15TH NATIONAL CONVENTION ON STATISTICS
03-05 OCTOBER 2022

Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

On Some Feature Ranking Procedures Applied to High-Dimensional Regression Problem

Aries P. Valeriano
Graduate Student

[Computational Statistics]
Crowne Plaza Manila Galleria
4 October 2022; [3:30-5:00] PM

Introduction

- **Statistical Modelling**
  - $X_1, X_2, \ldots, X_p \longrightarrow Y$ (continuous)
- **High-Dimensional Data (HDD)**
  - $p \gg n$
- **Feature Ranking Procedures**
  - Bayesian Feature Ranking (BFR) by Enes Makalic and Daniel Schmidt (2011)
  - Random Forest (RF) by Breiman (2001)
  - Independent Screening by Generalized Correlation (HM) by Peter Hall and Hugh Miller (2009)
- **Modified Forward Selection Procedure (MFSP)**

Objectives of the Study

i. **Function I and $SNR = 1$**, depicts a scenario where a dataset has y with higher level of noise v infused to it, and $x_j's$ that are independent to each other.

ii. **Function I and $SNR = 8$**, depicts a scenario where a dataset has y with lower level of noise v added to it, and $x_j's$ that are independent to each other.

iii. **Function II and $SNR = 1$**, depicts a scenario where a dataset has y with higher level of noise v added to it, and $x_j's$ that have varying levels of correlation ρ.

iv. **Function II and $SNR = 8$**, depicts a scenario where a dataset has y with lower level of noise v infused to it, and $x_j's$ that have varying levels of correlation ρ.

v. **Function III and $SNR = 1$**, depicts a scenario where a dataset has y with higher level of noise v added to it, and have certain $X_j's$ that are independent to each other and the rest $X_j's$ have varying levels of correlation ρ.

vi. **Function III and $SNR = 8$**, depicts a scenario where a dataset has y with lower level of noise v infused to it, and have certain $X_j's$ that are independent to each other and the rest $X_j's$ have varying levels of correlation ρ.

Datasets

- **Simulated Dataset**
  - 100 datasets per scenario with $n = 50$, $p = 100$, and y is continuous.
- **Real Dataset**
  - "eyedata" with $n = 120$ samples or rats, $p = 200$ gene probes, and Y, a vector with 120 expression levels of the TRIM32 gene.

Feature Ranking Procedures

- **Random forest (RF)** is mainly used for predictive modelling. However, it also features a variable importance that is quite popular. This feature is capable of including influential variables that are highly correlated to other significant variable, in which it is unlikely to happen with other prediction-based variable selection method.
- **Bayesian feature ranking (BFR)** is a variable selection method for any parametric model where samples from the posterior distribution of the parameters are available. This algorithm not only computes an importance ranking of all observed features but also provides a credible intervals that can be used in removing those features that has little influence to the target variable. Also, it does not require any user specified parameters and is applicable to both regression and classification problems.

- **Independent screening by generalized correlation (HM)** is a variable selection method in which it is based on ranking generalized empirical correlations between Y and $X_j; j = 1, 2, \ldots, p$. This method is not prediction based and can identify variables that are influential but not explicitly part of the predictive

Modified Forward Selection Procedure (MFSP)

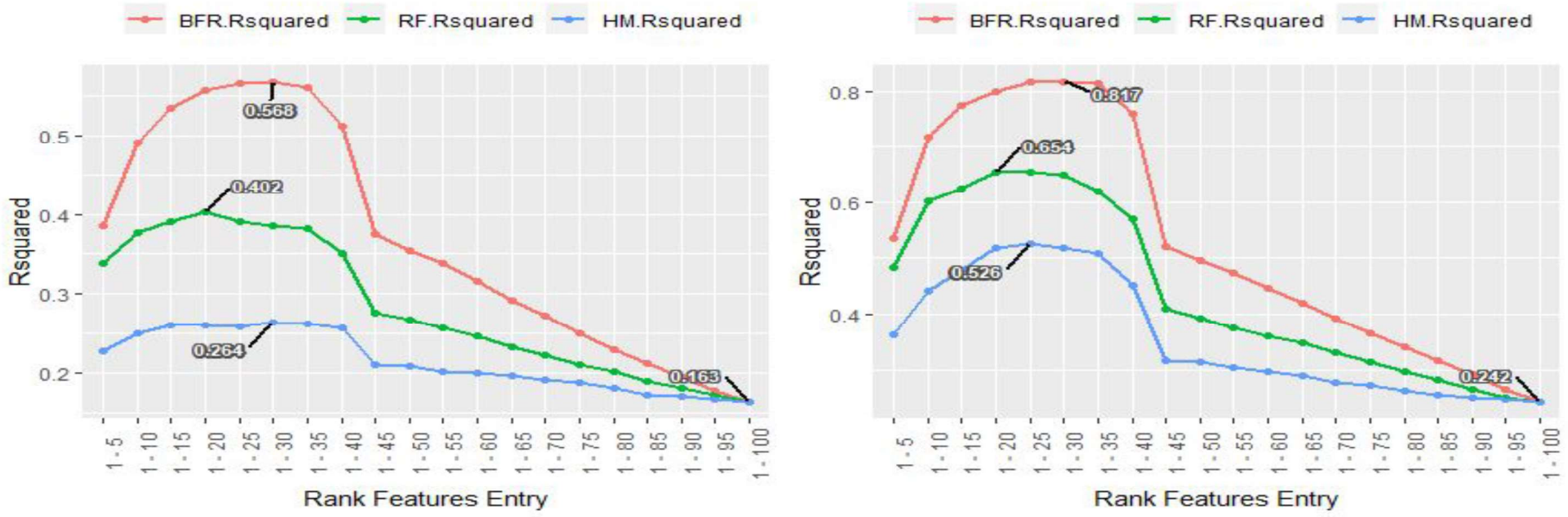
- $\Delta = 5$
- Fit ridge multiple linear regression model.
  - R^2
- Does not stop until the model includes all features.

Function I and $SNR \in \{1, 8\}$

(a) $SNR = 1$ (b) $SNR = 8$

Figure 3: *Performance of BFR, RF, and HM for Function I as rank features entry increases.*

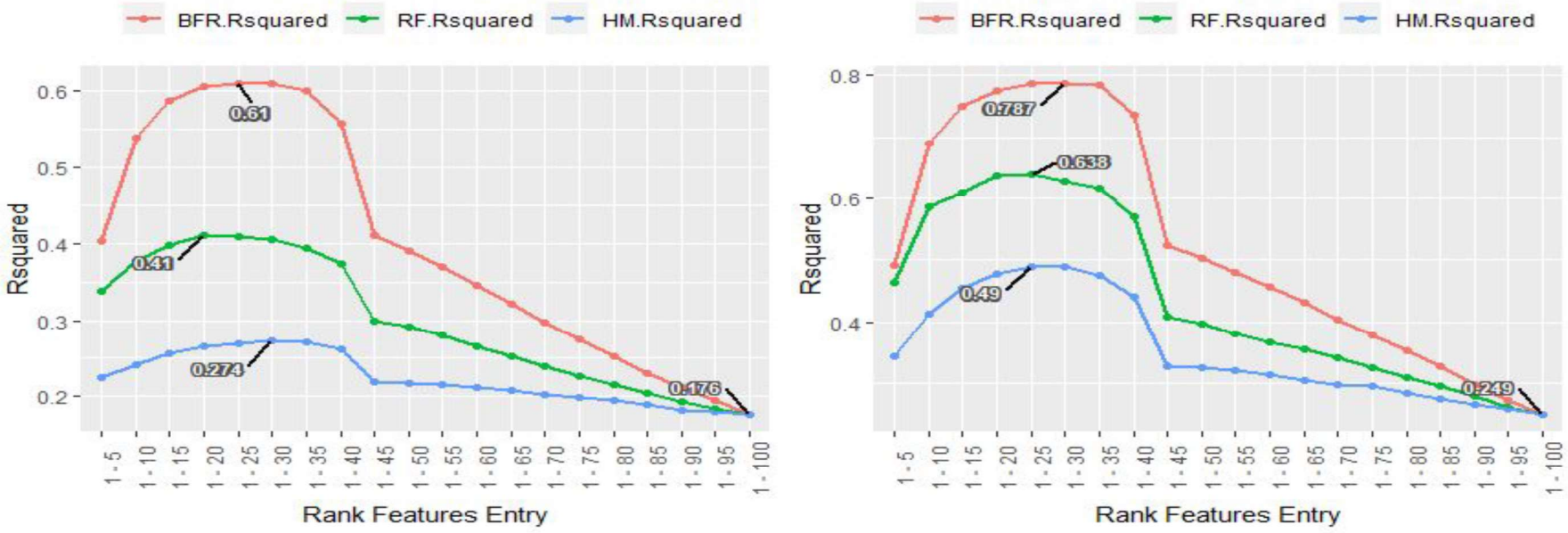
Function II and $SNR \in \{1, 8\}$



(a) $SNR = 1$ (b) $SNR = 8$

Figure 4: *Performance of BFR, RF, and HM for Function II as rank features entry increases.*

Function III and $SNR \in \{1, 8\}$



(a) $SNR = 1$ (b) $SNR = 8$

Figure 5: *Performance of BFR, RF, and HM for Function III as rank features entry increases.*

Real Dataset (eyedata)

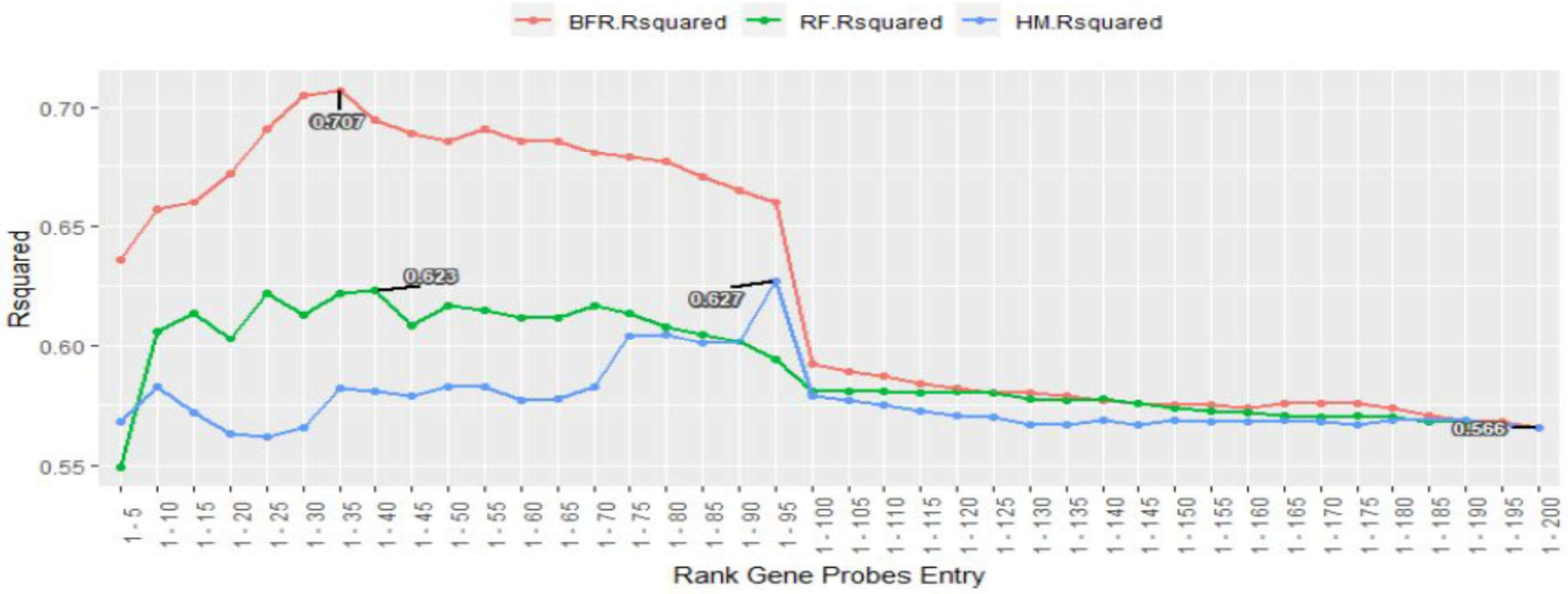


Figure 6: *Performance of BFR, HM, and RF as Rank Gene Probes Entry Increases.*

Table 11: *Top 35 Gene Probes.*

Rank	BFR	RF	HM	Rank	BFR	RF	HM
1	X87	X96	X153	19	X19	X5	X152
2	X62	X153	X23	20	X92	X178	X5
3	X180	X37	X55	21	X174	X187	X122
4	X153	X154	X52	22	X90	X11	X88
5	X140	X17	X67	23	X41	X165	X155
6	X76	X180	X96	24	X146	X87	X47
7	X134	X151	X120	25	X66	X131	X134
8	X200	X90	X199	26	X11	X135	X60
9	X187	X38	X24	27	X48	X158	X64
10	X155	X141	X42	28	X99	X52	X21
11	X71	X117	X40	29	X164	X139	X14
12	X102	X157	X43	30	X172	X110	X143
13	X50	X173	X140	31	X42	X177	X180
14	X54	X99	X71	32	X101	X169	X85
15	X184	X67	X168	33	X147	X172	X148
16	X13	X140	X10	34	X157	X186	X15
17	X96	X156	X171	35	X114	X196	X7
18	X185	X55	X45				

Thank you!

http://www.psa.gov.ph/ncs

http://openstat.psa.gov.ph

https://twitter.com/PSAgovph

https://www.facebook.com/PSAgovph