Combining High-dimensional Biomarkers with Limit of Detection

Jinjuan Wang[1], Yunpeng Zhao[2,*], Larry L Tang[2], Qizhai Li[1]

[1]LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

[2]

*Corresponding author,E-mail: yzhao15@gmu.edu

1 Introduction

The limit of detection (LOD) is often encountered when measuring proteomic markers. Due to the limited detecting ability of an equipment or instrument, it is difficult to measure markers at a relatively low level. If the marker value of a subject is beyond the range of the detection, the subject is assigned the measurement of NA (**REF**). The ultrasensitive immunosensors and arrays based on nanotechnology have been developed to improve the LOD. However, in point-of-care applications, the LOD is still a challenging issue in evaluating diagnostic accuracy with available statistical methods (**REF**). One may assume that non-NA observations follow multivariate normal distributions, and empirically estimate mean vector and variance covariance matrices for two populations using only non-NA observations. Another method is to replace NA with some constant, and then estimate parameters using all the data (**REF**). But both methods may lead to biased parameter estimators (**REF**). A maximum likelihood estimation method (MLE) can be used to estimate parameters in the distributions of two markers or more markers (**REF**). However, these existing methods can not be applied to the high dimensional proteomic biomarkers since the likelihood functions for estimating mean and covariance matrices can no longer be optimized without additional penalty terms or simplified assumptions on the covariance matrices.

High dimensional biomarker data are often encountered in contemporary studies when the dimension of the biomarkers greatly exceeds the sample sizes. Dimension reduction techniques have been proposed to deal with biomarkers with LOD (see for example, (**REF**).

However, the inference post dimension reduction is not trivial.

When assessing the discriminating diagnostic ablitity of a biomarker, receiver operating characteristic(ROC) curve is an important tool. The ROC curve of a biomarker can be obtained via plotting its sensitivity(true positive rate) versus 1 minus its specificity(true negative rate). The area and partial area under the ROC curve(AUC) are widely used for measuring the performance of a biomarker, with larger the area meaning better peformance. In some diagnostic studies, multiple biomarkers are measured. In this case, combining these biomarkers can improve diagnostic accuracy.

In this work, we propose a resample-replace procedure to handle this issue. We first estimate the mean and variance parameters for a biomarker with lowest missing rate and draw the sample from the estimated normal distribution. We then use conduct the linear regression on one biomarker with the second lowest missing value on the biomarker with full data and use the fitted values to replace its missing values. After that we optimize the penalized likelihood function for cases and controls using a graphical LASSO. Our method leads to valid ROC curve estimation without explicit parametric distribution assumption even for high dimensional proteomic markers.

2 Method

2.1 Notations

Suppose that there are $m+n$ subjects enrolled and p biomarkers are measured on each subject in a study, where m subjects are randomly sampled from case population and n subjects are randomly sampled from control population. Assume that the detection limit value of the jth biomarker is d_j ($d_j > 0$), $j = 1, \cdots, p$, that means, when the value of the jth biomarker on a subject is larger than d_j, it can be observed, otherwise it is missing, in which NA is used to denote it. Denote the value of the jth biomarker on the ith subject by x_{ij} in case population, $i = 1, \cdots, m$, and that on the kth individual by y_{kj} in control population, $k = 1, \cdots, n$, $j = 1, \cdots, p$. Let $\mathbf{X} = (x_{ij})_{m \times p}$, and m_j be the number of missing values among m case subjects for the jth biomarker, and $\mathbf{Y} = (y_{kj})_{n\times p}$, and n_j be the number of missing values among n control subjects for the jth biomarker, $j = 1, \cdots, p$. Without loss of generality, we assume that $0 \leq m_1 \leq m_2 \leq \cdots \leq m_p < m$, x_{gj} is not NA for $g = 1, \cdots, m - m_j$, $0 \leq n_1 \leq n_2 \leq \cdots \leq n_p < n$, and y_{gj} is not NA for $g = 1, \cdots, n - n_j$, $j = 1, \cdots, p$.

Suppose that there exist a and b such that $m_a > 0$ and $m_{a-1} = 0$, and $n_b > 0$ and $n_{b-1} = 0$, where $m_0 = n_0 = 0$. Let $\mu_x = (\mu_{x1}, \cdots, \mu_{xp})^\tau$ and $\mu_y = (\mu_{y1}, \cdots, \mu_{yp})^\tau$ be the population means of case and control populations, respectively, and U_x and U_y be the corresponding variance and covariance matrices, where τ denote the transpose of a vector or a matrix. Denoet a $L-$dimensional column vector with the unit being 1 by $\mathbf{1}_L$, where L in an positive integer.

2.2 Estimation Procedure

From Su and Liu (**REF**) and Liu et.al. (**REF**), the optimal linear combination coefficient maximizing AUC of the combined marker is given by

$$\eta = (U_x + U_y)^{-1}(\mu_x - \mu_y)$$

The corresponding AUC is

$$\text{AUC} = \Phi\left(\sqrt{(\mu_x - \mu_y)^\tau (U_x + U_y)^{-1}(\mu_x - \mu_y)}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution. Denote $\zeta(\eta) = \frac{\eta^\tau(\mu_x - \mu_y)}{(\eta^\tau U_x \eta)^{1/2}}$, and $\xi(\eta) = \frac{(\eta^\tau U_y \eta)^{1/2}}{(\eta^\tau U_x \eta)^{1/2}}$ The pAUC for α is

$$\text{pAUC}(\alpha) = \int_0^\alpha \Phi\left(\zeta(\eta) - \xi(\eta)\Phi^{-1}(1-p)\right) \mathrm{d}p$$

We now estimate μ_x, μ_y, U_x, and U_y. We propose to replace NAs using observations randomly sampled from an inferred distribution for the biomarker with the lowest missing rate based on its observed data. Then we use the fitted values by linear regression to replace the NAs of the other biomarkers based on the imputed ones. This is motivated the fact that the conditional mean of a variable give other variables is a linear function of means of other variables when they are jointly follow a multiple-variate normal distribution.

The likelihood function for the ath biomarker in case is

$$L_{xa}(\mu_{xa}, \sigma_{xa}^2) = \prod_{i=1}^{m-m_a} \frac{\phi\big(\frac{x_{ia}-\mu_{xa}}{\sigma_{xa}}\big)}{1-\Phi\big(\frac{d_a-\mu_{xa}}{\sigma_{xa}}\big)},$$

and the log-likelihood function is

$$l_{xa}(\mu_{xa}, \sigma_{xa}^2) = -\sum_{i=1}^{m-m_a} \frac{(x_{ia}-\mu_{xa})^2}{2\sigma_{xa}^2} - (m-m_a)\ln\left[1-\Phi\Big(\frac{d_a-\mu_{xa}}{\sigma_{xa}}\Big)\right] - (m-m_a)\ln\sigma_{xa} - \frac{m-m_a}{2}\ln(2\pi),$$

where $\phi(\cdot)$ denotes the probability density function of a standard normal distribution. Let

$$(\hat{\mu}_{xa}, \hat{\sigma}_{xa}^2) = \arg\max_{\mu_{xa},\sigma_{xa}^2} l_{xa}(\mu_{xa}, \sigma_{xa}^2).$$

The log-likelihood function for the bth biomarker in controls is

$$l_{yb}(\mu_{yb}, \sigma_{yb}^2) = -\sum_{k=1}^{n-n_b} \frac{(y_{kb}-\mu_{yb})^2}{2\sigma_{yb}^2} - (n-n_b)\ln\left[1-\Phi\Big(\frac{d_b-\mu_{yb}}{\sigma_{yb}}\Big)\right] - (n-n_b)\ln\sigma_{yb} - \frac{n-n_b}{2}\ln(2\pi),$$

and let

$$(\hat{\mu}_{yb}, \hat{\sigma}_{yb}^2) = \arg\max_{\mu_{yb},\sigma_{yb}^2} l_{yb}(\mu_{yb}, \sigma_{yb}^2).$$

The proposed estimation procedure is divided into four steps as follows:

1. *Step 1.*

   (1) Generate m_a observations from the truncated normal distribution with density function $f_a(x_a) = \phi\big(\frac{x_a-\hat{\mu}_{xa}}{\hat{\sigma}_{xa}}\big)/\Phi\big(\frac{d_a-\hat{\mu}_{xa}}{\hat{\sigma}_{xa}}\big), x_a \in (-\infty, d_a)$, denote them by $\tilde{x}_{1a}, \cdots, \tilde{x}_{m_a a}$. Using these values to replace NAs in $\mathbf{X}$ for the ath biomarkers in cases.

   (2) For j from $a+1$ to p,

      i. conduct the linear regression using $\{x_{ij}, i = 1, 2, \cdots, m-m_j\}$ and $\{(x_{i1}, \cdots, x_{i(j-1)}), i = 1, 2, \cdots, m-m_j\}$, that means,

$$x_{ij} = \beta_{j0} + \beta_{j1}x_{i1} + \cdots + \beta_{j(j-1)}x_{i(j-1)} + \epsilon_j,$$

where β_{j0} is the intercept, $\beta_{jl}, l = 1, \cdots, j-1$ are coefficients and ϵ_j is the error term following the normal distribution with mean 0 and unknown variance $\sigma^2_{\epsilon xj}$. Denote

$$\beta_j = \left(\beta_{j0}, \cdots, \beta_{j(j-1)}\right)^\tau, \quad \mathbf{x}_g = (x_{1g}, \cdots, x_{(m-m_g)g})^\tau, g = 1, \cdots, j$$

and $\mathbf{X}_j = \left(\mathbf{1}_{m-m_j}, \mathbf{x}_1, \cdots, \mathbf{x}_{j-1}\right)$.

The least square estimate(also the maximum likelihood estimate) of β_j and is

$$\hat{\beta}_j = \left(\hat{\beta}_{j0}, \cdots, \hat{\beta}_{j(j-1)}\right)^\tau = \left(\mathbf{X}_j^\tau \mathbf{X}_j\right)^{-1} \mathbf{X}_j^\tau \mathbf{x}_j.$$

and the maximum likelihood estimate of $\sigma^2_{\epsilon xj}$ is

$$\hat{\sigma}^2_{\epsilon xj} = \frac{||\mathbf{x}_j - \mathbf{X}_j \hat{\beta}_j||^2}{m - m_j}$$

ii. calculate

$$\hat{x}_{ij} = \hat{\beta}_{j0} + \hat{\beta}_{j1} x_{i1} + \cdots + \beta_{j(j-1)} x_{i(j-1)} + \epsilon_{ij} \text{ for } i = m - m_j + 1, \cdots, m,$$

where ϵ_{ij} is a random variate that follows the normal distribution $\mathbf{N}(0, \hat{\sigma}_{\epsilon xj})$ and less than d_j and replace NAs of the jth biomarker by $\{\hat{x}_{ij}, i = m - m_j + 1, \cdots, m\}$.

(3) Update $\mathbf{X}$ to $\tilde{\mathbf{X}}$ using the imputed data. Denote $\tilde{\mathbf{X}} = (\tilde{x}_{ij})_{m\times p} \widehat{=} (\tilde{\mathbf{x}}_1^\tau, \cdots, \tilde{\mathbf{x}}_m^\tau)^\tau$. The log-likelihood function is

$$l(\mu_x, U_x) = -\sum_{i=1}^{m} \frac{(\tilde{\mathbf{x}}_i - \mu_x)^\tau U_x^{-1} (\tilde{\mathbf{x}}_i - \mu_x)}{2|U_x|} - \frac{mp}{2} \ln(2\pi) - m \ln |U_x|.$$

Let $\Theta_x = U_x^{-1} = (\theta_{xij})_{p\times p}$. Using the graphical LASSO, the penalized log-likelihood is

$$l(\mu_x, \Theta_x) \propto \ln \det(\Theta_x) - \frac{1}{m} \sum_{i=1}^{m} (\tilde{\mathbf{x}}_i - \mu_x)^\tau \Theta_x (\tilde{\mathbf{x}}_i - \mu_x) - \lambda \sum_{i \neq j} \theta_{xij},$$

where $\lambda > 0$ is the tuning parameter. From

$$\frac{\partial l(\mu_x, \Theta_x)}{\partial \mu_x} = 0$$

we have $\hat{\mu}_x = \left(\frac{1}{m} \sum\limits_{i=1}^{m} \tilde{x}_{i1}, \cdots, \frac{1}{m} \sum\limits_{i=1}^{m} \tilde{x}_{ip} \right)^{\tau}$. By plugging in $\hat{\mu}_x$, maximizing $l(\mu_x, \Theta_x)$ over Θ_x is equivalent to maximizing

$$\log \det(\Theta_x) - \mathrm{tr}(V_x \Theta_x) - \lambda \sum_{i \neq j \in \{1, \cdots, p\}} \theta_{ij}$$

where $V_x = \frac{1}{m} \sum_{i=1}^{m} (\tilde{\mathbf{x}}_i - \hat{\mu}_x)(\tilde{\mathbf{x}}_i - \hat{\mu}_x)^{\tau}$. Using the graphical lasso algorithm (**REF**)(which has been implemented in the function *glasso* of the R package *glasso*), we can get the estimate of U_x, denote it by $\hat{U}_x$.

2. *Step 2.* Similarly perform step 1 for $\mathbf{Y}$ in controls, and denote the estimation of μ_y and U_y by $\hat{\mu}_y$ and $\hat{U}_y$, respectively.

3. *Step 3.* Plugging in the estimators $\hat{\mu}_x, \hat{\mu}_y, \hat{U}_x, \hat{U}_y$, we can use the expression in the notation section to estimate AUC, and pAUC(α) for some specific α ,denoting them as $\widehat{\mathrm{AUC}}$, and $\mathrm{p}\widehat{\mathrm{AUC}(\alpha})$.

4. *Step 4.* Perform Steps 1-3 r times (f.g. $r = 50$), we obtain r estimators of each parameter we want to estimate, which are denoted as $\widehat{\mathrm{pAUC}}(\alpha)_t, t = 1, \cdots, r$. So we get

$$\widehat{\mathrm{pAUC}}(\alpha) = \sum_{t=1}^{r} \widehat{\mathrm{pAUC}}(\alpha)_t$$

2.3 Mathematical Theory

To verify the rationalization of our new estimation procedure, we need to prove that the dataset we get after the filling section can be seen as an sample from the original unknown population. We do this by by proving that each filled element follows the normal distribution $\mathbf{N}(\mu_j, \sigma_j^2)$, where μ_j and σ_j^2 are the corresponding unknown but true mean and variation.

Taking $\mathbf{X} = (x_{ij})_{m \times p}$ for example. Suppose $i > m_j$, then x_{ij} is a missing data and $x_{i,1}, \cdots, x_{i,(j-1)}$ $\mathbf{x}_j = (x_{1j}, \cdots, x_{(m-m_j)j})^{\tau}$ and $\mathbf{X}_j = \big(\mathbf{1}_{m-m_j}, \mathbf{x}_1, \cdots, \mathbf{x}_{j-1}\big)$ are known. Instead of using only $\mathbf{x}_j$ to estimate the distribution of x_{ij} and get a possible value $\hat{x}_{ij}$ to replace x_{ij}, we also want to use the relationship information of x_{ij} between $x_{i,1}, \cdots, x_{i,(j-1)}$. So we use theory of conditional distribution of multivariate normal distribution.

Denote $j \times 1$ vector $(x_1, \cdots, x_{j-1}, x_j)$ follows normal distribution $\mathbf{N}(\mu_{1 \cdot j}, \Sigma_{1 \cdot j})$, where

$$\mu_{1 \cdot j} = (\mu_{x1}, \cdots, \mu_{xj})$$

and

$$\Sigma_{1 \cdot j} = \begin{pmatrix} \Sigma_{1 \cdot (j-1)} & \sigma^2_{1 \cdot (j-1),j} \\ \sigma^2_{1 \cdot (j-1),j} & \sigma^2_{jj} \end{pmatrix}$$

Given $x_{i,1 \cdot (j-1)} = (x_{i1}, \cdots, x_{i(j-1)})$, the conditional distribution of x_{ij} is $\mathbf{N}(\mu_{xj|1 \cdot (j-1)}, \sigma^2_{j|1 \cdot (j-1)})$, with

$$\mu_{xj|1 \cdot (j-1)} = \mu_{xj} - \sigma^2_{1 \cdot (j-1),j} \Sigma^{-1}_{1 \cdot (j-1)} (x_{i,1 \cdot (j-1)} - \mu_{1 \cdot (j-1)}) = \beta_{j0} + \beta_j^{\tau} x_{i,1 \cdot (j-1)}$$

$$\sigma^2_{j|1 \cdot (j-1)} = \sigma^2_{jj} - \sigma^2_{1 \cdot (j-1),j} \Sigma^{-1}_{1 \cdot (j-1)} \sigma^2_{1 \cdot (j-1),j}$$

That is,

$$x_{ij} = \beta_{j0} + \beta_j^{\tau} x_{i,1 \cdot (j-1)} + \epsilon_{ij}$$

where ϵ_{ij}'s mean is 0 and variance is $\sigma^2_{j|1 \cdot (j-1)}$. So we can maximize the likelihood function to estimate these unknown parameters. More specifically, consider the likelihood funtion

$$L(\mu_{xj|1 \cdot (j-1)}, \sigma^2_{j|1 \cdot (j-1)}) = \prod_{i=1}^{m-m_j} \frac{1}{\sqrt{2\pi}\sigma_{j|1 \cdot (j-1)}} exp(\frac{x_{ij} - \beta_{j0} - \hat{\beta}_j^{\tau} x_{i,1 \cdot j-1}}{2\sigma^2_{j|1 \cdot (j-1)}})$$

Maximize it, we can get the maximum likelihood estimate(MLE) of these parameters Denote the MLE of β_{j0},β_j and $\sigma^2_{j|1 \cdot (j-1)}$ are $\hat{\beta_{j0}}$, $\hat{\beta}_j$ and $\hat{\sigma}^2_{j|1 \cdot (j-1)}$, then the filled element is

$$\hat{x}_{ij} = \hat{\beta}_{j0} + \hat{\beta}_j^{\tau} x_{i,1 \cdot (j-1)} + \epsilon_{ij}$$

where ϵ_{ij} is a random variate that comes from $\mathbf{N}(0, \hat{\sigma}^2_{j|1 \cdot (j-1)})$. Now we prove that $\hat{x}_{ij}$'s mean is μ_{xj} and variance is σ^2_{jj}. Using the consistency property of maximum likelihood estimate, we have that

$$\begin{aligned} \mathbf{E}(\hat{x}_{ij}) &= \mathbf{E}(\mu_{xj} - \sigma^2_{1 \cdot (j-1),j} \Sigma^{-1}_{1 \cdot (j-1)} (x_{i,1 \cdot (j-1)} - \mu_{1 \cdot (j-1)})) \\ &= \mu_{xj} - \sigma^2_{1 \cdot (j-1),j} \Sigma^{-1}_{1 \cdot (j-1)} (\mathbf{E}(x_{i,1 \cdot (j-1)}) - \mu_{1 \cdot (j-1)}) \\ &= \mu_{xj} \end{aligned}$$

$$\begin{aligned}\mathbf{var}(\hat{x}_{ij}) &= \mathbf{var}(\hat{\beta}_j^\tau x_{i,1\cdot(j-1)} + \epsilon_{ij}) \\ &= \mathbf{var}(\hat{\beta}_j^\tau x_{1\cdot(j-1)}) + \mathbf{var}(\epsilon_{ij}) \\ &\to \sigma^2_{1\cdot(j-1),j}\Sigma^{-1}_{1\cdot(j-1)}\Sigma_{1\cdot(j-1)}\Sigma^{-1}_{1\cdot(j-1)}\sigma^2_{1\cdot(j-1),j} + (\sigma^2_{jj} - \sigma^2_{1\cdot(j-1),j}\Sigma^{-1}_{1\cdot(j-1)}\sigma^2_{1\cdot(j-1),j}) \\ &= \sigma^2_{jj}\end{aligned}$$

This means that the dataset we get after filling all the missing values can be seen as a random sample subtracted from the original population. Using this complete dataset, we obtain the estimate of the variance-covariance matrix through graphical glasso.

3 Numerical Studies

3.1 Simulation studies

In this section, we conduct simulation studies to demonstrate the performance of the proposed resample-input-graphical LASSO method (denote it by RIG) by comparing it with two methods: substituting NA with LOD value (denote it by SNL) and ignoring the NA values (denote it by IGN). We consider $p = 120$ and (m, n) is chosen from $\{(50, 50), (75, 75), (100, 100)\}$. We set $\mu_x = (11, 11, \cdots, 11)^\tau$ and $\mu_y = (10, 10, \cdots, 10)^\tau$. $U_x = U_y$ are block diagonal matrixes composing of U_0 with U_0 being a 30×30 matrix with diagonal elements 1 and off-diagonal elements ρ, which is chosen from $\{0.2, 0.5, 0.8\}$. The LOD vaules are set to be 7.5, 8.5 or 9.5. We set α changing in $\{0.05, 0.1, 0.2, 1\}$. For each setting, we repeat 1000 replicates to calculate the mean-squared error (MSE) of all the pAUCs.

The ratios of the MSEs between SNL and RIG are shown in Tables 1 and 2. It can be seen that in all settings, the proposed RIG performs better than SNL, since all the ratios are bigger than 1. From both tables, we can see

1) When the sample size is fixed, the ratios are increasing with ρ increasing. It is reasonable since the RIG is constructed based on the linear regression and the conditional mean of a biomarker is the linear function of regressed biomarkers. The imputed data is more close to its ”true” value when ρ is larger. For example, when $(m, n) = (100, 100)$ and the LOD

is 7.5, the ratio for pAUC(0.05) under $\rho = 0.2$ is 1.013, while it increases to be 1.413 under $\rho = 0.8$.
2) When ρ is fixed, the ratio becomes bigger as the sample size gets larger. Take $\rho = 0.5$ and the LOD=8.5 as an example, when (m, n) are (50,50),(75,75) and (100,100), the ratios for pAUC(0.05) are 1.003, 1.047 and 1.124, respectively.

The ratios of the MSEs of IGN over RIG are also given in Tables 1 and 2. It can be seen that under most of the considered scenarios, the proposed RIG performs better than IGN. In both tables, there are some NAs, which are made for IGN when more than 90% of the 1000 estimates for a single pAUC are NAs. From both tables, we find that
1) When the sample size is given, the bigger the ρ is, the better the proposed RIG is. For example, when $(m, n) = (75, 75)$ and the LOD is 8.5, the ratios for pAUC(0.1) changes from 0.987 to 1.189 as ρ changes from 0.2 to 0.8.
2) When the LOD value is bigger, the IGN dose not always works. The ratios can be less than 1 as $(m, n) = (50, 50)$. This can be explained by the fact that the estimations of mean and variance might be biased for RIG as the sample size is small. The ratios gets bigger as the sample size gets bigger. When ρ is set to be 0.2 and the LOD is 8.5, the ratio for pAUC(0.05) is NA, 0.993 and 1.01 respectively for the sample size being (50,50),(75,75) and (100,100). We would like to point that it is reasonable to compare IGN and RIG under large sample size since there are small percentages of NAs in the 1000 replicates. And under this situation, the ratios are bigger than 1, which means that the proposed RIG has better performance than IGN.

3.2 Application

To illustrate the application of our proposed method, we use a survival status dataset with 150 biomarkers. The survival status has two values, short and long. There are 15 observations belonging to the short section and 11 to the long section. We choose the biomarkers that have at least one missing value and do not have too many missing value, that is in each section, the number of missing value can not be more than 30 percent of the total observation

Table 1: The ratios of MSE for pAUC($\alpha = 0.05$) and pAUC($\alpha = 0.1$).

(m, n)	ρ	LOD	pAUC($\alpha = 0.05$)		pAUC($\alpha = 0.1$)	
			SNL/RIG	IGN/RIG	SNL/RIG	IGN/RIG
(50,50)	0.2	7.5	1.000	0.996	1	0.996
		8.5	1.000	NA	1	NA
		9.5	1.000	MA	1	NA
	0.5	7.5	1.003	0.994	1.002	0.995
		8.5	1.003	0.969	1.002	0.957
		9.5	1.002	NA	1.002	NA
	0.8	7.5	1.069	1.062	1.053	1.047
		8.5	1.063	0.993	1.048	0.975
		9.5	1.054	0.907	1.041	0.858]
(75,75)	0.2	7.5	1.002	1.002	1.002	1.002
		8.5	1.002	0.993	1.001	0.987
		9.5	1.001	NA	1.001	NA
	0.5	7.5	1.050	1.05	1.04	1.04
		8.5	1.047	1.042	1.038	1.031
		9.5	1.042	NA	1.034	NA
	0.8	7.5	1.243	1.243	1.192	1.192
		8.5	1.244	1.239	1.194	1.189
		9.5	1.216	1.122	1.173	1.058]
(100,100)	0.2	7.5	1.013	1.01	1.012	1.013
		8.5	1.012	1.01	1.011	1.011
		9.5	1.009	NA	1.009	NA
	0.5	7.5	1.126	1.144	1.105	1.119
		8.5	1.124	1.144	1.104	1.119
		9.5	1.103	NA	1.087	NA
	0.8	7.5	1.413	1.45	1.341	1.367
		8.5	1.400	1.472	1.331	1.383
		9.5	1.337	1.356	1.283	1.278

Table 2: the ratio of MSE for pAUC($\alpha = 0.2$) and AUC

(m,n)	ρ	LOD	pAUC($\alpha = 0.2$)		AUC	
			SNL/RIG	IGN/RIG	SNL/RIG	IGN/RIG
(50,50)	0.2	7.5	1.000	0.996	1	0.995
		8.5	1.000	NA	1	NA
		9.5	1.000	NA	1	NA
	0.5	7.5	1.002	0.995	1.002	0.995
		8.5	1.002	0.937	1.002	0.819
		9.5	1.001	NA	1.001	NA
	0.8	7.5	1.041	1.037	1.033	1.03
		8.5	1.038	0.957	1.03	0.908
		9.5	1.032	0.783	1.026	0.268
(75,75)	0.2	7.5	1.001	1.001	1.001	1.001
		8.5	1.001	0.975	1.001	0.892
		9.5	1.001	NA	1.001	NA
	0.5	7.5	1.034	1.034	1.03	1.03
		8.5	1.032	1.023	1.028	1.005
		9.5	1.029	NA	1.025	NA
	0.8	7.5	1.156	1.156	1.127	1.127
		8.5	1.158	1.153	1.129	1.122
		9.5	1.142	0.988	1.117	0.649
(100,100)	0.2	7.5	1.011	1.01	1.009	1.01
		8.5	1.010	1.01	1.008	1.006
		9.5	1.008	NA	1.007	NA
	0.5	7.5	1.091	1.102	1.081	1.09
		8.5	1.090	1.102	1.081	1.09
		9.5	1.076	NA	1.069	NA
	0.8	7.5	1.286	1.305	1.238	1.252
		8.5	1.279	1.317	1.233	1.263
		9.5	1.241	1.211	1.205	1.041

Table 3: The ratios of MSE for pAUC($\alpha = 0.05$) and pAUC($\alpha = 0.1$).

(m, n)	ρ	LOD	pAUC($\alpha = 0.05$)		pAUC($\alpha = 0.1$)	
			SNL/RIG	IN/RIG	SNL/RIG	IN/RIG
(50,50)	0.2	7.5	1.000	0.993	1.000	0.992
		8.5	1.000	NA	1.000	NA
		9.5	1.000	NA	1.000	NA
	0.5	7.5	1.007	0.990	1.005	0.990
		8.5	1.007	0.943	1.005	0.923
		9.5	1.008	NA	1.006	NA
	0.8	7.5	1.141	1.125	1.107	1.095
		8.5	1.139	0.991	1.106	0.955
		9.5	1.142	0.832	1.112	0.751
(75,75)	0.2	7.5	1.003	1.003	1.003	1.003
		8.5	1.003	0.979	1.003	0.965
		9.5	1.003	NA	1.002	NA
	0.5	7.5	1.101	1.102	1.081	1.081
		8.5	1.094	1.085	1.075	1.064
		9.5	1.092	NA	1.077	NA
	0.8	7.5	1.552	1.555	1.429	1.430
		8.5	1.534	1.519	1.419	1.405
		9.5	1.502	1.308	1.414	1.189
(100,100)	0.2	7.5	1.026	1.028	1.023	1.025
		8.5	1.026	1.027	1.023	1.024
		9.5	1.020	NA	1.018	NA
	0.5	7.5	1.270	1.314	1.224	1.257
		8.5	1.258	1.301	1.216	1.248
		9.5	1.224	NA	1.196	NA
	0.8	7.5	1.976	2.084	1.784	1.857
		8.5	1.894	2.100	1.732	1.874
		9.5	1.829	1.885	1.723	1.713

Table 4: the ratio of MSE for pAUC($\alpha = 0.2$) and AUC

(m, n)	ρ	LOD	pAUC($\alpha = 0.2$)		AUC	
			SNL/RIG	IN/RIG	SNL/RIG	IN/RIG
(50,50)	0.2	7.5	1.000	0.992	1.000	0.991
		8.5	1.000	NA	1.000	NA
		9.5	1.000	NA	1.000	NA
	0.5	7.5	1.004	0.990	1.003	0.990
		8.5	1.004	0.891	1.003	0.755
		9.5	1.005	NA	1.004	NA
	0.8	7.5	1.084	1.075	1.067	1.060
		8.5	1.084	0.920	1.067	0.834
		9.5	1.092	0.648	1.076	0.668
(75,75)	0.2	7.5	1.003	1.003	1.002	1.002
		8.5	1.003	0.946	1.002	1.091
		9.5	1.002	NA	1.002	NA
	0.5	7.5	1.068	1.068	1.060	1.060
		8.5	1.063	1.047	1.056	1.017
		9.5	1.066	NA	1.060	NA
	0.8	7.5	1.344	1.345	1.277	1.278
		8.5	1.339	1.325	1.276	1.258
		9.5	1.351	1.069	1.301	0.747
(100,100)	0.2	7.5	1.021	1.023	1.018	1.019
		8.5	1.021	1.021	1.018	1.015
		9.5	1.017	NA	1.016	NA
	0.5	7.5	1.193	1.219	1.171	1.194
		8.5	1.187	1.212	1.167	1.186
		9.5	1.176	NA	1.163	NA
	0.8	7.5	1.644	1.696	1.526	1.563
		8.5	1.612	1.715	1.509	1.586
		9.5	1.644	1.570	1.581	1.252

numbers. So we finally choose 45 biomarkers. Using this dataset, we compare the estimated pAUCs of the proposed method and that of the replacing LOD method. When α changes in $\{0.05, 0.1, 0.2, 1\}$, the corresponding pAUCs of the former are 0.04995,0.09994,0.19994 and 0.99994,respectively, while that of the latter is 0.03989,0.08372,0.17503 and 0.95428. So our proposed method performs better.