Covariate Information Number for Feature Screening in Ultrahigh-Dimensional Supervised Problems

Debmalya Nandy
dnandy@psu.edu
[Joint work with Francesca Chiaromonte & Runze Li]
The Pennsylvania State University

ASA Symposium on Data Science & Statistics
Bellevue, WA | June 01, 2019

1

Response variable
n random sample units
p-dimensional feature vector
p >> n

2

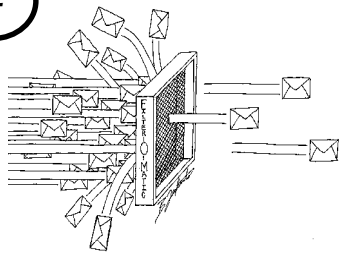
Computational burden
Statistical inaccuracy
Algorithmic instability

[Fan, J. *et al.* (2009), **JMLR**, 10(Sep): 2013-38]

3

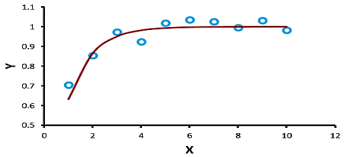
~~Finding needles in a haystack?~~
Let's find the right haystack first!

4



Feature Screening

5



1. Compute marginal utility

2. Rank the features

3. Screen the top d

Google Images

Covariate Information Screening

- **X** : p-dimensional feature vector (**p >> n**), Y: response, n: sample size
- $f_j(\cdot)$: marginal density of feature X_j
- A = { j : F(y | **x**) depends on x_j } : "**active**" set
- Fisher information for "parameter" $X_j = x_j$ in $f(y \mid x_j)$: $I_j = \int \left[\frac{d}{dx_j} \log f(y|x_j) \right]^2 f(y|x_j)\, dy$
- Covariate Information Number for feature X_j in $f(y, x_j)$: $\boldsymbol{\omega_j = \int I_j\, f_j(x_j)\, dx_j\ (\geq 0)}$
- Define [4]: $J_{X_j} = \int \left[\frac{d}{dx_j} \log f(x_j) \right]^2 f(x_j)\, dx_j$

  $J_{X_j \mid Y=y} = \int \left[\frac{d}{dx_j} \log f(x_j|y) \right]^2 f(x_j|y)\, dx_j$ and $J_{X_j \mid Y} = \int J_{X_j \mid Y=y}\, f(y)\, dy$
- We can write $\hat{\omega}_j = \hat{J}_{X_j \mid Y} - \hat{J}_{X_j}$, j = 1, 2, …, p. Feature screening with $\hat{\omega}_j$ is **CIS**

Implementation of CIS

Let $(y_i, \mathbf{x}_i^T)$, i = 1, 2, …, n, be a random sample from the distribution of (Y, **X**).
For each j = 1, 2, …, p :

- Compute $\hat{J}_{X_j}$, $\hat{J}_{X_j \mid Y}$, and $\hat{\omega}_j = \hat{J}_{X_j \mid Y} - \hat{J}_{X_j}$ using "**slicing**" and **kernel density estimation** (Gaussian kernel, rule of thumb bandwidth; different number of slices)

- **Normalize** $\hat{\omega}_j$'s relative to $\hat{J}_{X_j}$: $\hat{\omega}_j^* = \hat{\omega}_j \,/\, \hat{J}_{X_j}$

- **Rank** $\hat{\omega}_j^*$'s in **decreasing** order

- **Screen** X_j's corresponding to the **top d** $\hat{\omega}_j^*$'s. Use "hard" and/or "soft" **thresholding** [5] to determine d

Properties of CIS

- **Model-free** feature screening approach [5, 6, 7]
- CIS has the **sure screening** property [2]
- ω_j has a **Fisher informational** interpretation
- Implementation is **fast**, uses **both** $f(x_j|y)$ and $f_j(x_j)$, is **robust** to **response outliers**, and allows **any type** of **response** with continuous features

Simulation Study

$$Y = c\left(X_1 + 0.75\,X_2^2 + 2.25\cos(X_5)\right) + \varepsilon$$

- $\varepsilon \sim N(0,1)$, **n = 200** and **600**, **p = 2000**
- $c = 0.5, 1.25,$ and 2.5 (SNR $\approx 0.8, 5,$ and 20)
- $A = \{1, 2, 5\}$
- r : minimum number of ranked features to include active X_j's, $j \in A$
- Independent $\mathbf{X}$: $X_j \sim N(0,1)$, $j = 1, \ldots, p$; $X_i \perp\!\!\!\perp X_j, i \neq j$
- Dependent $\mathbf{X}$: MVN($\mathbf{0}$, $((\sigma_{ij}))$); $\sigma_{ii} = 1$; $\sigma_{ij} = 0.2, i \neq j$ (compound-symmetric)
- SNR: Signal-to-Noise Ratio; Var(E[Y | $\mathbf{X}$])/E[Var(Y | $\mathbf{X}$)]
- **SIS**: **S**ure **I**ndependence **S**creening [2];
  SIRS: **S**ure **I**ndependent **R**anking and **S**creening [5];
  DC-SIS: **D**istance **C**orrelation-based **SIS** [6]; **MDC-SIS**: **M**artingale **DC**-based **SIS** [7]

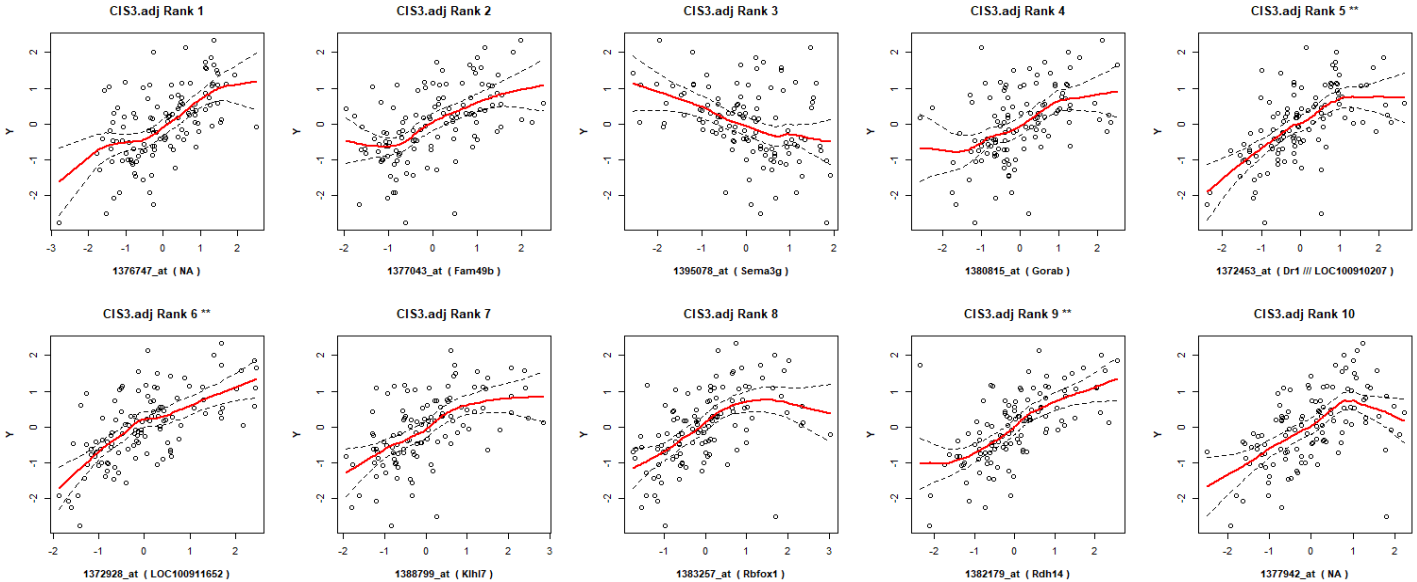
Simulation Study [Contd.]

SNR	0.8	5	20	0.8	5	20
X	Independent			Dependent		
	n = 200					
SIS	1279 (406)	1230 (460)	1231 (465)	1398 (379)	1394 (382)	1449 (379)
SIRS	1425 (365)	1423 (375)	1414 (398)	1451 (350)	1485 (337)	1562 (294)
DC-SIS	28 (19)	**3** (0)	**3** (0)	80 (61)	10 (7)	6 (3)
MDC-SIS	**22** (14)	**3** (0)	**3** (0)	**65** (51)	11 (7)	7 (4)
CIS	185 (158)	4 (1)	**3** (0)	221 (182)	**5** (2)	**3** (0)
	n = 600					
SIS	1333 (411)	1183 (473)	1246 (439)	1477 (352)	1493 (372)	1561 (346)
SIRS	1402 (362)	1390 (402)	1420 (379)	1505 (332)	1636 (266)	1721 (227)
DC-SIS	**3** (0)	**3** (0)	**3** (0)	4 (1)	**3** (0)	**3** (0)
MDC-SIS	**3** (0)	**3** (0)	**3** (0)	4 (1)	**3** (0)	**3** (0)
CIS	**3** (0)	**3** (0)	**3** (0)	**3** (0)	**3** (0)	**3** (0)

Table 1: Median (median absolute deviation) of r in 1000 replications. **Better = Closer to 3**. Smallest value for each scenario is **bold**-faced. Number of CIS slices = 5

Transcriptomic Data Application [n = 120, p = 18975]

- **Affymetrix GeneChip Rat Genome 230 2.0 Array** data on *Rattus norvegicus* [1, 7];
- Screened **gene expressions** (features **X**) associated with expression level of **Trim32 gene** (response Y); **CIS (3 slices)**; fit **G**eneralized **A**dditive **M**odel (**GAM**) with **top d = 10** genes



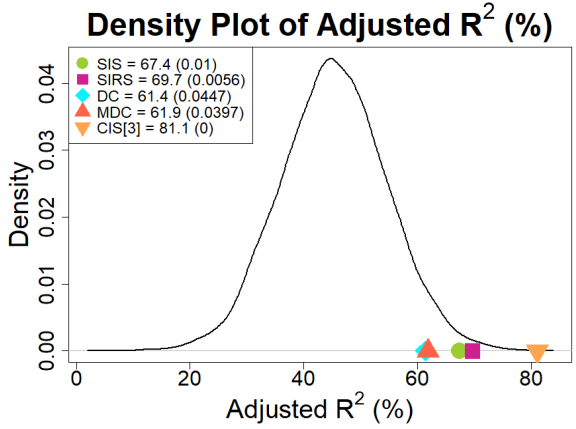
Genes with ranks 8 and 10 vary **non-linearly** with Y. Ranks 6 and 9 are **common** to **all methods. Klhl7** (rank 7) has effects in **retinitis pigmentosa** in rats. Mutations/alleles of **Fam49b** (rank 2) cause **abnormal retinal morphology. Rbfox1** (rank 8) is expressed in **retina ganglion cell layer**; **Dr1** (rank 5) in inner and outer retinal layers. See for refs: **https://goo.gl/bQ4rDy**

Figure 1: Expression levels scatterplots for the **response Trim32** against each of the **top d = 10** genes ranked by **CIS**. Red solid lines are LOESS smoothers; dashed lines are prediction bands (2 SD)

Transcriptomic Data Application [Contd.]

	SIS	SIRS	DC-SIS	MDC-SIS	**CIS**
Adjusted R^2 (%)	67.4	69.7	61.4	61.9	**81.1**
Deviance Exp. (%)	73.7	75.7	65.6	68.2	**89.7**

Table 2: Adjusted R^2 (%) and deviance explained (%) by GAM fits with the **top d = 10** genes



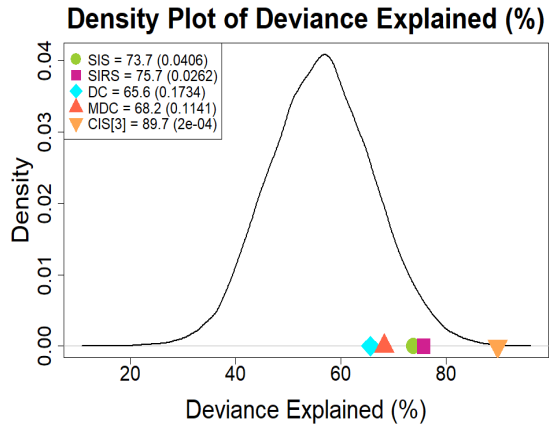


Figure 2: Distributions of **adj. R^2** and **dev. Explained** from about 10000 GAM fits using **top d = 10 randomly selected** genes. Colored dots are values for **CIS**, **SIS**, **SIRS**, **DC-SIS**, and **MDC-SIS**. Empirical p-values in legends.

Ongoing/Future research avenues

- Construct a graphical diagnostic to determine d
- Prove ranking consistency for CIS theoretically
- Use $\omega_\lambda = \lambda\,\omega_{CIS} + (1-\lambda)\,\omega_S$, $\lambda \in [0,1]$; S: another screening procedure
- Implement iterative CIS using the notion of predictor residual matrix [5]
- Extend CIS for cases with (i) multivariate response and
  (ii) response and features all discrete/categorical

References

1. Scheetz, T. E., et al. (2006), **PNAS**, 103(39):14429-14434
2. Fan, J. and Lv, J. (2008), **JRSS-B**, 70(5): 849-911
3. Fan, J., Samworth, R., and Wu, Y. (2009), **JMLR**, 10(Sep):2013-2038
4. Hui, G. and Lindsay, B. G. (2010), **Sankhya B**, 72(2):123-153
5. Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011), **JASA**, 106(496):1464-1475
6. Li, R., Zhong, W., and Zhu, L. (2012), **JASA**, 107(499):1129-1139
7. Shao, X. and Zhang, J. (2014), **JASA**, 109(507):1302-1318

Acknowledgments

NSF Grant DMS-1407639

Bharath Sriperumbudur. Penn State Statistics

Amal Agarwal and Mauricio Nascimento. Penn State Statistics

Xiaofeng Shao. UIUC Statistics

Jingsi Zhang. NorthwesternU Statistics

Binglan Li [Victoria]. Genomics & Computational Biology, UPenn Perelman School of Medicine

Makova Lab & Collaborators. Penn State

Summary

Covariate Information Number – Sure Independence Screening [CIS]

1. Model-free marginal utility with Fisher informational notion
2. Utility with both inverse regressional and marginal density information
3. Sure screening property
4. Any response type with continuous features and discrete/categorical features with continuous response
5. Desired performance in simulations, competitive computational time
6. Interpretable results in real data application

~~Finding needles in a haystack?~~ Let's find the right haystack first! ☺