Finding best linear combination of markers for a medical diagnostic with restricted false positive rate

Yuan-chin I. Chang

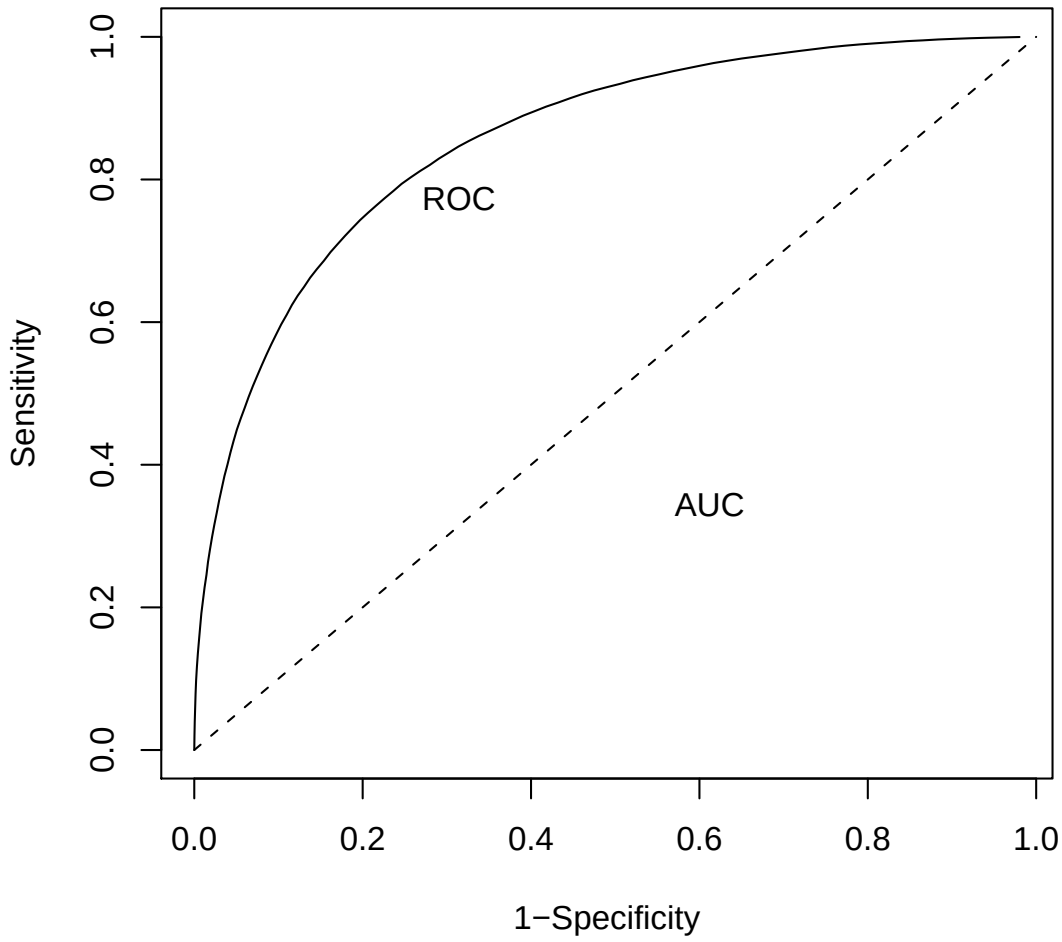
Academia Sinica, Taipei, Taiwan
ycchang@sinica.edu.tw

ROC curve

- Let Z be continuous output of a classifier for a given subject.
- If $Z > c$, for a pre-specified threshold c, then we classify the subject to be positive.
- The true positive rate (TPR) is $S_D(c) = Pr(Z > c|\text{diseased})$.
- The false positive rate (FPR) is $S_{\bar{D}}(c) = Pr(Z > c|\text{non-diseased})$.
- ROC curve is a plot of $\{(S_{\bar{D}}(c), S_D(c)), c \in (-\infty, \infty)\}$;
- Or, $ROC(u) = S_D(S_{\bar{D}}^{-1}(u))$.
- The area under ROC curve,

$$AUC = \int_0^1 ROC(t)dt. \tag{1}$$

ROC curve plot



Notations

- Consider the two-group classification problem with sizes n and m for the diseased and normal groups, respectively.
- Suppose each subject has p features. Let Y and X denote, respectively, the p dimensional feature vectors of the normal and diseased groups.
- Let $l \in R^p$ be the vector of linear combination coefficients and $l'X$ and $l'Y$ are then called linear risk scores.
- We want to construct a classifier based on those linear risk scores such that the classifier satisfies some performance measure.

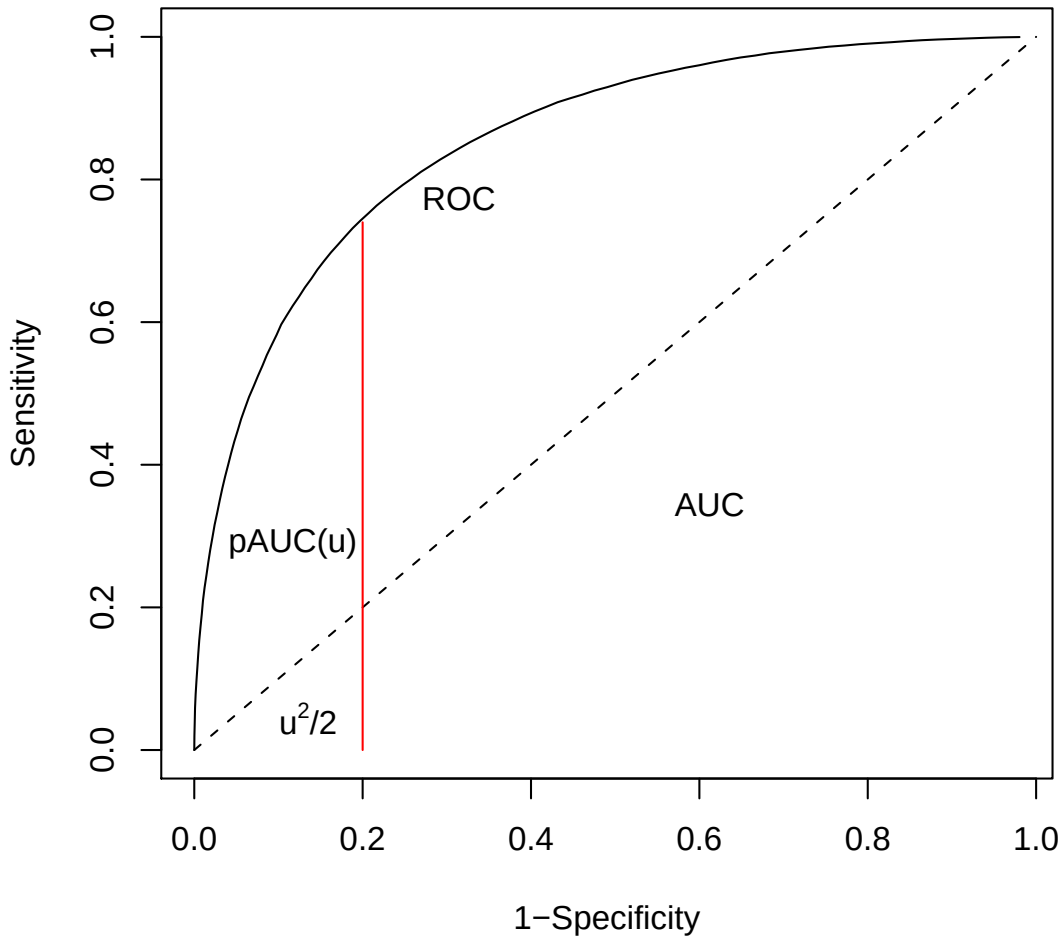
Motivation

- Not the whole range of FPRs is of interest; especially, when there is only certain value of FPRs is allowed (i.e. FPR$\in$ (0, 0.1)).
- The partial area under ROC curve is defined as

$$pAUC(u) = \int_0^u ROC(t)dt$$

when the maximum of FPR is no greater than u, for a given u.

pAUC plot



Disadvantage of partial AUC?

- Threshold dependent (the specificity need to be determined in advance.)

Parametric Approach

- Under normality assumption, and let D and $\bar{D}$ be random variables from disease and non-disease group and assume that they follow from multivariate normal distributions with means μ_D and $\mu_{\bar{D}}$ and covariance matrices Σ_D an $\Sigma_{\bar{D}}$, then the vector of the best linear combination coefficients of markers that maximizing AUC is (Su and Liu, 1993)

$$\ell_a = (\Sigma_D + \Sigma_{\bar{D}})^{-1}(\mu_D - \mu_{\bar{D}})$$

- Select markers based on the absolute value of coefficients of ℓ_a.

Linear Combination that maximizes the partial AUC

1. As an extension, we found that the best linear combination that maximizes the partial area under the curve for a given specificity is

$$\ell_p = (\omega_D \Sigma_D + \omega_{\bar{D}} \Sigma_{\bar{D}})^{-1} (\mu_D - \mu_{\bar{D}})$$

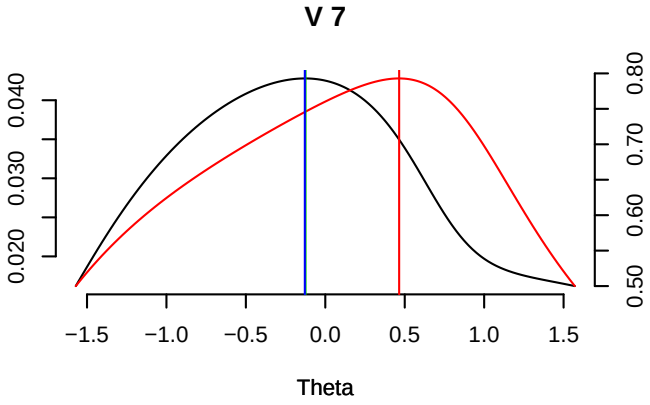
   where ω_D and $\omega_{\bar{D}}$ depends on the given specificity and also function of ℓ_p.

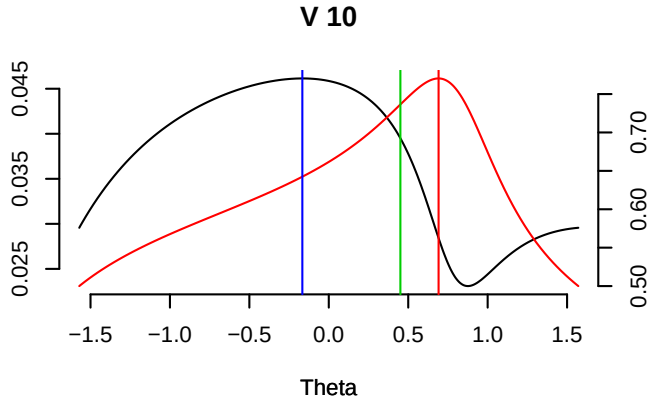
2. $\omega_D = c_1 \frac{\ell_p' \Delta_\mu}{Q_D + Q_{\bar{D}}} + c_2 Q_{\bar{D}}$ and $\omega_{\bar{D}} = c_1 \frac{\ell_p' \Delta_\mu}{Q_D + Q_{\bar{D}}} - c_2 Q_D$ and $\Delta_\mu = \mu_D - \mu_{\bar{D}}$, $Q_A = \ell' \Sigma_A \ell$ for $A \in \{D, \bar{D}\}$.

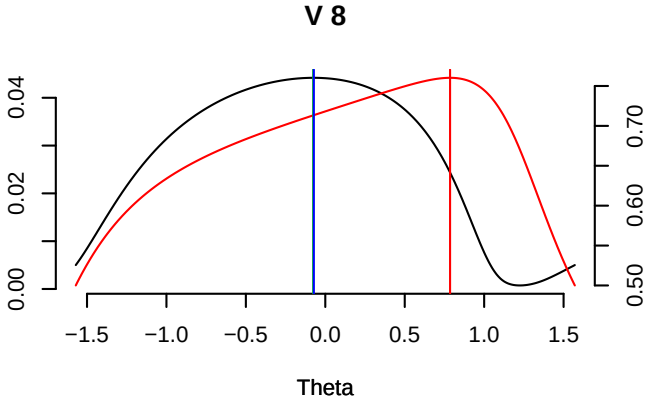
3. Both c_1 and c_2 depend on ℓ_p.

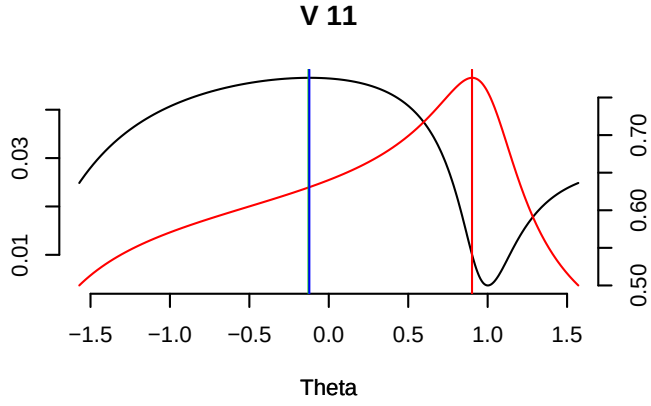
4. When the given specificity is 1, c_1 is a constant and $c_2 = 0$, and then the formula above is identical to the result of Su and Liu (1993).

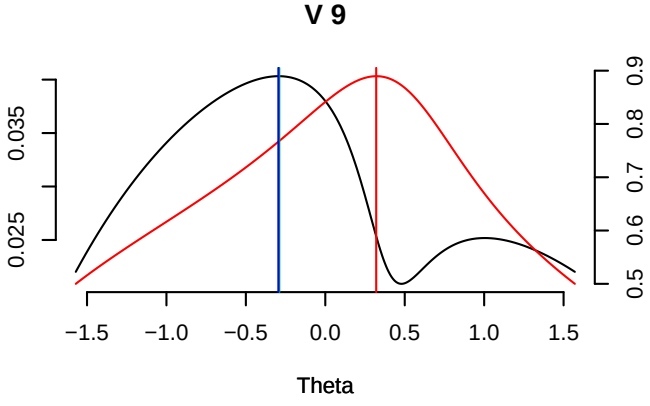
   Thus, the solution of ℓ_p requires some iteration procedure.

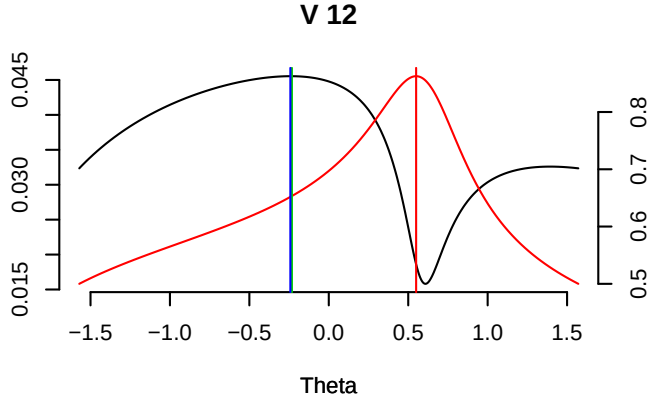

V 7
0.020
0.030
0.040
0.50
0.60
0.70
0.80
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta

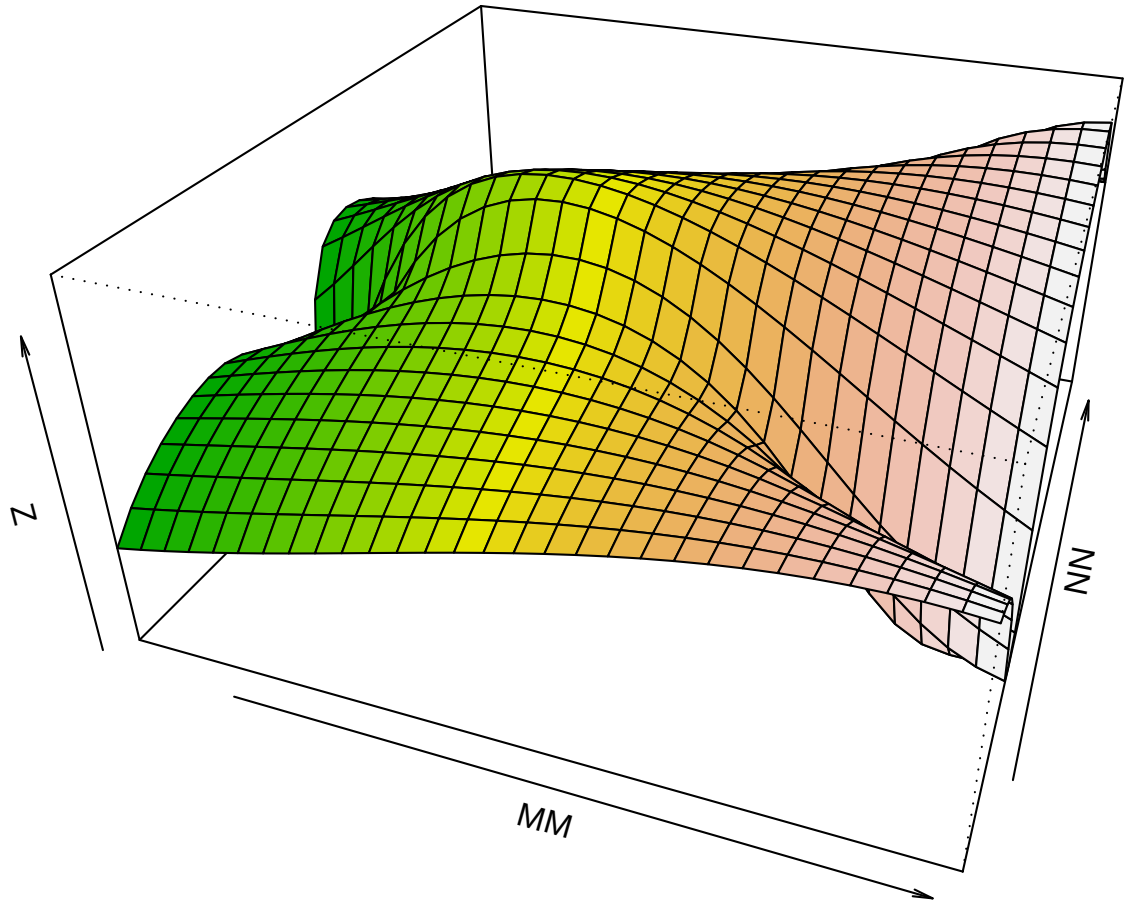

V 10
0.025
0.035
0.045
0.50
0.60
0.70
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta

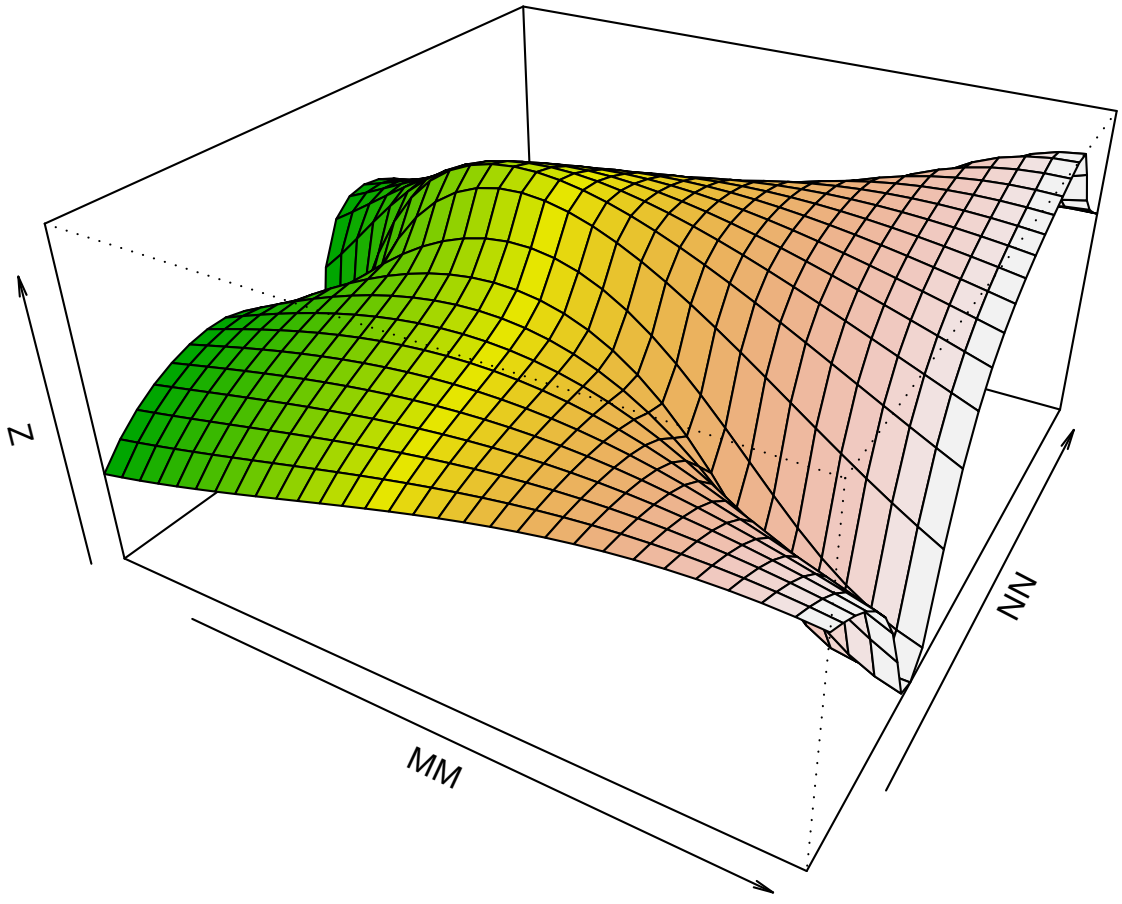

V 8
0.00
0.02
0.04
0.50
0.60
0.70
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta

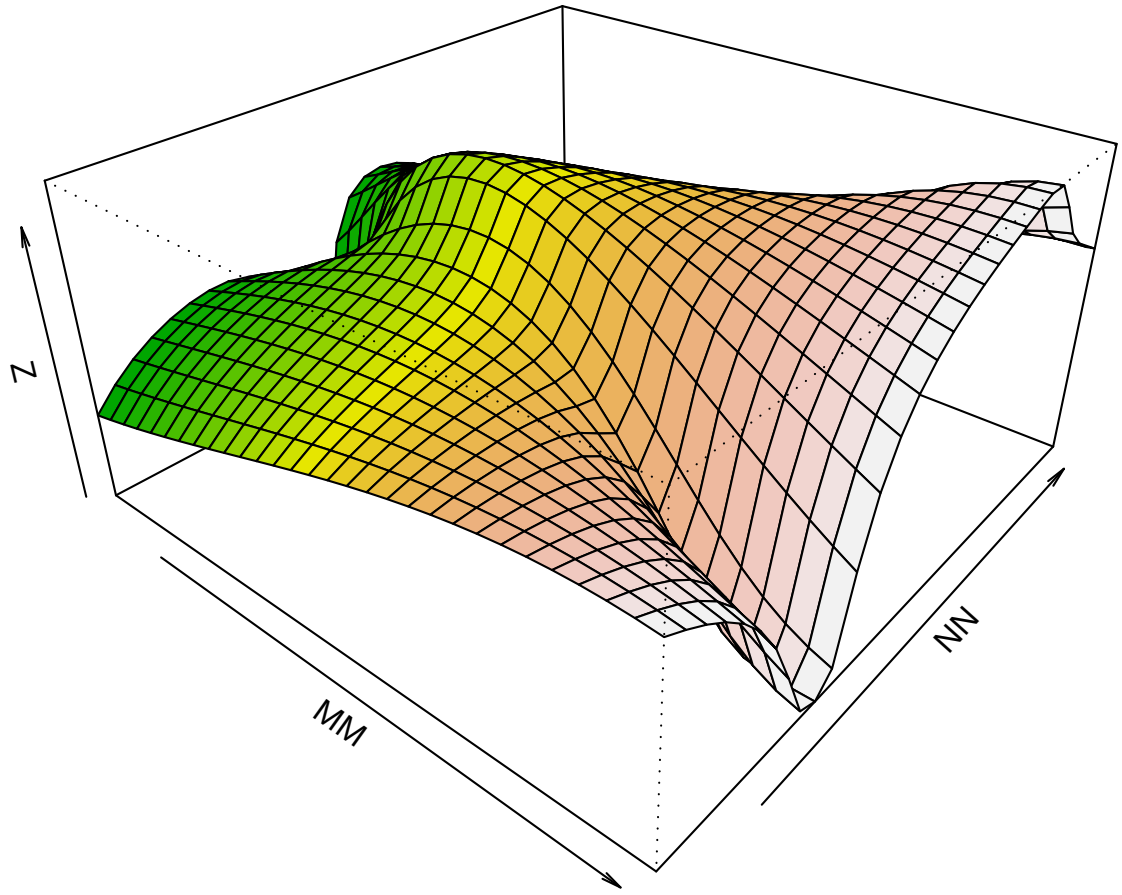

V 11
0.01
0.03
0.50
0.60
0.70
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta

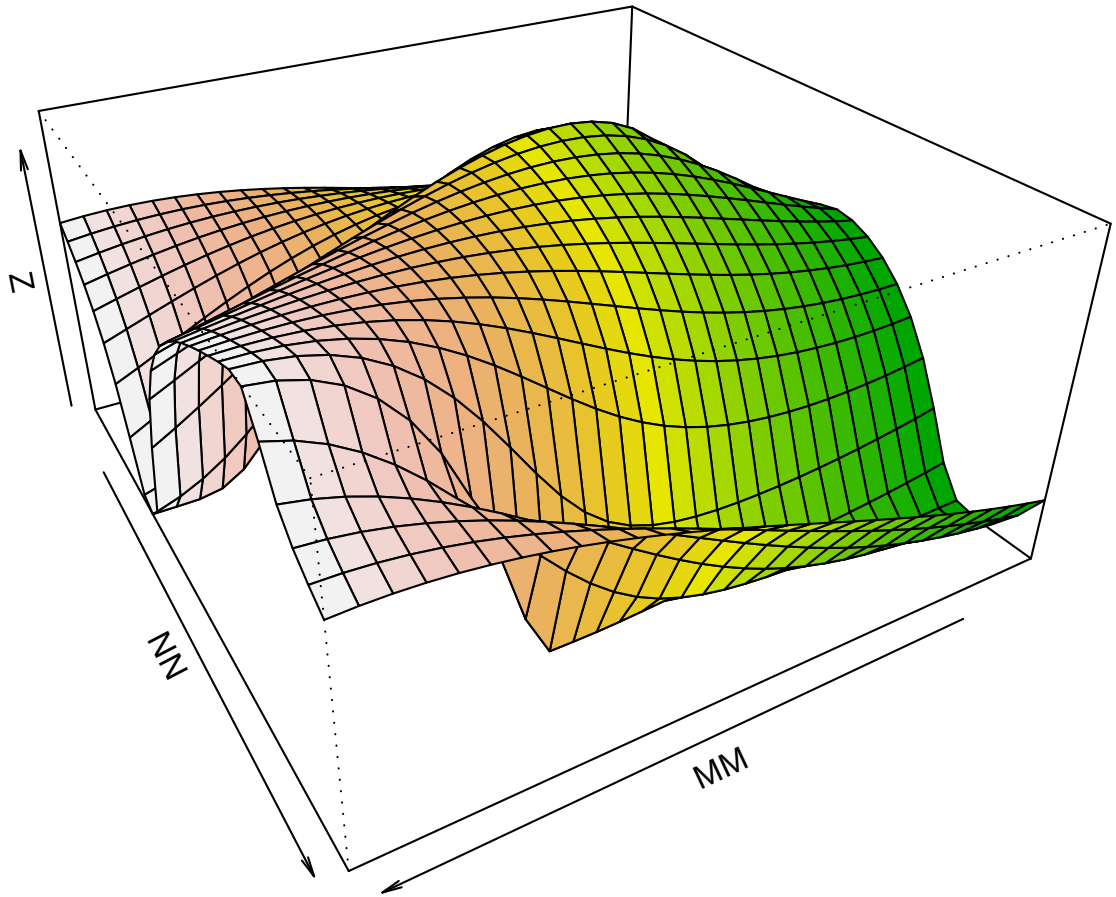

V 9
0.025
0.035
0.5
0.6
0.7
0.8
0.9
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta

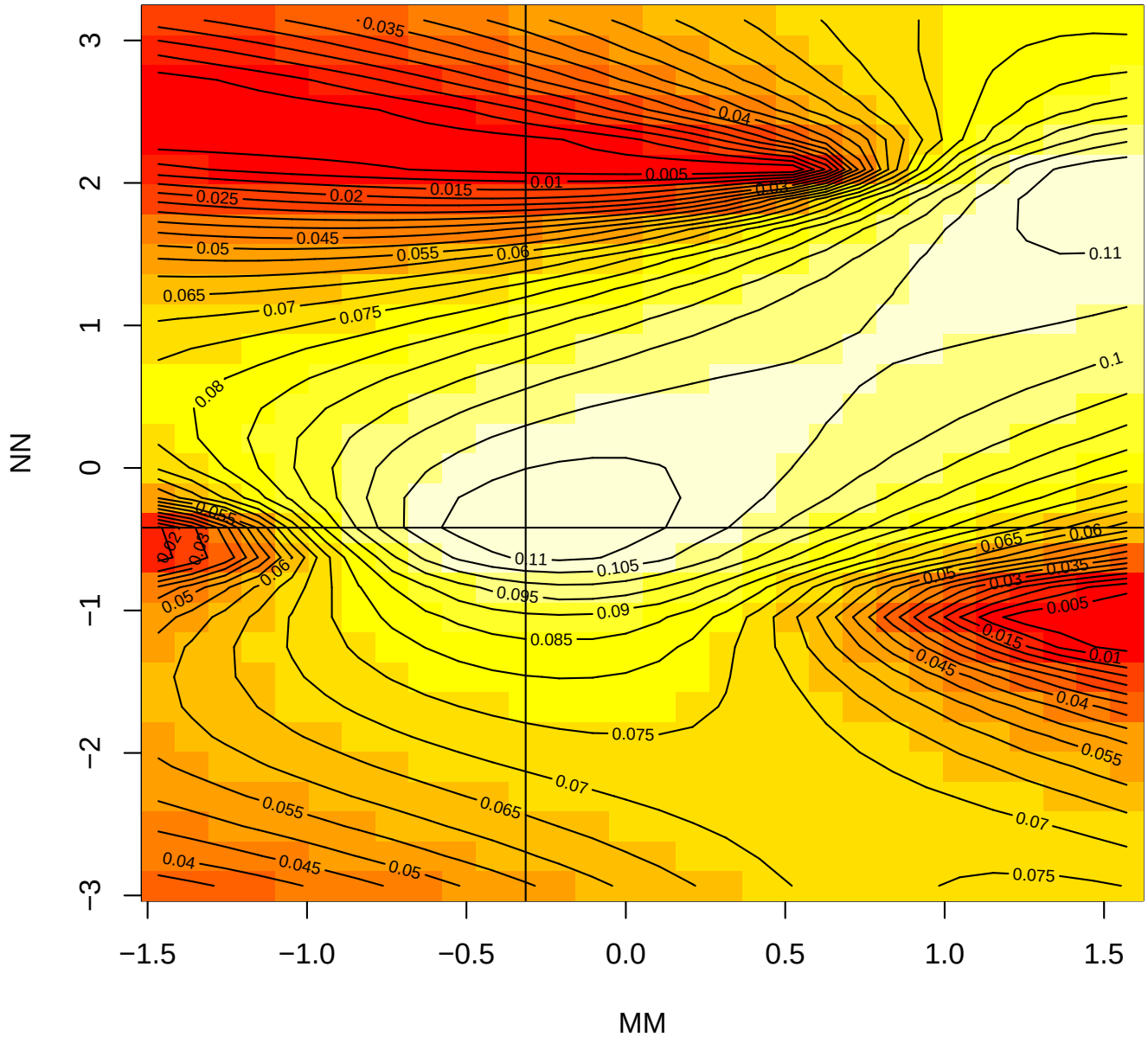

V 12
0.015
0.030
0.045
0.5
0.6
0.7
0.8
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theta


Z
MM
NN


Z
MM
NN


Z
MM
NN


Z
NN
MM


NN
MM
3
2
1
0
−1
−2
−3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
0.035
0.04
0.005
0.01
0.015
0.02
0.025
0.045
0.05
0.055
0.06
0.065
0.07
0.075
0.08
0.11
0.1
0.105
0.095
0.09
0.085
0.03
0.035
0.045
0.04
0.055
0.065
0.07
0.075

Some computational issues

The computational issues:

- The partial AUC “may not” be unique as in AUC case.
- When the length of marker is large or the covariance matrices are nearly singular then the results become unstable.

Coronary Heart Disease (Ex. in Liu, et al (2005) Statist. in Med.)

Dimension $p = 4$. $\mu_x^t = (0.1275, 0.8845, 4.0776, 6.7724)$ and $\mu_y^t = (0.1402, 0.9337, 4.1225, 6.9112)$. False Positive Rate=0.2. $a*$ is based on AUC's combination as its initial vector.

- Variance:

$$\Sigma_x = \begin{pmatrix} 0.0034 & -0.0004 & -0.0002 & -0.0051 \\ -0.0004 & 0.0285 & 0.0029 & 0.0417 \\ -0.0002 & 0.0039 & 0.0488 & 0.0268 \\ -0.0051 & 0.0417 & 0.0268 & 0.2846 \end{pmatrix}, \quad \Sigma_y = \begin{pmatrix} 0.0043 & -0.0004 & -0.0002 & -0.0051 \\ 0.0033 & 0.0415 & 0.0019 & 0.0426 \\ 0.0006 & 0.0019 & 0.0389 & 0.0010 \\ 0.0067 & 0.0426 & 0.0010 & 0.1504 \end{pmatrix}$$

- The cutoff for pAUC, $t = 0.2$.
- $\epsilon = 10^{-6}$.

Results:

1. Marginal pAUC: 0.0331,0.0392,0.0230,0.0176
2. Not robust to the initial value.

Initial	Iterations	pAUC	a_0
$V_{(1)}$	4	0.0480	(0.9475,0.3150,0.0431,0.0320)
$V_{(2)}$	3	0.0440	(0.8927,0.4395,-0.0992,-0.0029)
$V_{(4)}$	2	0.0230	(0.0000,0.0000,1.0000,0.0000)
a^*	3	0.0480	(0.9476,0.3150,0.0432,0.0321)
Su and Liu		0.0384	(1.4600,0.3400,0.4117,0.2216)
Su and Liu(scaled)			(0.9298,0.2165,0.2622,0.1411)
Liu et al.		0.0470(0.019?)	(-0.8436,-3.2269,0.2918,-0.1181)
Liu et al.(scaled)			(-0.2518,-0.9632,0.0871,-0.0352)
Liu et al.(change sign)		0.0402	

Figure: Liu, et al (2005): Coronary Heart Disease Example

Nonparametric Method (SpAUC)

- The feature selection is performed by maximizing $pAUC(u)$ rather than AUC if high specificities (1-FPR) are paid more attention to.

- By using integration by parts we know

$$\begin{aligned} pAUC(u) &= uROC(u) - \int_0^u tdROC(t) \\ &= u - \left[\int_0^u tdROC(t) + u(1 - ROC(u)) \right] \\ &= u - E\{\min[X, u]\}, \end{aligned} \tag{2}$$

where probability distribution of the random variable X is $ROC(x)$.

Method

- we can choose $S_{\bar{D}}(Y_D)$ as a representation of X since

$$Pr(S_{\bar{D}}(Y_D) \leq u) = Pr(Y_D \geq S_{\bar{D}}^{-1}(u)) = ROC(u).$$

- We can use empirical partial area of ROC curve to approximate to pAUC(u) based on sample set $\{x_1, \cdots, x_n, y_1, \cdots, y_m\}$,

$$\widehat{pAUC}(u) = \frac{1}{n} \sum_{i=1}^{n} \left[u - \min \left\{ \frac{1}{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{I}(\mathrm{y_j} > \mathrm{x_i}), \mathrm{u} \right\} \right]. \quad (3)$$

Method

- We take sigmoid function $K(t) = 1/(1+\exp(-t))$ and

$$\begin{aligned} M(t;u) &= \int_0^t 1/(1+\exp((x-u)/h_1))dx \\ &= h_1 \log \frac{1+\exp(u/h_1)}{1+\exp((u-t)/h_1)}. \end{aligned} \tag{4}$$

For sufficiently small h_1 and h_2, we have
$K((y-x)/h_2) \approx I(y>x)$ and $M(t;u) \approx \min(\mathrm{t}, \mathrm{u})$.

- We obtain a smoothed estimator of $pAUC(u)$,

$$\widehat{pAUC}_s(u) = \frac{1}{n}\sum_{i=1}^{n}\left[u - h_1 \log \frac{1+\exp(u/h_1)}{1+\exp((u-\frac{1}{m}\sum_{j=1}^{m} K((y_j-x_i)/h_2))/h_1)}\right]. \tag{5}$$

Method

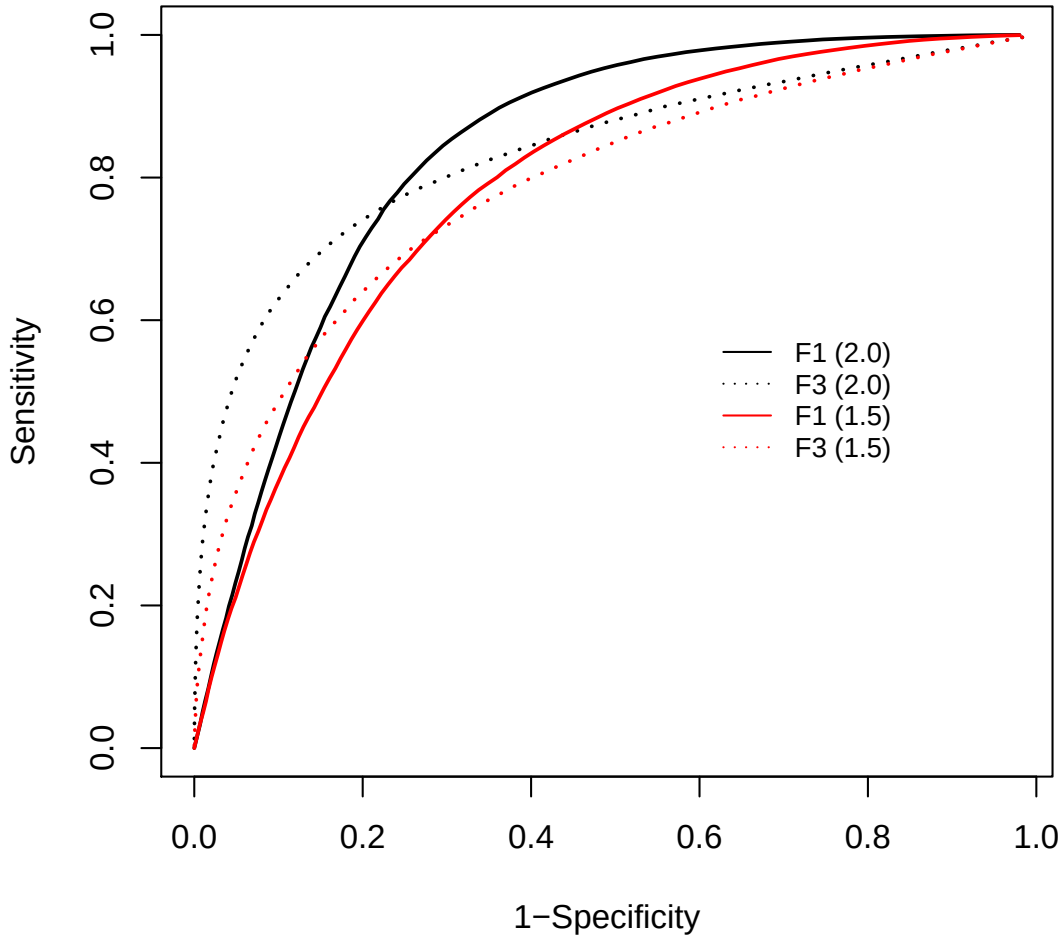
- For a vector of constants $l \in R^p$, the linear combination of features $l^{'}X$ and $l^{'}Y$ are called linear risk scores. We want to construct a classifier for the diseased and normal groups based on such linear risk scores.
- We obtain the object function

$$R(l;u) = \frac{1}{n}\sum_{i=1}^{n}\left[u - h_1 \log \frac{1+\exp(u/h_1)}{1+\exp((u-\frac{1}{m}\sum_{j=1}^{m}K(l^{'}(y_j - x_i)/h_2))/h_1)}\right]. \quad (6)$$

Simulation

- Assume that only 4 features are differently expressed between disease and control groups, denoted by $F1, F2, F3$ and $F4$. Sample sizes take $n = m = 50$.

Figure: ROC curves of $F1$ and $F3$.



Simulation results

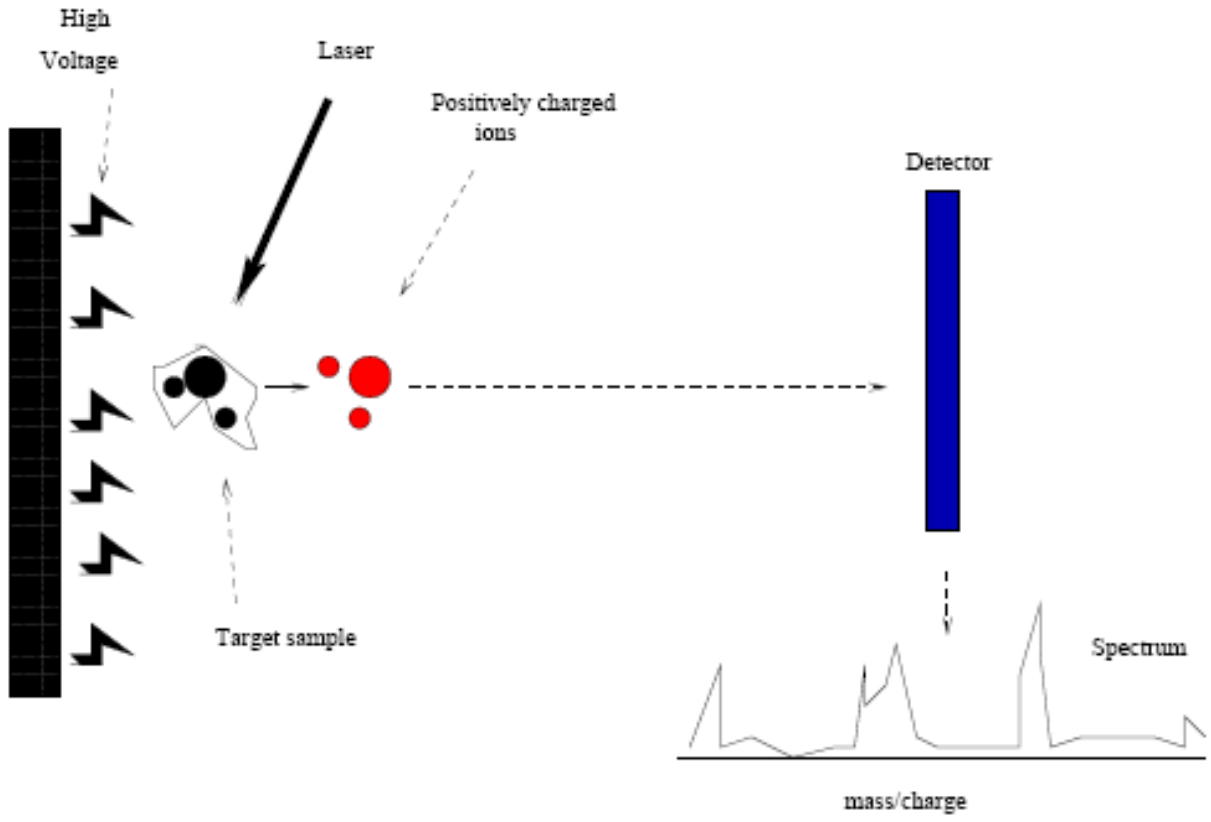
Table: Frequencies of selected features and comparisons of area under curve.

	$(\delta,\ u)$	F1	F2	F3	F4
PAUC	(1.5, 0.1)	0.367	0.372	0.810	0.783
	(1.5, 0.2)	0.441	0.458	0.683	0.698
	(2.0, 0.1)	0.357	0.361	0.867	0.866
	(2.0, 0.2)	0.460	0.502	0.679	0.651
AUC	(1.5, - -)	0.646	0.691	0.672	0.669
	(2.0, - -)	0.678	0.668	0.679	0.671

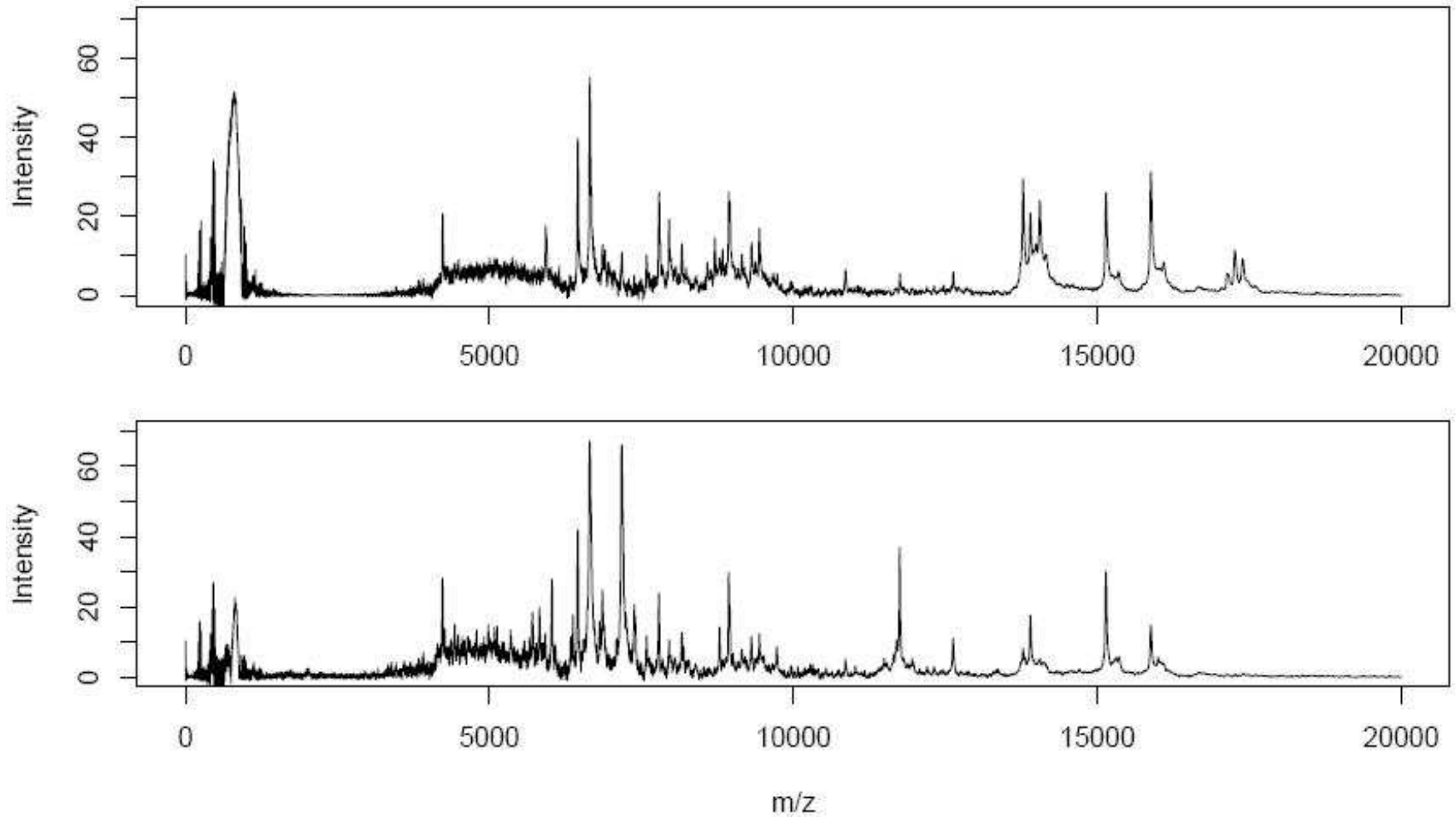
	$(\delta,\ u)$	pAUC(0.1)	pAUC(0.2)	AUC
PAUC	(1.5, 0.1)	0.066(0.014)	–	0.903(0.053)
	(1.5, 0.2)	–	0.146(0.018)	0.903(0.044)
	(2.0, 0.1)	0.076(0.011)	–	0.928(0.044)
	(2.0, 0.2)	–	0.162(0.013)	0.934(0.034)
AUC	(1.5, - -)	0.067(0.014)	0.154(0.021)	0.927(0.037)
	(2.0, - -)	0.077(0.012)	0.168(0.016)	0.953(0.027)

*Standard deviations are in parentheses.

Mass spectrometer



MS data examples



Liver cancer

Table: Selection frequencies of the top five features for liver cancer (I).

	NO vs. LDH			NO vs. LD		
Mz	$pAUC(0.1)$	$pAUC(0.2)$	AUC	$pAUC(0.1)$	$pAUC(0.2)$	AUC
4271.37			0.468			
9004.74	0.258	0.336				
9154.26				0.139	0.024	
9262.76		0.075	0.254			0.133
9441.30				0.102	0.058	0.043
11488.09	0.426	0.388	0.207			
11545.90	0.256					
11675.92			0.373			
11781.41	0.495	0.149		0.385	0.281	0.308
11866.94	0.744	0.887	0.983	0.412	0.606	0.850
13893.91				0.265	0.046	
15397.91						0.056

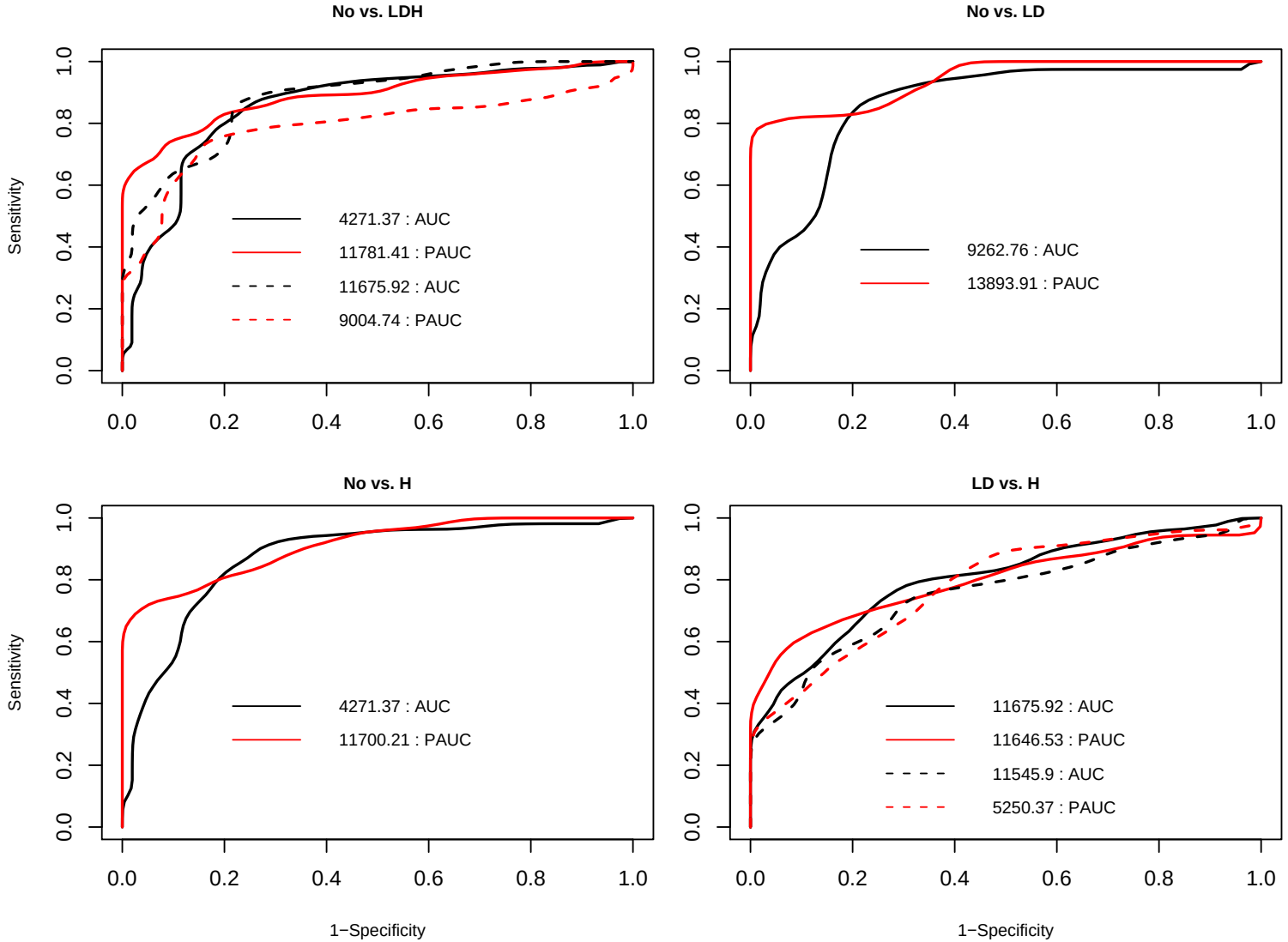
Liver cancer

Table: Selection frequencies of the top five features for liver cancer (II).

	NO vs. H			LD vs. H		
Mz	$pAUC(0.1)$	$pAUC(0.2)$	AUC	$pAUC(0.1)$	$pAUC(0.2)$	AUC
4271.37			0.307			
5209.75						0.379
5250.37				0.336		
10862.00				0.617	0.550	0.752
11488.09	0.415	0.564	0.378	0.263		
11519.47	0.148	0.185	0.360			
11545.90	0.320	0.158	0.261		0.231	0.427
11646.53				0.798	0.753	
11675.92	0.334	0.153	0.479		0.188	0.428
11700.21	0.317	0.088				
11781.41				0.280	0.228	
14097.74						0.209

Liver cancer

Figure: ROC curves comparison of selected features for liver cancer.



Liver cancer

Table: Comparisons of the area under curve for liver cancer by using pAUC (u=0.1 and 0.2) and AUC (u=1.0).

		training			testing		
	u	pAUC(0.1)	pAUC(0.2)	AUC	pAUC(0.1)	pAUC(0.2)	AUC
No vs.	0.1	0.088(0.009)*	-	0.961(0.036)	0.071(0.015)	-	0.918(0.054)
LDH	0.2	-	0.175(0.011)	0.953(0.021)	-	0.154(0.022)	0.914(0.046)
	1.0	0.087(0.013)	0.182(0.020)	0.970(0.049)	0.071(0.016)	0.159(0.023)	0.926(0.054)
No vs.	0.1	0.088(0.005)	-	0.971(0.016)	0.072(0.016)	-	0.934(0.049)
LD	0.2	-	0.180(0.006)	0.969(0.015)	-	0.167(0.020)	0.947(0.045)
	1.0	0.090(0.006)	0.187(0.009)	0.983(0.012)	0.078(0.014)	0.170(0.020)	0.949(0.040)
No vs.	0.1	0.085(0.007)	-	0.961(0.022)	0.065(0.013)	-	0.910(0.044)
H	0.2	-	0.171(0.009)	0.951(0.017)	-	0.150(0.019)	0.912(0.042)
	1.0	0.088(0.009)	0.183(0.013)	0.976(0.021)	0.070(0.013)	0.156(0.019)	0.921(0.041)
LD vs.	0.1	0.077(0.015)	-	0.922(0.064)	0.046(0.020)	-	0.826(0.083)
H	0.2	-	0.159(0.021)	0.914(0.055)	-	0.114(0.031)	0.814(0.082)
	1.0	0.080(0.016)	0.172(0.024)	0.955(0.050)	0.043(0.018)	0.115(0.027)	0.829(0.072)

* Standard deviations are in parentheses.

Liver cancer

Table: Classification results for liver cancer by using pAUC (u=0.1 and 0.2) and AUC (u=1.0).

		number	training		testing	
	u	of selection	sensitivity	specificity	sensitivity	specificity
No vs.	0.1	4.080(2.127)*	0.922(0.067)	0.914(0.000)	0.874(0.088)	0.807(0.107)
LDH	0.2	2.402(1.129)	0.932(0.043)	0.800(0.004)	0.891(0.067)	0.728(0.126)
	1.0	4.705(2.799)	0.908(0.064)	0.947(0.057)	0.858(0.091)	0.843(0.105)
No vs.	0.1	1.441(0.863)	0.936(0.043)	0.916(0.009)	0.865(0.115)	0.858(0.093)
LD	0.2	1.028(0.193)	0.968(0.027)	0.814(0.037)	0.943(0.085)	0.786(0.110)
	1.0	1.703(0.986)	0.970(0.031)	0.933(0.040)	0.908(0.111)	0.891(0.075)
No vs.	0.1	2.629(1.711)	0.910(0.060)	0.915(0.004)	0.830(0.092)	0.828(0.108)
H	0.2	1.505(0.888)	0.941(0.036)	0.801(0.011)	0.892(0.079)	0.746(0.120)
	1.0	3.362(2.145)	0.928(0.049)	0.932(0.054)	0.838(0.094)	0.846(0.101)
LD vs.	0.1	6.251(4.254)	0.844(0.120)	0.920(0.004)	0.723(0.143)	0.768(0.142)
H	0.2	4.571(3.264)	0.876(0.095)	0.800(0.004)	0.767(0.121)	0.676(0.149)
	1.0	6.533(3.757)	0.895(0.077)	0.914(0.078)	0.762(0.118)	0.701(0.142)

*Standard deviations are in parentheses.

Thanks!