and β_j

Wrapper feature selection

Extending feature selection by including interaction terms

1. SFFS: Sequential (Floating) Forward Selection [Pudil et al., 1994]
2. ~~SFBS: Sequential (Floating) Backward Selection~~[1994]
3. GA: Genetic Algorithm [Siedlecki and Sklansky, 2011]
4. SP-FSR [2016]

SP-FSR algorithm

- Introduced by Aksakalli and Malekipirbazari [2016]
- Based on Simultaneous Perturbation Stochastic Approximation [Spall, 1992]
- Refined by Yenice et al. [2018]
- Pseudo gradient descent method on the loss function
- `spFSR` package is now available in R [Aksakalli et al., 2018]

SP-FSR algorithm to identify interactions

- Assume a strong hierarchy
- Simplified version of **two-step SP-FSR** algorithm:

1. Identify the optimal set of k main effects using SP-FSR
2. Search k', number of interactions from k main effects with SP-FSR

- k and k' can be determined via grid search or automatically

Experimental setup

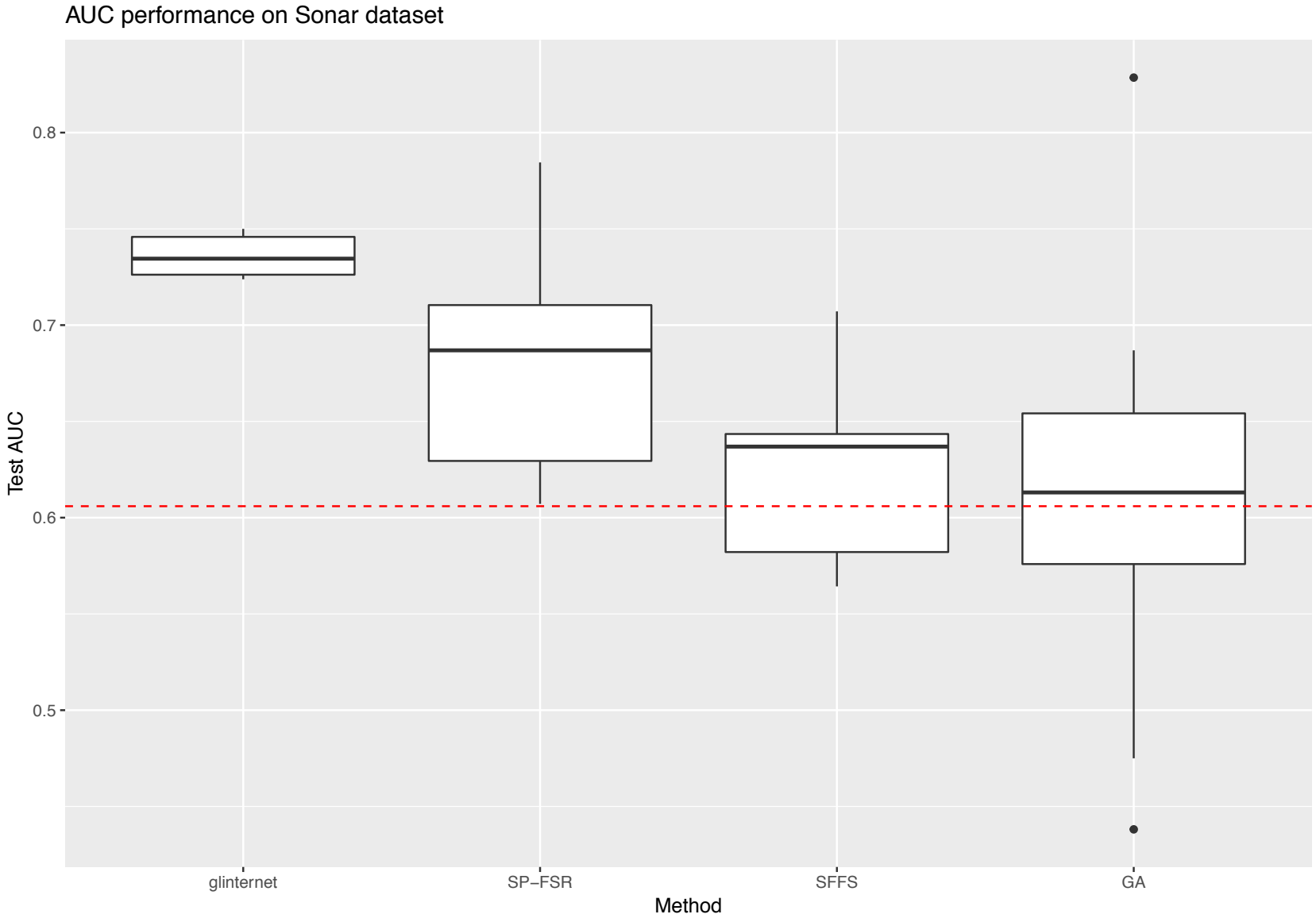
- Datasets with binary Y: Ionosphere and Sonar
- Source: Lichman [2013]
- 80-20 training/test split
- Model: logistic regression
- Evaluation: AUC, AIC, and BIC
- Assume strong hierarchy
- Methods: GA, SFFS, SP-FSR, `glinternet`
- Run on 6 to 12 random seeds

Experimental results: Sonar dataset

Summary statistics

- $n = 208, p = 60$
- Baseline test AUC: 0.6059524
- Baseline train AIC and BIC: 122.0000 and 312.1976
- SFFS tends to run into errors
- Test AUC, Train AUC, AIC, and BIC are mean values

Method	Count	Test AUC	Train AUC	k	k'	AIC	BIC
GA	12	0.6117063	1.0000000	22.00000	110.416667	243.3333	622.6892
SFFS	6	0.6257937	0.9647436	11.66667	7.166667	118.3868	180.2271
SP-FSR	12	0.6810516	1.0000000	16.41667	65.500000	133.0000	340.3466
glinternet	12	0.7367063	1.0000000	45.75000	82.833333	277.5137	678.4358

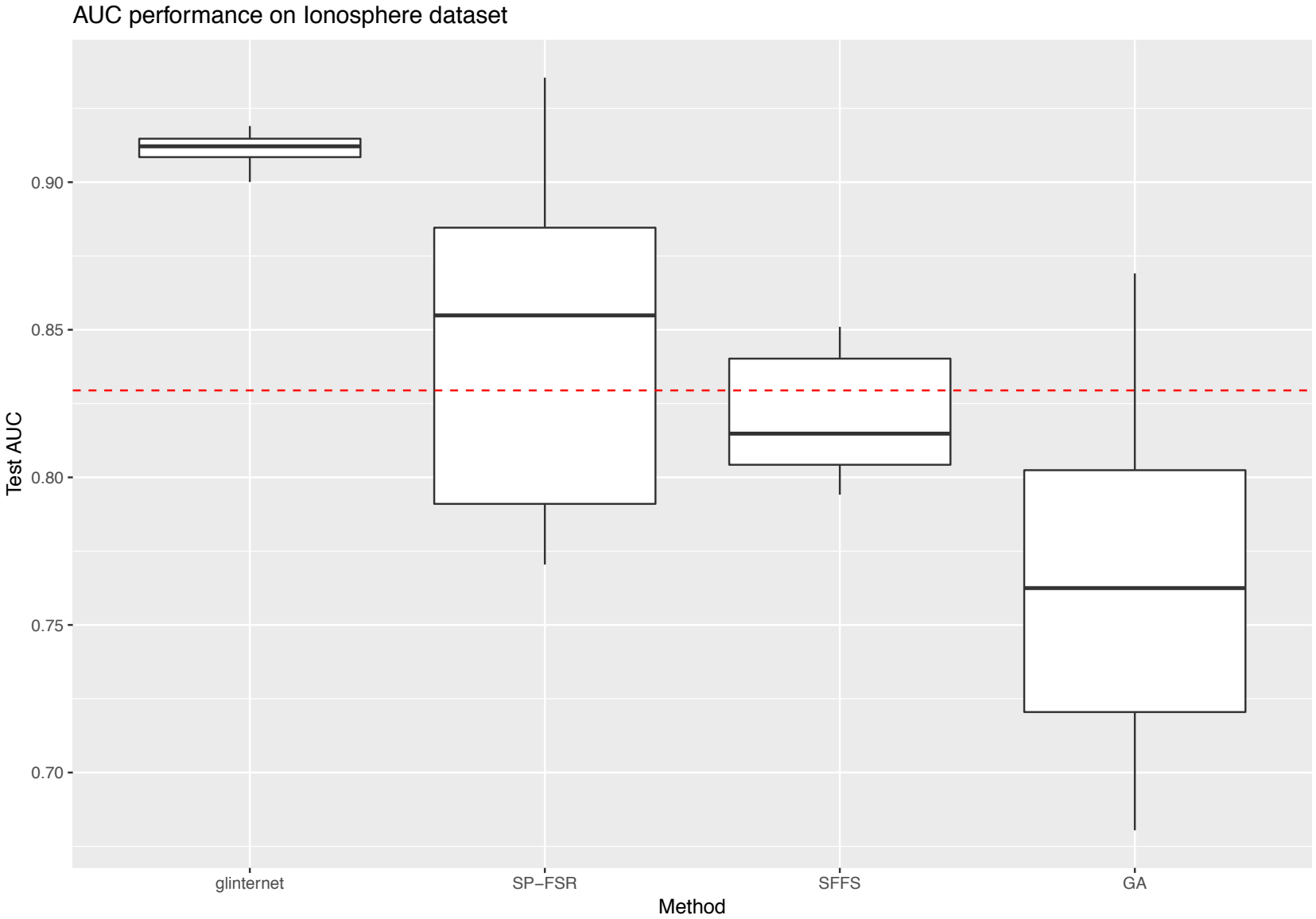

AUC performance on Sonar dataset
Test AUC
0.8
0.7
0.6
0.5
glinternet
SP−FSR
SFFS
GA
Method

Experimental results: Ionosphere dataset

Summary statistics

- $n = 351, p = 33$
- Baseline test AUC: 0.8294574
- Baseline train AIC and BIC: 101.0075 and 224.7115
- SFFS tends to run into errors
- Test AUC, Train AUC, AIC, and BIC are mean values

Method	Count	Test AUC	Train AUC	k	k'	AIC	BIC
GA	12	0.7678366	0.9813844	15.83333	59.416667	517.8325	795.2571
SFFS	6	0.8207005	0.9580975	8.50000	5.666667	155.8369	211.0186
SP-FSR	12	0.8432386	0.9797910	12.08333	28.416667	138.3111	245.3394
glinternet	12	0.9109963	0.9913929	27.16667	37.833333	206.9426	443.4357


AUC performance on Ionosphere dataset
Test AUC
0.90
0.85
0.80
0.75
0.70
glinternet
SP−FSR
SFFS
GA
Method

Conclusion and Future works

Conclusion

- In identifying interactions, SP-FSR trails behinds `glinterent` in terms of accuracy measures, but it selects a smaller model on average based on information criteria.

Future works

- Extend experiment to continuous and multinomial Y
- Include assessment of underlying assumptions?
- Include quadratic and higher-order terms
- Include overidentifcation mechanism

Acknowledgement

- Dr Vural Aksakalli (VA) as my supervisor and mentor
- VA, Dr Babak Abbasi, and Zeren D. Yenice for `spFSR` package
- Zeren D. Yenice, Niranjan Adhikari, Vural Aksakalli, Dr Alev Taskin Gumus, and Babak Abbasi for their previous works on SP-FSR algorithm

References I

Vural Aksakalli and Milad Malekipirbazari. Feature selection via binary simultaneous perturbation stochastic approximation. *Pattern Recognition Letters*, 75(Supplement C):41 – 47, 2016. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2016.03.002. URL http://www.sciencedirect.com/science/article/pii/S0167865516000702.

Vural Aksakalli, Babak Abbasi, and Yong Kai Wong. *spFSR: Feature Selection and Ranking by Simultaneous Perturbation Stochastic Approximation*, 2018. https://www.featureranking.com/, https://arxiv.org/abs/1804.05589.

D. R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1–24, 1984. URL http://www.jstor.org/stable/1403235.

References II

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v033/i01.

R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 1(2):273–324, 1997.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Michael Lim and Trevor Hastie. *glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization*, 2018. URL https://CRAN.R-project.org/package=glinternet. R package version 1.0.7.

References III

P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, volume 2, pages 279–283 vol.2, Oct 1994. doi: 10.1109/ICPR.1994.576920.

W. Siedlecki and J. Sklansky. *A Note on Genetic Algorithm for Large-Scale Feature Selection*, pages 88–107. 2011. doi: 10.1142/9789814343138_0005. URL https://www.worldscientific.com/doi/abs/10.1142/9789814343138_0005.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE*, 37(3):322–341, 3 1992.

Zeren D. Yenice, Niranjan Adhikari, Yong Kai Wong, Vural Aksakalli, Alev Taskin Gumus, and Babak Abbasi. Spsa-fsr: Simultaneous perturbation stochastic approximation in feature selection and ranking. 2018. URL https://arxiv.org/abs/1804.05589.