Regularized ROC Estimation:

with Applications to Classification Using Microarray Data

Shuangge Ma

Department of Biostatistics, University of Washington

(Joint with Jian Huang, University of Iowa)

MICROARRAY IN MEDICAL RESEARCH

- Microarrays are capable of monitoring expressions on a large scale.
- An important application: discover biomarkers associated with different phenotypes.
- A typical study:

  Response: cancer type/survival time (usually $<= 200$);

  Covariate: gene expressions (usually > 1000).

CLASSIFICATION USING MICROARRAY: COLON DATA.

- Colon study: Princeton University Gene Expression Project.
- Observations: 40 tumor and 22 normal colon tissues.
- Covariates: 2000 human genes measured using the Affymetrix gene chip.
- Goal:

  identify genes associated with tumor;

  predict tumor risk based on gene measurements;

GOAL

GOAL

- Biologically: only a small number of genes are related to cancer;
- Statistically: more stable estimate with fewer variables;

GOAL

- Biologically: only a small number of genes are related to cancer;
- Statistically: more stable estimate with fewer variables;
- A statistical model/approach with weaker assumptions.
- Well-behaved sparse estimates with built-in gene selection.
- Computationally affordable.

Two Sample Classification: ROC

- Binary outcome $Y = 0, 1$; Continuous, high dimensional covariates X;
- Generalized linear model:

  $Pr(Y = 1|X) = G(\beta' X)$, for an unknown, monotone link function G.
- Classification based on $\beta' X$: e.g., $Y = 1$ if $\beta' X > c$.
- Evaluation: true and false positive rates (TPR and FPR):

  $$TPR(c) = P(\beta' X \geq c|Y = 1) \ \ and \ \ FPR(c) = P(\beta' X \geq c|Y = 0).$$

Two Sample Classification: ROC (Cont.)

- Receiver Operating Characteristic (ROC) Curve: $\{(FPR(c), TPR(c)) : -\infty < c < \infty\}$.
- Classification performance can be evaluated using the area under curve (AUC).

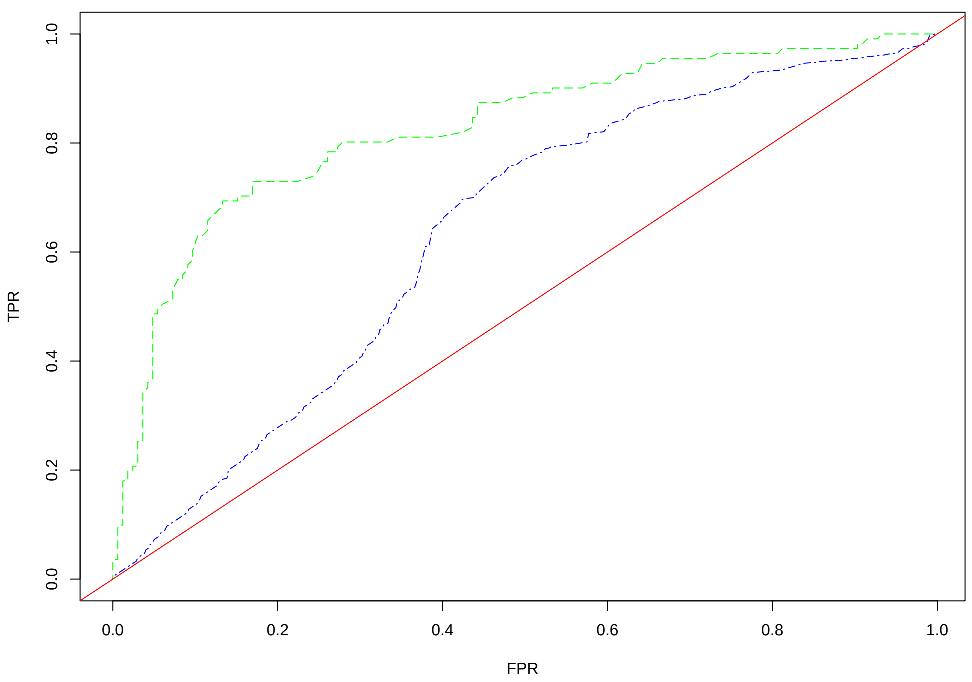

Figure 1: ROC plot.

TWO SAMPLE CLASSIFICATION: ROC (CONT.)

- Denote $\mathbb{D}$ and $\mathbb{H}$ as the index sets for diseased and healthy subjects.
- the empirical AUC is

$$AUC(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} I(\beta' \mathbb{X}_i > \beta' \mathbb{X}_j). \tag{1}$$

- Define the ROC estimate as the maximizer of $AUC(\beta)$.
- Identifiable only up to a scale.
- The objective function not differentiable. With high dimensional covariates, direct maximization of $AUC(\beta)$ is difficult.

Sigmoid Rank Estimator

Approximate the indicator function with the Sigmoid function:
$s(x) = 1/(1 + exp(-x))$.

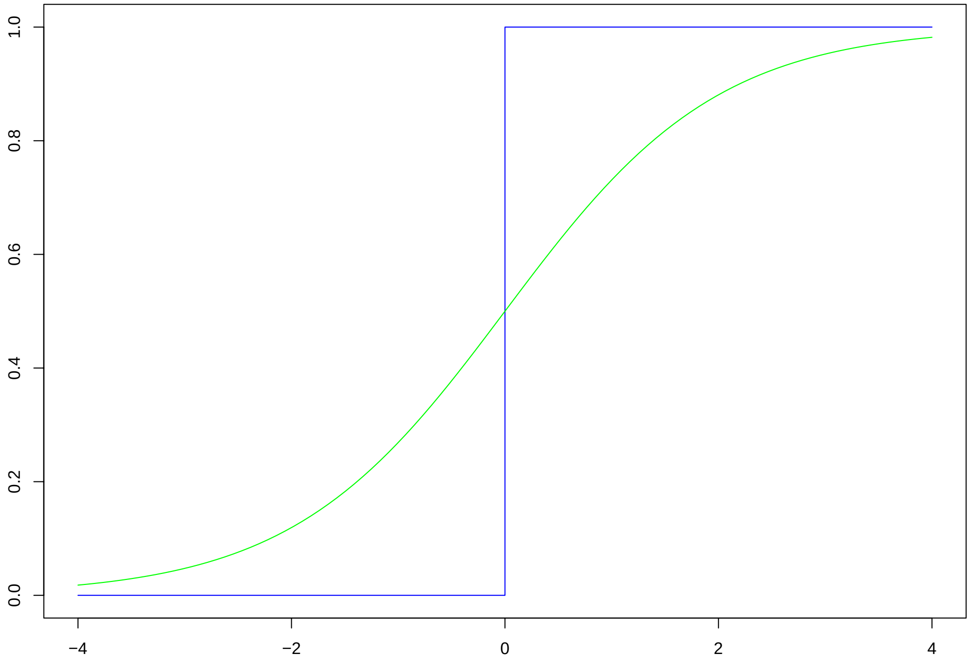

Figure 2: Sigmoid function.

Sigmoid Rank Estimator (Cont.)

Sieve approximation:

Scaled Sigmoid function $s_n(x) = s(x/\sigma_n) = 1/(1 + exp(-x/\sigma_n))$ with $\sigma_n \longrightarrow 0$. Tuning parameter σ_n.

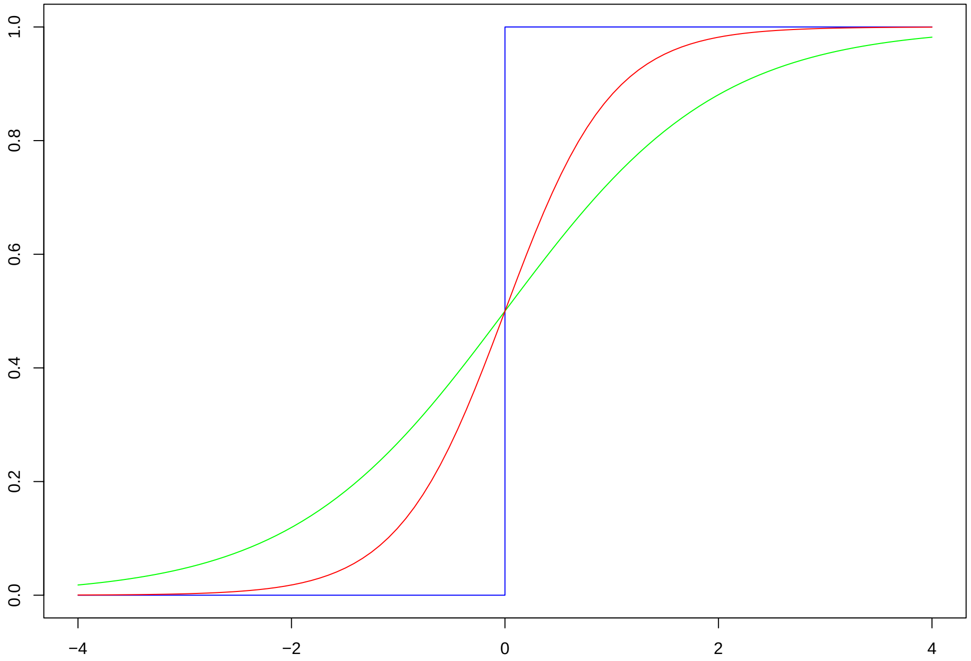

Figure 3: Scaled sigmoid function.

Sigmoid Rank Estimator (Cont.)

- The sigmoid maximum rank correlation (SMRC) estimator:

$$\hat{\beta} = argmax \left\{ R_n(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\}. \tag{2}$$

- For identifiability: we assume $|\hat{\beta}_{(1)}| = 1$.

- The sigmoid function can be replaced by any continuously differentiable K: $\lim_{x \to -\infty} K(x) = 0$ and $\lim_{x \to \infty} K(x) = 1$.

- Similar approximation has been investigated in machine learning studies.

REGULATED SIGMOID ESTIMATE.

Regulated Sigmoid Estimate.

- With the sigmoid approximation, the ”perfect fit” problem is still unsolved.

REGULATED SIGMOID ESTIMATE.

- With the sigmoid approximation, the ”perfect fit” problem is still unsolved.
- Desired properties of an estimating procedure: unique estimate and sparsity.

REGULATED SIGMOID ESTIMATE.

- With the sigmoid approximation, the ”perfect fit” problem is still unsolved.
- Desired properties of an estimating procedure: unique estimate and sparsity.
- Solution ⟶ regulated estimates: the LASSO and the TGDR.

Regulated Sigmoid Estimate: LASSO

REGULATED SIGMOID ESTIMATE: LASSO

- LASSO: least absolute shrinkage and selection operator (Tibshirani, 1996).

REGULATED SIGMOID ESTIMATE: LASSO

- LASSO: least absolute shrinkage and selection operator (Tibshirani, 1996).
- Definition

$$\hat{\beta} = argmax \left\{ R_n(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\},$$

under the L_1 constraint $|\hat{\beta}|_{L_1} \leq u$.

Regulated Sigmoid Estimate: LASSO

- LASSO: least absolute shrinkage and selection operator (Tibshirani, 1996).
- Definition

$$\hat{\beta} = argmax \left\{ R_n(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\},$$

under the L_1 constraint $|\hat{\beta}|_{L_1} \leq u$.

- Tuning parameter u: determines the sparsity of the estimate.

REGULATED SIGMOID ESTIMATE: LASSO

- LASSO: least absolute shrinkage and selection operator (Tibshirani, 1996).
- Definition

$$\hat{\beta} = argmax \left\{ R_n(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}; j \in \mathbb{H}} s_n(\beta'(\mathbb{X}_i - \mathbb{X}_j)) \right\},$$

under the L_1 constraint $|\hat{\beta}|_{L_1} \leq u$.

- Tuning parameter u: determines the sparsity of the estimate.
- Properties: unique and usually sparse estimates. Constraint function not differentiable.

REGULATED SIGMOID ESTIMATE: LASSO (CONT.)

Computational algorithm:

- Quadratic programming (Tibshirani, 1996): not applicable to ”small n, large d” cases.
- LARS (least angle regression, Efron et al. 2004): not directly applicable; number of iterations depends on d.
- We propose using a L_1 boosting based algorithm: simpler computations, faster convergence.

L_1 BOOSTING BASED LASSO

1. Initialization $\beta = (0, \ldots, 0)$ and $m = 0$.
2. Compute $g(\beta)$, the negative derivative of $R_n(\beta)$ w.r.t. β. Denote the p^{th} component of $g(\beta)$ as $g_{(p)}(\beta)$.
3. Find p^* that minimizes $min_p(g_{(p)}(\beta), -g_{(p)}(\beta))$.
4. Denote $\gamma = -sign(g_{(p^*)}(\beta))$. Find $\hat{\alpha} \in [0, 1]$ that minimizes $R_n((1-\alpha)\beta + \alpha \times u \times \gamma\eta_{p^*})$, where η_{p^*} has the p^*th element equals to 1 and the rest components 0.
5. $\beta_{(p)} = (1-\hat{\alpha})\beta_{(p)}$ for $p \neq p^*$, and $\beta_{(p^*)} = (1-\hat{\alpha})\beta_{(p^*)} + \gamma u\hat{\alpha}$. Let $m = m + 1$.
6. Repeat steps 2–5 until convergence.

REGULATED SIGMOID ESTIMATE: TGDR

As an alternative:

Threshold Gradient Directed Regularization (TGDR, Friedman and Popescu, 2004).

http://www-stat.stanford.edu/˜ jhf/PathSeeker.html

- Originally developed for linear regression;
- Now being used in survival analysis;
- Gradient directed; iterative;

Regulated Sigmoid Estimate: TGDR

For any fixed threshold value $0 \leq \tau \leq 1$:

1. Initialize $\beta(0) = 0$ and $\nu_0 = 0$.
2. Compute the negative gradient $g(\nu) = -\partial M(\beta)/\partial\beta$. Denote the j^{th} component of $g(\nu)$ as $g_j(\nu)$.
3. Compute the vector $f(\nu)$ of length p, where the j^{th} component of $f(\nu)$: $f_j(\nu) = I\{|g_j(\nu)| \geq \tau \cdot \max_j |g_j(\nu)|\}$.
4. Update $\beta(\nu + \Delta_\nu) = \beta(\nu) + \Delta_\nu \times g(\nu) \times f(\nu)$ and $\nu = \nu + \Delta_\nu$.
5. Steps 2–4 are repeated S times. S is taken to be a large number to guarantee a full parameter path.

Regulated Sigmoid Estimate: TGDR

Tuning parameters: number of iterations – k and threshold – τ.

Properties of the TGDR:

- $\tau \longrightarrow 0$, the TGDR is close to the ridge regression;
- $\tau \longrightarrow 1$, the TGDR yields sparse estimate (like the LASSO);
- $0 < \tau < 1$, the TGDR produces a full path connecting the ridge regression and the LASSO;

LASSO vs TGDR: Empirical Comparison

	LASSO	TGDR
Model	smaller	
Theoretical properties	clearer	
Classification performance	similar	
Computational burden	similar	
Flexibility		better
Grouping effect		better

COLON DATA.

Data pre-processing:

- Fill in missing values with sample medians;
- Threshold the raw data with a floor of 100 and a ceiling of 16000;
- Genes with $max(expression)/min(expression) < 10$ and/or $max(expression) - min(expression) < 1000$ are also excluded;
- A base 2 logarithmic transformation is then applied;
- Normalize to zero mean and unit variance.

COLON DATA: TGDR.

Tuning parameter selection features.

τ	k	variable	AUC
0.0	448	500	0.943
0.2	440	500	0.946
0.4	479	467	0.959
0.6	638	266	0.946
0.8	1410	74	0.964
1.0	4280	29	0.954

COLON DATA: LASSO AND TGDR ESTIMATES.

- LASSO: $u = 22.5$; identifies 19 genes; AUC = 0.954;
- TGDR: $k = 4280$, $\tau = 1.0$; identifies 29 genes; AUC = 0.954;

COLON DATA: LASSO AND TGDR ESTIMATES.

Note: half table ONLY.

Gene ID	LASSO	TGDR	Gene ID	LASSO	TGDR
Hsa.467	-0.867	-0.922	Hsa.1013	-0.164	–
Hsa.18664	0.236	–	Hsa.8147	–	-1.142
Hsa.81	–	0.534	Hsa.36689	-0.862	-1.619
Hsa.24506	–	-0.456	Hsa.37937	–	-1.332
Hsa.949	0.564	–	Hsa.2487	–	1.428
Hsa.3306	0.819	0.775	Hsa.10047	–	0.448
Hsa.2856	–	0.561	Hsa.692	-0.778	–
Hsa.549	2.127	–	Hsa.8214	–	0.310
Hsa.3016	–	0.930			

Questions? Comments?

Thank You!