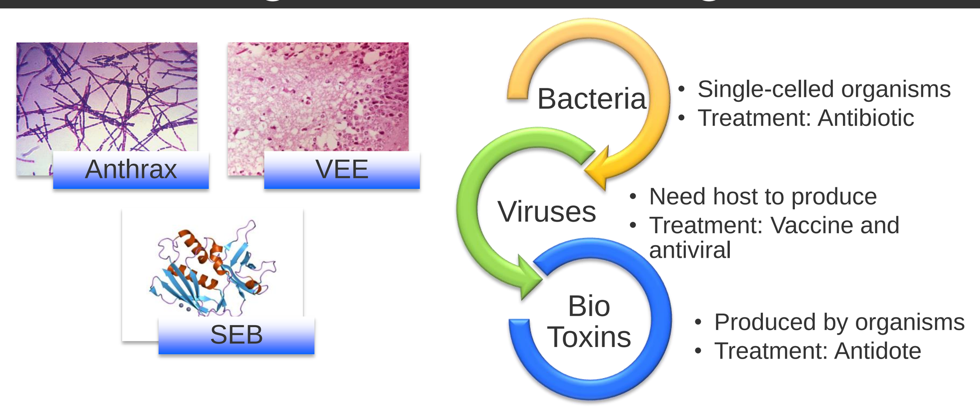
University of Georgia, Big Data Analytics Lab

Weighted Leverage Score for Genetic Marker Selection

Yiwen Liu[1], Peng Zeng[2] , Wenxuan Zhong[1]

1 Department of Statistics, University of Georgia. 101 Cedar Street, Athens, GA. 2 Department of Statistics, Auburn University.

Background: biothreat agents



❖ **Characteristic of biothreat agents**

- Transfer fast from person to person
- Most agents have no vaccine
- Difficult to detect in their early stage

❖ **Detection of biothreat agents**

A set of **phenotypical measurements** on a host are highly unreliable in the early biothreat detection. Certain genes in infected cells show different expression levels for different pathogens. (Das et al. (2008)). Thus **genomic markers** one of the most reliable indicators and are widely used in the past decades (Lim et al. (2005)). Our goal becomes to **identify the differentially expressed genes for different pathogens**.

❖ **Challenges in biothreat detection.**

Pathogens

Sample	Types
1	Anthrax
.	.
.	.
n	Plague

Gene expressions

	Gene 1		Gene p
1	1.3	…	2.7
.	.	.	.
.	.	.	.
n	4.1	…	6.4

Dimension Reduction Framework

Let $Y \in \mathbb{R}$ be the response variable and $X = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ be the predictors with $E(X) = 0$ and $cov(X) = \Sigma_X$. $(x_i, y_i)|_{i=1}^n$ is an observation from the ith subject $i = 1, \dots, n$. Throughout the poster, We assume the following model (Li 1991):

$$Y = f(\beta_1^T X, \dots, \beta_K^T X, \epsilon) \qquad (1)$$

where $f(\cdot)$ is an unspecified link function on $\mathbb{R}^{K+1}$, $\beta_1, \dots, \beta_K$ are p-dimensional vectors and ϵ is the random error independent of X with mean 0 and finite variance. If $\beta_{kj} = 0$ for all $k = 1, \dots, K$, X_j is referred to as a irrelevant predictor, otherwise, it is a relevant predictor. We further developed the notation T as the set of relevant predictors and T^c as the set of irrelevant predictors. When model (1) holds, p-dimensional variable X is projected onto a K-dimensional subspace $\mathcal{S}$ spanned by $\beta_1, \dots, \beta_K$, which captures all the information in Y,

$$Y \perp X | P_{\mathcal{S}} X \qquad (2)$$

Where $P_{\mathcal{S}}$ is the projection matrix.

When $f(\cdot)$ is unknown, consider the profile correlation function,

$$R^2(\beta_i) = \max_{\beta, T} Corr^2(T(Y), \beta^T X)$$
$$s.t.\ cov(\beta_i^T X, \beta_j^T X) = 0, \qquad i \neq j$$

Intuitively, β_1 is a direction in $\mathbb{R}^p$ along which the transformed Y and $\beta_1^T X$ have the largest correlation coefficient. β_2, orthogonal to β_1, is a direction that produce the second largest correlation coefficient between $T(Y)$ and $\beta_2^T X$. Under the assumption of model (1) or (2), the procedure can be continued until all K directions are found that are orthogonal to each other and have nonzero $R^2(\beta)$ resulting in $\beta_1, \dots, \beta_K$ that spanned the K-dimensional subspace $\mathcal{S}$.

Weighted Leverage Score

❖ **Derivation of weighted leverage score**

The solution of the profile correlation problem is,

$$\beta^* = \arg\max_\beta \frac{\beta^T \overbrace{Var[E(X|Y)]}^{MSG} \beta}{\beta^T \underbrace{Var(X)}_{MST} \beta}$$

$$Var[E(X|Y)] = \frac{1}{n} \bar{X}_H^T \bar{X}_H, \qquad Var(X) = \frac{1}{n} X^T X = \Sigma$$

where $\bar{X}_H$ is the sliced mean as described in Sliced Inverse regression. Consider the rank d singular value decomposition $X = U\Lambda V^T$, we have $Z = \Sigma^{-1/2} X = UV^T$ as normalized version of X. The solution to the profile correlation problem then is,

$$\beta^* = \arg\max_\beta \beta^T (\bar{Z}_H^T \bar{Z}_H) \beta\ ,$$

$$\bar{z}_{hj} = \frac{1}{n_h} \sum_{i=1}^n \Big(\sum_{k=1}^d u_{ik} v_{jk}\Big) I(y_i \in S_h) = \sum_{k=1}^d \left(\sum_{i=1}^n \frac{1}{n_h} u_{ik} I(y_i \in S_h) \right) v_{jk} = \sum_{k=1}^d \omega_k^h v_{jk}$$

The weighted leverage score of jth variable is defined as follows.

Weighted Leverage Score

$$WLS_j = V_j^T \left(\sum_{h=1}^H p_h \bar{U}_h \bar{U}_h^T \right) V_j$$

where $\bar{U}_h = (\omega_1^h, \dots, \omega_d^h)^T$, $V_j = (v_{j1}, \dots, v_{jd})^T$ and $p_h = n_h/n$.

❖ **Theoretical properties of weighted leverage score.**

The weighted leverage score guarantees the rank consistency given the following conditions.

C1. Linearity condition.

$$E(X|\beta^T X_{\mathcal{T}}) \text{ is linear in } \beta^T X_{\mathcal{T}}.$$

C2. Let $x_1, \dots, x_n \in \mathbb{R}^p, n \geq p$ be independent random vectors that have sub-Gaussian distribution for some L,

$$P(|\langle X, x \rangle| > t) \leq 2e^{-t^2/L^2}$$

for some $t > 0$ *and* $x \in S^{p-1}$.

C3. Covariance matrix.

$$lim_{p\to\infty} \min_{1 \leq i \leq p} \lambda_i > b > 0.$$

where λ_i be the eigenvalues of Σ.

Theorem 1. Given the condition above, we have $\mathrm{WLS}_{j \in \mathcal{T}^c} = 0$ and the following inequality,

$$\max_{j \in \mathcal{T}^c} WLS_j < \min_{j \in \mathcal{T}} WLS_j$$

holds uniformly for $j = 1, \dots, p$.

Theorem 2. With C1 and Theorem 1, under the null hypothesis that given $\beta^T X_{\mathcal{T}}, Y$ is independent of X *for* $j \in \mathcal{T}^c$, $n\hat{w}_j$ follows a weighted χ^2 distribution.

To have the rank consistency of weighted leverage score, we need the following corollary. Vershynin (2012) proved that with C2, for every $\delta > 0$, with probability at least $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^n x_i x_i^T - E(XX^T) \right\| \leq C(L, \delta)(p/n)^{1/2}$$

which guarantees the convergence of $\hat{\Sigma}$. With Weyl's inequality, we have the convergence of $\hat{\Sigma}^{-1/2}$.'

Corollary 1. Let λ_i *and* $\hat{\lambda}_i$ *be the eigenvalues of* Σ *and* $\hat{\Sigma}$. Given condition 3 and when n is large enough, $\lim_{p\to\infty} \min_{1 \leq i \leq p} \hat{\lambda}_i > \frac{b}{2} > 0$ by Weyl's inequality, we have that

$$\left\| \hat{\Sigma}^{-1/2} - \hat{\Sigma}^{-1/2} \right\| = O((p/n)^{-1/2})$$

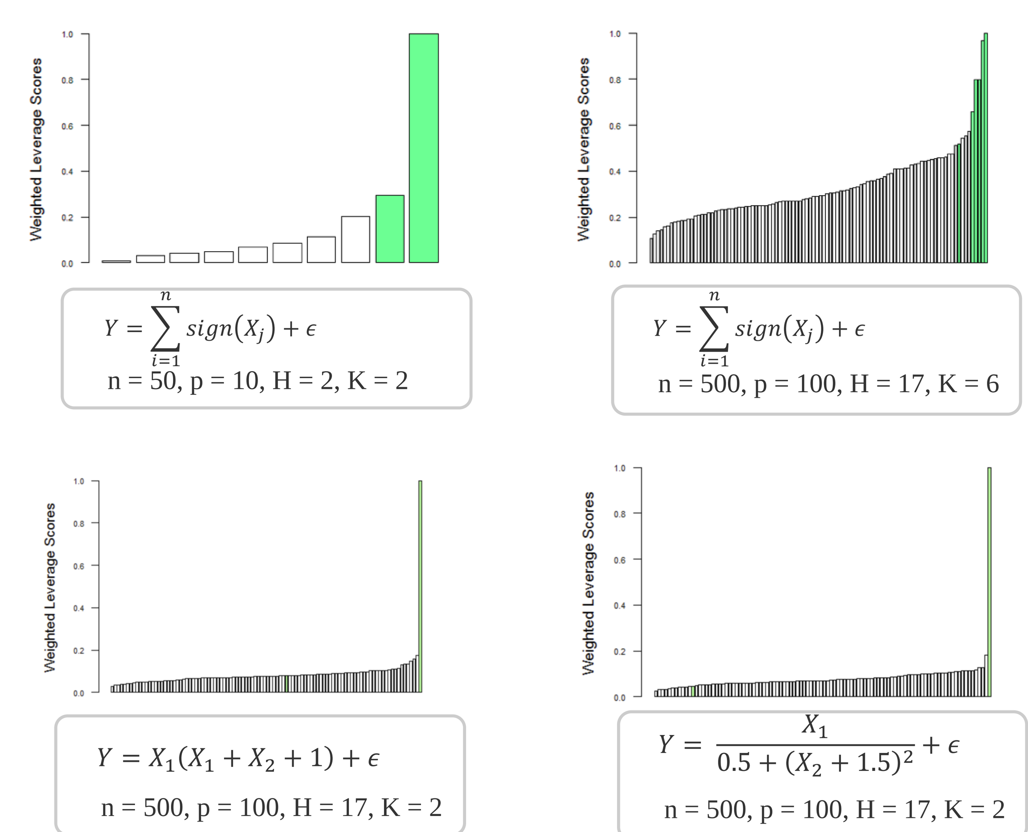
Theorem 3. For any $\epsilon > 0$, there exists a sufficient small constant s such that

$$P\left(\sup_{j=1,\dots,p} \left| \widehat{\boldsymbol{WLS}}_j - \boldsymbol{WLS}_j \right| > \epsilon \right) \leq 2pexp\left(\frac{n}{2} \log(1 - \epsilon s) \right)$$

Thus denote $\delta = \min_{j \in \mathcal{T}} w_j - \max_{j \in \mathcal{T}^c} w_j$, there exists s_δ such that

$$P\left(\min_{j \in \mathcal{T}} \widehat{\boldsymbol{WLS}}_j > \max_{j \in \mathcal{T}^c} \widehat{\boldsymbol{WLS}}_j \right) \leq 1 - 4pexp\left[\frac{n}{2} \log(1 - \delta s_\delta / 2) \right].$$

Simulation Results



Real Data Example

❖ **Background**

To examine the gene expression host responses to different biological threat agents in human peripheral blood mononuclear cells (PBMCs), PBMCs were exposed to various pathogens with different time duration.

❖ **Data structure**

1. Sample: human peripheral blood mononuclear cells
2. 8 biological threat agents (BTAs)

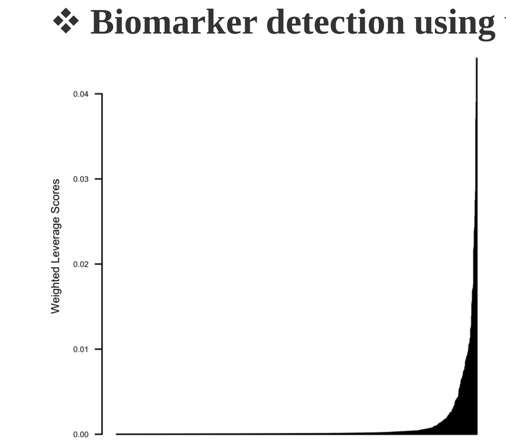
--Toxin: SEB, CT, BoNT-A

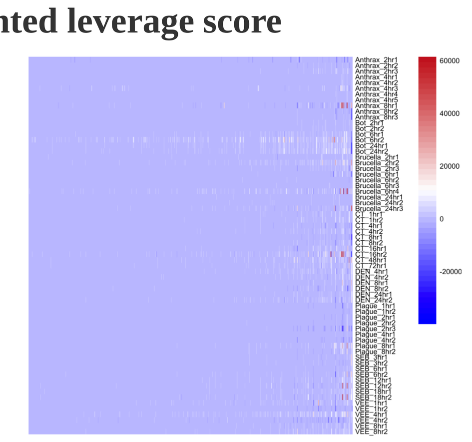
--Bacteria: Anthracis, Yersinia pestis, Brucella melitensis

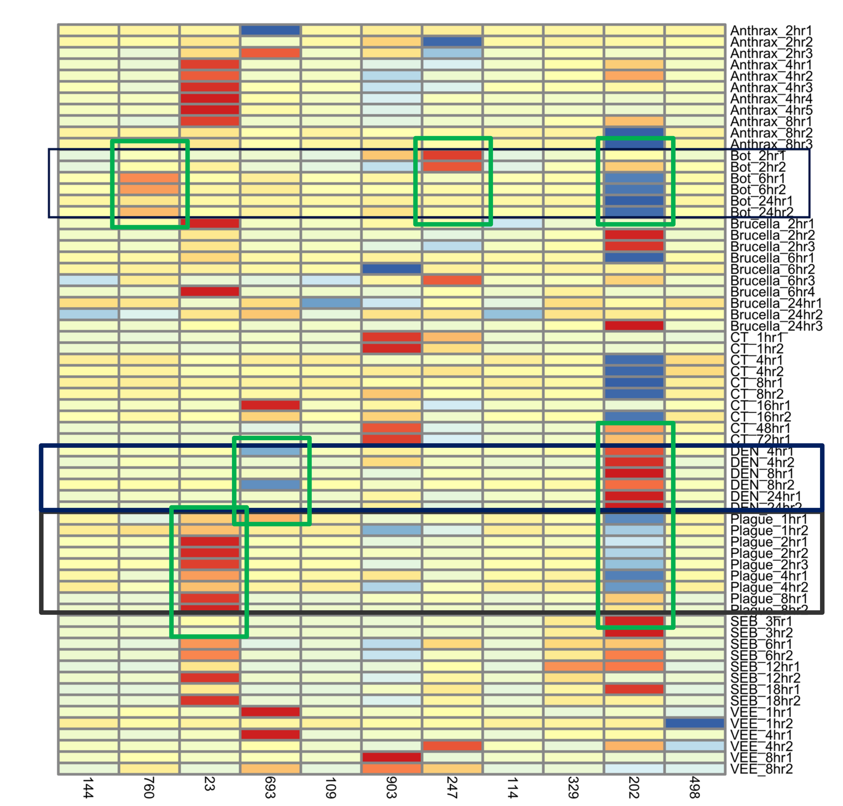
--Viruses: VEE, DEN-2

3. For each pathogen, 3-6 successive time periods were studied. Both infected and uninfected cells were maintained for further analysis.

❖ **Biomarker detection using weighted leverage score**







Ten genes with the highest scores are plotted for illustration.

Discussion

Four distinguished features of the our newly proposed variable screening methods for high dimensional regression analysis are:

- Weighted leverage score screening procedure does not impose assumptions on the relationship between the response variable and the predictors, thus it is considered as model-free variable selection method that is applicable in high dimensional data analysis;
- Weighted leverage score screening procedure enjoys a great computational and theoretical advantage, which is highly desirable for high dimensional data. WLS screening procedure only includes one-time singular value decomposition, which has the computation complexity of $O(np^2)$, while a matrix inversion has a computation complexity of $O(n^3)$.
- There is no need to pre-specify the number of linear combination K.
- Weighted leverage score screening procedure is designed to include both the information from columns of X and the relationship between X and Y.

References

[1] Das R, Hammamieh R, Neill R, Ludwig GV et al. Early indicators of exposure to biological threat agents using host gene profiles in peripheral blood mononuclear cells. *BMC Infect Dis* 2008 Jul 30;8:104

[2] Li, Ker-Chau. "Sliced inverse regression for dimension reduction." *Journal of the American Statistical Association* 86.414 (1991): 316-327.

[3] Chen, Chun-Houh, and Ker-Chau Li. "Can SIR be as popular as multiple linear regression?." *Statistica Sinica* 8.2 (1998): 289-316.

[4] Vershynin, Roman. "How close is the sample covariance matrix to the actual covariance matrix?." *Journal of Theoretical Probability* 25.3 (2012): 655-686.

Acknowledgement

NIH NIGMS

NSF DMS