A Comparison of Hamming Errors of Representative Variable Selection Methods

Authors: Tracy Ke, Longlin Wang

Presenter: Longlin Wang (`lwang2@fas.harvard.edu`)

March 14, 2022

Motivation of Our Work

Variable selection is a well-studied area, but limitations remain.

- Choice of the criterion: model selection consistency (i.e. $\mathrm{P}[\mathrm{Supp}(\hat{\beta}) = \mathrm{Supp}(\beta)]$) might be idealistic.
- Lack of comparability: Existing methods are analysed under various and distinct models.

How we address these:

- Study the Hamming error;

$$\mathrm{H}(\hat{\beta}, \beta) = \sum_{j=1}^{p} \mathbb{I}_{\left\{\hat{\beta}_j \neq 0, \beta_j = 0\right\}} + \mathbb{I}_{\left\{\hat{\beta}_j = 0, \beta_j \neq 0\right\}} \tag{1}$$

- Use a unified framework to comprehensively study different methods.

High-level Summary of Our Model

Our model features a triplet of parameters (ϑ, r, ρ)

- ϑ: signal sparsity
- r: signal strength
- ρ: correlation (sign and level)

Our goal: The explicit Hamming error rates of different methods in a unified framework.

$$\mathrm{H}(\hat{\beta}, \beta) = \sum_{j=1}^{p} \mathbb{I}_{\{\hat{\beta}_j \neq 0, \beta_j = 0\}} + \mathbb{I}_{\{\hat{\beta}_j = 0, \beta_j \neq 0\}}$$

$$\mathsf{E}[\mathrm{H}(\hat{\beta}, \beta)] = L_p \cdot p^{1-h(\vartheta, r, \rho)} \text{ where } L_p \text{ is a multi-log term} \quad (2)$$

Our goal is to exactly study $h(\vartheta, r, \rho)$.

Rare/Weak Signal Model with Correlated Design

- Linear model: $y = X\beta + z$, $\|x_j\| = 1$, $z \sim \mathcal{N}(0, I_n)$
- The "rare/weak" signals: (Donoho & Jin, 2004; Arias-Castro et al., 2011; Jin & Ke, 2016)

$$\beta_j = \begin{cases} \tau_p, & \text{with prob. } \epsilon_p, \\ 0, & \text{with prob. } 1 - \epsilon_p, \end{cases} \quad \epsilon_p = p^{-\vartheta},\ \tau_p = \sqrt{2r \log(p)}. \tag{3}$$

- Blockwise Correlated Designs:

$$X'X = \mathrm{diag}(B, B, \cdots, B),\ B = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{4}$$

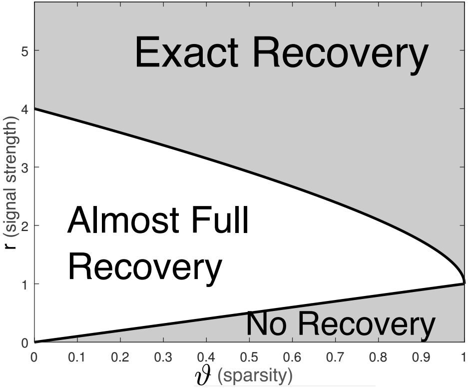
connection to real data in biological and financial scenarios. (Dehman et al., 2015; Fan et al., 2015)

Visualization with Phase Diagrams

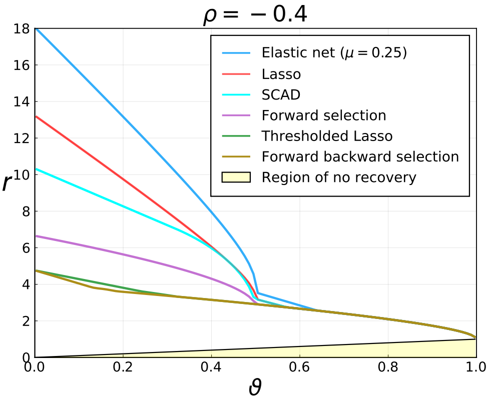
Phase transition:

- *Exact Recovery*: $\mathrm{E}[H(\hat{\beta},\beta)] = o(1)$.
- *Almost Full Recovery*: $\mathrm{E}[H(\hat{\beta},\beta)] = \Omega(1)$ but $\mathrm{E}[H(\hat{\beta},\beta)] = o(p \cdot p^{-\vartheta})$.
- *No Recovery*: $\mathrm{E}[H(\hat{\beta},\beta)] = \Omega(p \cdot p^{-\vartheta})$

An example of phase diagrams (plotted at $\rho = 0$):



A Peek into Our Main Results



Phase curves separating the regions $\implies$ the lower the better;
Compare and contrast $\implies$ pros and cons.
– See more in our paper!

References

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962-994, 2004.

Ery Arias-Castro, Emmanuel J Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, pp. 2533-2556, 2011.

Jiashun Jin and Zheng Tracy Ke. Rare and weak effects in large-scale inference: methods and phase diagrams. *Statistica Sinica*, pp. 1-34, 2016.

Alia Dehman, Christophe Ambroise, and Pierre Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(1):1-14, 2015.

Jianqing Fan, Yuan Liao, and Xiaofeng Shi. Risks of large portfolios. *Journal of Econometrics*, 186 (2):367-387, 2015.