Generalized t-statistic and AUC for binary classification

Osamu Komori
University of Fukui
Shinto Eguchi
The Institute of Statistical Mathematics

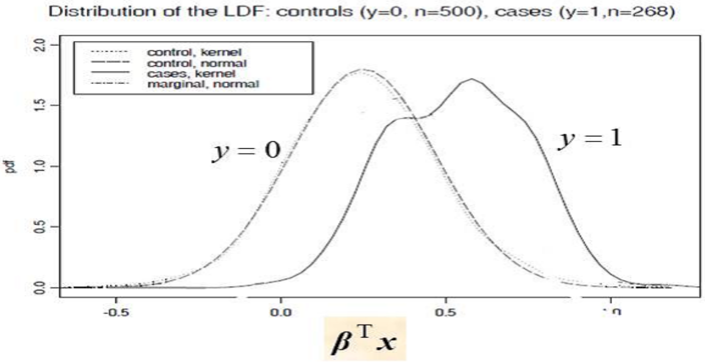
Humanities and Social Sciences Center, Academia Sinica, Taiwan
December 11, 2015

Contents

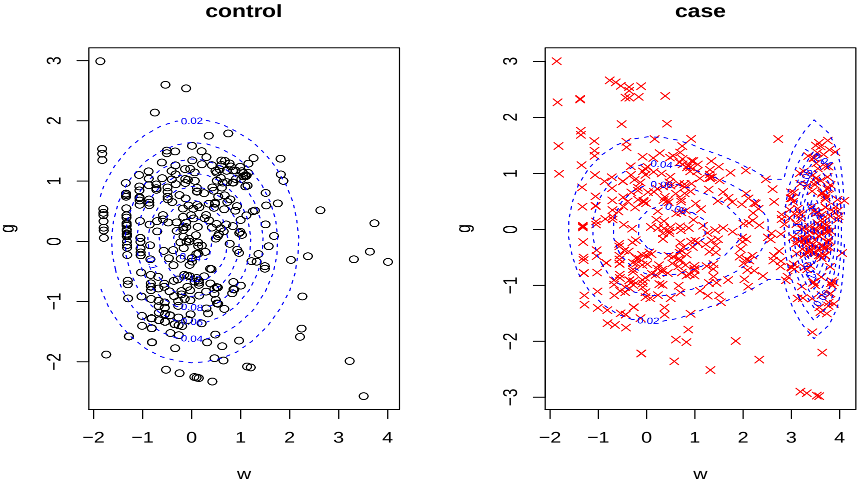
How to deal with the heterogeneity of sample of disease group ?

$y = 0$ (control); $y = 1$ (diabetes)

$\boldsymbol{x} =$ (glucouse level, BMI, diabetes measure)



Gene expression data of asthmatic markers (Dottorini *et al.*, 2011)



Problem setting

- We focus on a linear discriminant function: $F(x) = \beta^{\mathrm{T}} x$, where $x \in \mathbb{R}^p$
- We assume that the sample of controls ($y = 0$) are normally distributed:
$$x_0 \sim N(\mu_0, \Sigma_0),$$
where we apply log transformation if necessary.
  - We summarize the information of the sample into the sample mean $\overline{x_0}$, and sample variance S_0.
- However, we recognize the distribution of the cases sample ($y = 1$) is far from normality.
  - We take it into consideration more flexibly.
- We propose the generalized t statistic based on U function.

Contents

Generalization of t-statistic

For two samples $\{x_{0i} : i = 1, \ldots, n_0\}$ and $\{x_{1j} : j = 1, \ldots, n_1\}$, we consider the following statistic based on $F(x) = \beta^{\mathrm{T}} x$.

Generalized t-statistic (Komori *et al.*, 2015)

$$L_U(\beta) = \frac{1}{n_1} \sum_{j=1}^{n_1} U\left\{ \frac{\beta^{\mathrm{T}}(x_{1j} - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}} \right\},$$

where U is an arbitrary function: $\mathbb{R} \to \mathbb{R}$; $\bar{x}_y$, S_y are conditional sample mean and sample variance given y.

$$\mathbb{L}_U(\beta) = E_1\left[U\left\{ \frac{\beta^{\mathrm{T}}(x - \mu_0)}{\beta^{\mathrm{T}} \Sigma_0 \beta} \right\} \right],$$

where E_y, μ_y, Σ_y are conditional mean and variance given y.

t-statistic, AUC, Fisher, K-L divergence

1. t statistic: if $U(w) = w$, then

$$L_{\mathrm{I}}(\beta) = \frac{\beta^{\mathrm{T}}(\bar{x}_1 - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}}.$$

2. AUC: If $U(w) = \Phi(w)$(Su and Liu, 1993), then

$$L_{\Phi}(\beta) = \frac{1}{n_1} \sum_{j=1}^{n_1} \Phi\left\{\frac{\beta^{\mathrm{T}}(x_{1j} - \bar{x}_0)}{(\beta^{\mathrm{T}} S_0 \beta)^{1/2}}\right\} \to \mathrm{AUC}(\beta),$$

3. Fisher: If $\hat{U}(w) = -(w - \hat{c}_0)^2$, then

$$\operatorname*{argmax}_{\beta \in \mathbb{R}^p} L_{\hat{U}}(\beta) \propto (\hat{\pi}_0 S_0 + \hat{\pi}_1 S_1)^{-1}(\bar{x}_1 - \bar{x}_0).$$

4. K-L divergence: If $U(w) = U_{\mathrm{opt}}(w)$, then

$$\mathbb{L}_{U_{\mathrm{opt}}}(\beta) = \int f_1(w) \log \frac{f_1(w)}{\phi(w, \mu_w, \sigma_w^2)} dw.$$

Three assumptions

- normality assumption of data for $y = 0$: $x_0 \sim N(\mu_0, \Sigma_0)$.
- consistency

$$\text{(A)} \qquad E_1(g \mid w = a) = 0 \quad \text{for all } a \in \mathbb{R},$$

- asymptotic variance

$$\text{(B)} \qquad \mathrm{var}_1(g \mid w = a) = Q_0 \quad \text{for all } a \in \mathbb{R},$$

where $w = \beta_0^{\mathrm{T}}(x - \mu_0)$, $g = Q_0(x - \mu_0)$, $Q_0 = I_p - \beta_0\beta_0^{\mathrm{T}}$. The target parameter is defined as.

$$\beta_0 = \frac{\Sigma_0^{-1}(\mu_1 - \mu_0)}{\{(\mu_1 - \mu_0)^{\mathrm{T}}\Sigma_0^{-1}(\mu_1 - \mu_0)\}^{1/2}} \approx \beta_{\mathrm{F}},$$

where $\beta_{\mathrm{F}} = (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)/\{(\mu_1 - \mu_0)^{\mathrm{T}}(\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)\}^{1/2}$ and we can assume $\Sigma_0 = I_p$ and $\mu_0 = 0$ in general.

Consistency

We consider the estimator that maximizes the generalized t-statistic:

$$\widehat{\beta}_U = \underset{\beta \in \mathbb{R}^p}{\text{argmax}}\, L_U(\beta).$$

Theorem 1

Under Assumption (A) $\widehat{\beta}_U$ is consistent to β_0 for any U.

Proof.
Using $w = \beta_0^{\mathrm{T}}(x - \mu_0)$ and $g = Q_0(x - \mu_0)$, we have

$$\frac{\partial}{\partial \beta}\mathbb{L}(\beta) = E_1[U'(w)g],$$

which is 0 if $\beta = \beta_0$ on account of the assumption (A). Hence from the strong law of large numbers, $\widehat{\beta}_U$ is asymptotically consistent with β_0.

What is assumption (A)?

Proposition 1

If it holds that

$$p_1(x) = p_1((I - Q_0)(x - \mu_0) + \mu_0),$$

then (A) is satisfied where $p_1(x)$ is the probability density function of x given $y = 1$.

This means that $p_1(x)$ is symmetrical with respect to $\mu_0 + a\beta_0$ ($a \in \mathbb{R}$). That is, it covers a wide range of distributions such as the elliptical distributions including the multivariate t-distribution with mean μ_1 and the precision matrix I_p.

Gaussian mixture model

$$p_1(x) = \sum_{k=1}^{\infty} \epsilon_k \phi(x, \nu_k, V_k).$$

Proposition 2

Assumptions (A) and (B) under the infinite mixture model are rewritten as

(A′) $$\sum_{k \in K_\ell} \epsilon_k (Q_0 - Q_k) = 0, \quad \sum_{k \in K_\ell} \epsilon_k Q_k (\nu_k - \mu_0) = 0 \text{ for any } \ell \in \mathbb{N}$$

(B′) $$\sum_{k \in K_\ell} \epsilon_k \Big\{ Q_k V_k Q_0 - Q_0 \Big\} = 0 \text{ for any } \ell \in \mathbb{N}$$

where $Q_k = I_p - V_k \beta_0 \beta_0^{\mathrm{T}} / (\beta_0^{\mathrm{T}} V_k \beta_0)$ and $K_\ell = \{k \mid \beta_0^{\mathrm{T}} \nu_k = \beta_0^{\mathrm{T}} \nu_\ell,\ \beta_0^{\mathrm{T}} V_k \beta_0 = \beta_0^{\mathrm{T}} V_\ell \beta_0\}$.

Illustration of typical examples

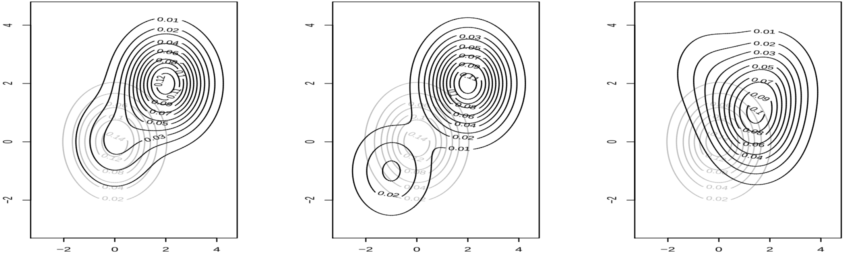


Figure 1: Contour plots of probability densities of $y = 0$ in gray and $y = 1$ in black, which satisfy Assumptions (A) and (B). For all three panels, $\mu_0 = (0, 0)^{\mathrm{T}}$, $\Sigma_k = \Sigma_0 = \mathrm{diag}(1, 1)$ for all k. $\nu_1 = (0, 0)^{\mathrm{T}}$, $\nu_2 = (2, 2)^{\mathrm{T}}$, $\epsilon_1 = 0 \cdot 2$ and $\epsilon_2 = 0 \cdot 8$ in the left panel; $\nu_1 = (-1, -1)^{\mathrm{T}}$, $\nu_2 = (2, 2)^{\mathrm{T}}$, $\epsilon_1 = 0 \cdot 2$ and $\epsilon_2 = 0 \cdot 8$ in the middle panel; $\nu_1 = (1, 1)^{\mathrm{T}}$, $\nu_2 = (1 \cdot 5, 0 \cdot 5)^{\mathrm{T}}$, $\nu_3 = (0 \cdot 5, 1 \cdot 5)^{\mathrm{T}}$, $\nu_4 = (-0 \cdot 5, 2 \cdot 5)$, $\nu_5 = (2, 2)^{\mathrm{T}}$, $\epsilon_1 = \epsilon_3 = \epsilon_4 = 0 \cdot 1$, $\epsilon_2 = 0 \cdot 4$ and $\epsilon_5 = 0 \cdot 3$ in the right panel

Semiparametric model

Theorem 2

Under the assumption $x_0 \sim N(\mu_0, \Sigma_0)$, if it holds that

$$p_1(x) = \psi(c + \beta^\top x) p_0(x),$$

then

1. *β is proportional to the target parameter β_0*
2. *assumptions (A) and (B) hold.*

If we consider $\psi(z) = \exp(z)$, it corresponds to logistic linear model.
ψ is an arbitrary non-parametric function.

Contents

Asymptotic variance

Let $f_1(w)$ be a probability density function of $w = \beta_0^{\mathrm{T}}(x - \mu_0)$ given $y = 1$.

Theorem 3

Under assumptions (A) and (B), $n_1^{1/2}(\widehat{\beta}_U - \beta_0)$ *is asymptotically distributed as* $N(0, \Sigma_U)$*, where*

$$\begin{aligned} \Sigma_U &= c_U Q_0^-, \\ c_U &= \frac{E_1\{U'(w)^2\} + \pi_1/\pi_0[E_1\{U'(w)w\}]^2 + \pi_1/\pi_0[E_1\{U'(w)\}]^2}{\left[E_1\{U'(w)S(w)\} + E_1\{U'(w)w\}\right]^2}, \end{aligned}$$

where Q_0^- *is the generalized inverse of* Q_0*,* $S(w) = \partial \log f_1(w)/\partial w$ *and* U' *is the first derivative of* U*.*

Optimal U function

Theorem 4

The optimal U that minimizes the asymptotic variance of $\widehat{\beta}_U$ is given as

$$U_{\mathrm{opt}}(w) = \log \frac{f_1(w)}{\phi(w, \mu_w, \sigma_w^2)},$$

where $\mu_w = E(w)$, $\sigma_w^2 = \mathrm{var}(w)$. Moreover, we have

$$\min_U c_U = \frac{\sigma_w^2}{\mu_{1,S^2} - 1 + (\pi_0 \mu_{1,w}^2 + \sigma_{1,w}^2 - 1)(\pi_0 + \pi_1 \mu_{1,S^2})},$$

where $\mu_{1,w} = E_1(w)$, $\sigma_{1,w}^2 = E_1\{(w - \mu_{1,w})^2\}$, $\mu_{1,S^2} = E_1\{S(w)^2\}$.

Optimality in terms of AUC

The expected generalized t-statistic asymptotically satisfies

$$\mathrm{E}\{\mathbb{L}_U(\hat\beta_1)\} - \mathrm{E}\{\mathbb{L}_U(\hat\beta_2)\} = \frac{1}{2n_1}\mathrm{tr}[H_U(\beta_0)\{\mathrm{var}_{\mathrm{A}}(\hat\beta_1) - \mathrm{var}_{\mathrm{A}}(\hat\beta_2)\}] \geq 0$$

optimality of $\hat\beta_{U_{\mathrm{opt}}}$

$$\mathrm{E}\{\mathbb{L}_U(\hat\beta_{U_{\mathrm{opt}}})\} \geq \mathrm{E}\{\mathbb{L}_U(\hat\beta)\}.$$

For example, if we take $\Phi(w)$ as $U(w)$, then we have

$$H_\Phi(\beta_0) = -2\int \phi(w) w f_1(w) dw Q_0.$$

Here we have $\int w f_1(w)dw = E_1(w) = (\mu_1^{\mathrm{T}}\mu_1)^{1/2} > 0$. This implies that $\int \phi(w) w f_1(w) dw > 0$ because of the symmetry of $\phi(w)$ with respect to the original point. Hence, the estimator $\hat\beta_{U_{\mathrm{opt}}}$ asymptotically has a maximum value of AUC.

Algorithm for estimation of β_0

1. Initialize as $\beta^{(1)} = S_0^{-1}(\bar{x}_1 - \bar{x}_0)$.
2. For $t = 2, \cdots, T$,
   - Estimate $f_1(w)$ based on kernel method to produce $\hat{U}_{\rm opt}(w)$.
   - Update $\beta^{(t-1)}$ to $\beta^{(t)}$ as

$$\beta^{(t)} = \underset{\beta \in \mathbb{R}^p}{\rm argmax} \frac{1}{n_1} \sum_{j=1}^{n_1} \hat{U}_{\rm opt}\left\{ \frac{\beta^{\rm T}(x_{1j} - \bar{x}_0)}{(\beta^{\rm T} S_0 \beta)^{1/2}} \right\}$$

3. Output $\widehat{\beta}_U = \beta^{(T)}$.

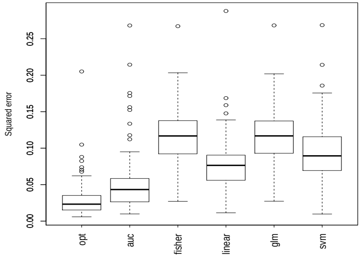
Note that the initial value $\beta^{(1)}$ in step 1 above could be replaced by any other value, so avoiding the need to calculate the inverse of S_0.

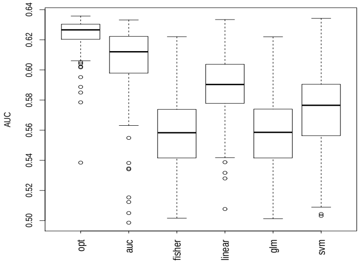
Simulation

Setting

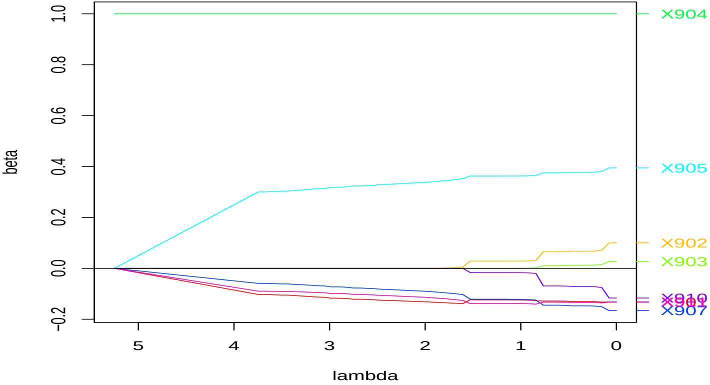
$$x_0 \sim N(\mathbf{0}, \boldsymbol{I}_p),\ \ x_1 \sim (1-\epsilon_1-\epsilon_2)N(\boldsymbol{\nu}_0, \boldsymbol{V}_0) + \epsilon_1 N(\boldsymbol{\nu}_1, \boldsymbol{V}_1) + \epsilon_2 N(\boldsymbol{\nu}_2, \boldsymbol{V}_2),$$

where $\epsilon_1 = \epsilon_2 = 0.1$, $\boldsymbol{\nu}_0 = (-1, -0.1, \ldots, -0.1)^\top \in \mathbb{R}^p$, $\boldsymbol{\nu}_1 = (1, 0.1 \ldots, 0.1)^\top$
$\boldsymbol{\nu}_2 = (3, 0.3 \ldots, 0.3)^\top \in \mathbb{R}^p$, $\boldsymbol{V}_0 = \boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{I}_p$ and $\pi_0 = \pi_1$, $p = 10$ and $n = 200$.





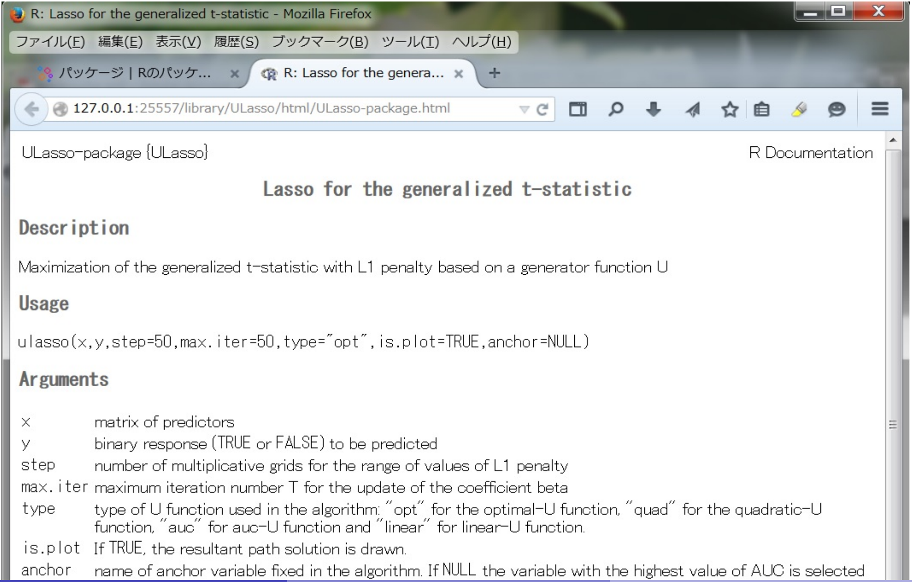
U-lasso



Solution paths by $U_{\rm opt}$-lasso for variables in X group

$$L_U^{\lambda}(\beta) = L_U(\beta) - \lambda \sum_{k=1}^{p} |\beta_k|,$$

R package of U-lasso



ULasso-package {ULasso} R Documentation

Lasso for the generalized t-statistic

Description

Maximization of the generalized t-statistic with L1 penalty based on a generator function U

Usage

```
ulasso(x,y,step=50,max.iter=50,type="opt",is.plot=TRUE,anchor=NULL)
```

Arguments

x	matrix of predictors
y	binary response (TRUE or FALSE) to be predicted
step	number of multiplicative grids for the range of values of L1 penalty
max.iter	maximum iteration number T for the update of the coefficient beta
type	type of U function used in the algorithm: "opt" for the optimal-U function, "quad" for the quadratic-U function, "auc" for auc-U function and "linear" for linear-U function.
is.plot	If TRUE, the resultant path solution is drawn.
anchor	name of anchor variable fixed in the algorithm. If NULL the variable with the highest value of AUC is selected

Contents

Generalization of Area under the ROC curve (AUC)

Fro two samples $\{x_{0i} : i = 1, \ldots, n_0\}$ and $\{x_{1j} : j = 1, \ldots, n_1\}$, we consider a linear predictor $F(x) = \beta^{\mathrm{T}} x$ and propose

Generalized AUC

$$L_U(\beta) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} U\left\{\frac{\beta^{\mathrm{T}}(x_{1j} - x_{0i})}{(\beta^{\mathrm{T}} S \beta)^{1/2}}\right\},$$

where U is a generator function: $\mathbb{R} \to \mathbb{R}$; $\overline{x_y}$ is a conditional sample mean of x given y; $S = S_0 + S_1$.

$$\widehat{\beta}_U = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmax}} \, L_U(\beta)$$

Two assumptions

- consistency

$$\text{(A)} \qquad E_y(g_y \mid w_y = a) = 0 \quad \text{for all } a \in \mathbb{R}, \ \text{for } y = 0, 1$$

- asymptotic variance

$$\text{(B)} \qquad \mathrm{var}_y(g_y \mid w_y = a) = \Sigma_y^* \quad \text{for all } a \in \mathbb{R}, \ \text{for } y = 0, 1$$

where $w_y = \beta_{\mathrm{F}}^{\mathrm{T}} x_y$, $g_y = Q x_y$, $Q = I - \beta_{\mathrm{F}} \beta_{\mathrm{F}}^{\mathrm{T}}$, $\Sigma_y^* = Q \Sigma_y Q^{\mathrm{T}}$. And we define a target parameter of β as

$$\beta_{\mathrm{F}} = \frac{\Sigma^{-1}(\mu_1 - \mu_0)}{\{(\mu_1 - \mu_0)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_0)\}^{1/2}},$$

where we assume $\Sigma = \Sigma_0 + \Sigma_1 = I_p$ and $\mu_0 + \mu_1 = 0$ without loss of generality

The target parameter is the coefficient of Fisher linear predictor.
No normality assumption of x_0.

Gaussian mixture

We consider Gaussian mixture such that

$$p_y(x) \;=\; \sum_{k=1}^{\infty} \epsilon_{yk}\phi(x, \nu_{yk}, V_{yk}) \text{ for } y = 0, 1.$$

Proposition 3

Then assumption (A) and (B) are rewritten

$$\text{(A}'\text{)} \qquad \sum_{k \in K_{y\ell}} \epsilon_k (Q - Q_{yk}) = 0, \;\; \sum_{k \in K_{y\ell}} \epsilon_{yk} Q_{yk} \nu_{yk} = 0, \;\text{ for } {}^\forall \ell \in \mathbb{N}, \; y = 0, 1$$

$$\text{(B}'\text{)} \qquad \sum_{k \in K_{y\ell}} \epsilon_{yk} \left\{ Q_{yk} V_{yk} Q - Q \Sigma_y Q \right\} = 0, \;\text{ for } {}^\forall \ell \in \mathbb{N}, y = 0, 1$$

where $Q_{yk} = I_p - V_{yk}\beta_{\mathrm{F}}\beta_{\mathrm{F}}{}^\top / (\beta_{\mathrm{F}}{}^\top V_{yk}\beta_{\mathrm{F}})$, $K_{y\ell} = \{k \mid \beta_{\mathrm{F}}{}^\top \nu_{yk} = \beta_{\mathrm{F}}{}^\top \nu_{y\ell}, \; \beta_{\mathrm{F}}{}^\top V_{yk}\beta_{\mathrm{F}} = \beta_{\mathrm{F}}{}^\top V_{y\ell}\beta_{\mathrm{F}}\}$.

Semiparametric model

Theorem 5

Let ψ_y be a function: $\mathbb{R} \to \mathbb{R}_+$ such that

$$p_y(x) = \psi_y(c + \beta^\top x)\phi(x, 0, \Sigma_y), \text{ for } y = 0, 1,$$

and

$$\Sigma_y \beta = \lambda_y \beta, \text{ for } y = 0, 1.$$

where $\lambda_y (\neq 0)$. Then

1. *β is proportional to β_{F}*
2. *assumption (A) and (B) are satisfied,*

where $\phi(x, \mu, \Sigma)$ is a normal distribution of mean μ and variance Σ. ψ_y is an arbitrary non-parametric function.

Asymptotic variance

Let $f(w)$ be a density function of $w = w_1 - w_0 = \beta_{\mathrm{F}}^{\mathrm{T}}(x_1 - x_0)$.

Theorem 6

Under assumptions (A) and (B) with $\Sigma_0^*/\pi_0 = \Sigma_1^*/\pi_1$, $n^{1/2}(\widehat{\beta}_U - \beta_{\mathrm{F}})$ is asymptotically distributed as $N(0, \Sigma_U)$

$$\Sigma_U = c_U Q^-,$$

$$c_U = \frac{E_0\Big[E_1\{U'(w)\}\Big]^2 + E_1\Big[E_0\{U'(w)\}\Big]^2 + 2\rho E\{U'(w)\}E\{U'(w)w\} - \Big[E\{U'(w)w\}\Big]^2}{\Big[E\{U'(w)S(w) + U'(w)w\}\Big]^2}.$$

where Q^- is the generalized inverse of Q; $\Sigma_y^* = Q\Sigma_y Q^{\mathrm{T}}$; $S(w) = \partial \log f(w)/\partial w$; U' is the first derivative of U and $\rho = E(w)$.

Optimal U function

By variational method, the optimal-U minimizing the asymptotic variance should satisfy

$$E_0[U'(w)] + E_1[U'(w)] = \lambda S(w) + aw + b,$$

where $w = w_1 - w_0$; $S(w) = \partial \log f_1(w)/\partial w$; λ, a, b are some constants.

Remark 1

Note that there does not exist $U(w)$ if $S(w)$ is a non-linear function.

$\Rightarrow$ 「No optimal-U for generalized AUC in general」

- β_0 is easy to estimate efficiently (generalized t-statistic)
- β_{F} is difficult to estimate efficiently (generalized AUC)

upper-U

The scalar term c_U in asymptotic variance is upper-bounded by

$$c_U \leq \frac{2E\{U'(w)^2\} + 2\rho E\{U'(w)\}E\{U'(w)w\} - \left[E\{U'(w)w\}\right]^2}{\left[E\{U'(w)S(w) + U'(w)w\}\right]^2},$$

where the equality holds when $U(w) = aw + b$.

Proposition 4

The upper-bound is minimized by

$$U_{\mathrm{upper}}(w) = \log f(w) + \frac{1}{2}w^2 - \frac{\rho^3}{2+\rho^2}w.$$

Based on $U_{\mathrm{upper}}(w)$ we construct optimal-U by polynomial approximation

$$U_{\mathrm{opt}}(w) = U_{\mathrm{upper}}(w) + a_1 w + a_2 w^2 + \cdots + a_m w^m,$$

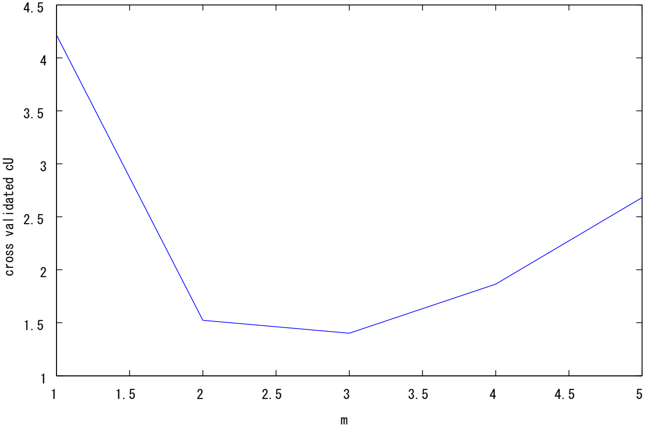
Optimal order of polynomial approximation

$$c_{U_m^{(k)}} = \frac{\overline{E}_0^{(k)}\left[\overline{E}_1^{(k)} U_m^{(k)\prime}(w)\right]^2 + \overline{E}_1^{(k)}\left[\overline{E}_0^{(k)} U_m^{(k)\prime}(w)\right]^2 + 2\hat{\rho}\overline{E}^{(k)}\{U_m^{(k)\prime}(w)\}\overline{E}^{(k)}\{U_m^{(k)\prime}(w)w\} - \left[\overline{E}\{U_m^{(k)\prime}(w)w\}\right]^2}{\left[\overline{E}^{(k)}\{U_m^{(k)\prime}(w)S(w) + U_m^{(k)\prime}(w)(w)\}\right]^2}$$

where $\overline{E}^{(k)}U'(w) = 1/(n_0^{(k)}n_1^{(k)})\sum_{i\in I_k}\sum_{j\in J_k} U'(w_{1j} - w_{0i})$, $\overline{E}_0^{(k)}U'(w) = 1/n_0^{(k)}\sum_{i\in I_k} U'(w_{1j} - w_{0i})$, $\overline{E}_1^{(k)}U'(w) = 1/n_1^{(k)}\sum_{j\in J_k} U'(w_{1j} - w_{0i})$. And $n_0^{(k)} and n_1^{(k)}$ are numbers of elements of I_k and J_k, respectively, where

$$\begin{aligned} I_k \cap I_{k'} &= \varnothing\ (k \neq k'),\ \bigcup_{k=1}^{K} I_k = \{1, \ldots, n_0\} \\ J_k \cap J_{k'} &= \varnothing\ (k \neq k'),\ \bigcup_{k=1}^{K} J_k = \{1, \ldots, n_1\}. \end{aligned}$$

Cross validation



Plot of $c_{U_m^{(k)}}$ against m

Summary

1. We propose generalized t-statistic and derive an optimal-U minimizing asymptotic variance. The lasso-type method is also considered to allow for high dimensional data analysis.
2. In order to allow for heterogeneity for both populations, we consider generalized AUC and its approximated optimal U.
3. We have confirmed that our proposed methods work well in simulation studies as well as real data analysis (not shown in details).

Contents

Discussion 1

Fisher linear discriminant analysis

$$F(x) = \widehat{\beta}_{\mathrm{F}}^{\top} x + c,$$

where $\widehat{\beta}_{\mathrm{F}} = (S_0 + S_1)^{-1}(\overline{x}_1 - \overline{x}_0)$ and c is a constant.

1. It is proposed by Ronald A. Fisher (Fisher, 1936).
2. It is derived by maximizing the ratio of the variance between the two classes to the variance within the classes.
3. It is still valid and useful in real data analysis (Dudoit *et al.*, 2002; Hess *et al.*, 2006)
4. Regularized LDA (Guo *et al.*, 2007; Witten and Tibshirani, 2011), LDA in the reproducing kernel Hilbert space (Mika *et al.*, 1999) and LDA with Lasso (Trendafilov and Jolliffe, 2007)

Discussion 2: Breast cancer data analysis

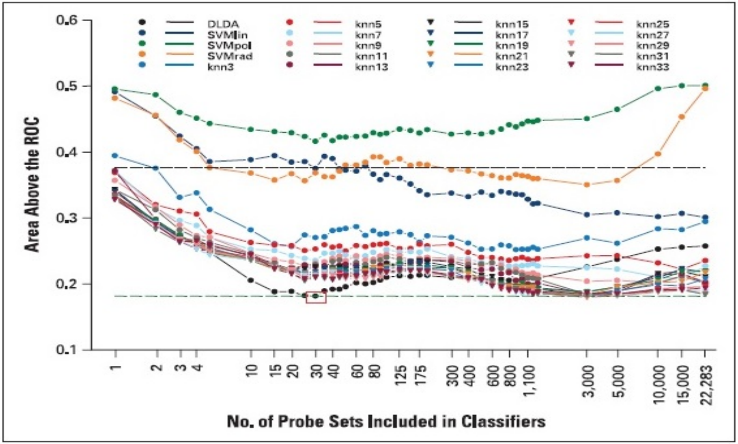


Figure 2: Mean area above the ROC curves plotted against the number of top genes included in the classifiers (Hess *et al.*, 2006)

Discussion 3: Consistency of $\widehat{\beta}_U$ to β_{F}

The important assumption is the one about consistency

(A) $$E_y(g_y \mid w_y = a) = 0 \quad \text{for all } a \in \mathbb{R}, \text{ for } y = 0, 1$$

For practical purpose, we can omit assumption (B)

(B) $$\mathrm{var}_y(g_y \mid w_y = a) = \Sigma_y^* \quad \text{for all } a \in \mathbb{R}, \text{ for } y = 0, 1$$

In that case we need the optimization regarding the asymptotic variance (matrix)

$$U_{\mathrm{opt}} = \operatorname*{argmin}_U |\Sigma_U|,$$

where U can be modeled using natural cubic spline or sigmoid function with some scale parameter.

Discussion 4: open problems

1. How far can Fisher linear discriminant analysis be extended by $F(x) = \widehat{\beta}_U^\top x$? Especially in high dimensional data analysis?
2. What are conditions of probability density function $p_0(x)$ and $p_1(x)$ such that $\widehat{\beta}_U$ has consistency to β_F?
3. How do we derive the optimal-U to estimate β_{F}?

Bibliography I

- Dottorini, T., Sole, G., Nunziangeli, L., Baldracchini, F., Senin, N., Mazzoleni, G., Proietti, C., Balaci, L. and Crisanti, A. (2011). Serum IgE reactivity profiling in an asthma affected cohort. *PLoS ONE* **6**, e22319.
- Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**, 77–87.
- Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* **7**, 179–188.
- Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100.
- Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N. and Pusztai, L. (2006). Pharmacogenomic prediction of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology* **24**, 4236–4244.
- Komori, O., Eguchi, S. and Copas, J. B. (2015). Generalized *t*-statistic for two-group classification. *Biometrics* **71**, 404–416.

Bibliography II

- Mika, S., Fitscht, G., Weston, J., Scholkopft, B. and Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. *In Proceedings, IEEE Workshop on Neural Networks for Signal Processing.*
- Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.
- Trendafilov, N. T. and Jolliffe, I. T. (2007). DALASS: Variable selection in discriminant analysis via the LASSO. *Computational Statistics & Data Analysis* **51**, 3718–3736.
- Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society* **73**, 753–772.