A resample-replace lasso procedure for combining high-dimensional markers with limit of detection

Jinjuan Wang\textsuperscript{a}, Yunpeng Zhao\textsuperscript{b}, Larry L. Tang\textsuperscript{c,d}, Claudius Mueller\textsuperscript{e}\textsuperscript{*} and Qizhai Li\textsuperscript{f}

\textsuperscript{a}School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, People's Republic of China; \textsuperscript{b}School of Mathematical and Natural Sciences, Arizona State University, Tempe, AZ, USA; \textsuperscript{c}Department of Statistics and National Center for Forensic Science, University of Central Florida, Orlando, FL, USA; \textsuperscript{d}Department of Statistics, Rehabilitation Medicine Department, NIH Clinical Center, Bethesda, MD, USA; \textsuperscript{e}George Mason University, Fairfax, VA, USA; \textsuperscript{f}LSC Academy of Mathematics and Systems Science, Chinese Academy of Sciences and University of Chinese Academy of Sciences, Beijing, People’s Republic of China

ABSTRACT

In disease screening, a biomarker combination developed by combining multiple markers tends to have a higher sensitivity than an individual marker. Parametric methods for marker combination rely on the inverse of covariance matrices, which is often a non-trivial problem for high-dimensional data generated by modern high-throughput technologies. Additionally, another common problem in disease diagnosis is the existence of limit of detection (LOD) for an instrument – that is, when a biomarker’s value falls below the limit, it cannot be observed and is assigned an NA value. To handle these two challenges in combining high-dimensional biomarkers with the presence of LOD, we propose a resample-replace lasso procedure. We first impute the values below LOD and then use the graphical lasso method to estimate the means and precision matrices for the high-dimensional biomarkers. The simulation results show that our method outperforms alternative methods such as either substitute NA values with LOD values or remove observations that have NA values. A real case analysis on a protein profiling study of glioblastoma patients on their survival status indicates that the biomarker combination obtained through the proposed method is more accurate in distinguishing between two groups.

ARTICLE HISTORY

Received 19 July 2019
Accepted 2 September 2021

KEYWORDS

Limit of detection (LOD); graphical lasso; precision matrix; area under the receiver operating characteristic curve (AUC); high-dimensional data; imputation

2020 MATHEMATICS SUBJECT CLASSIFICATION

97K80

1. Introduction

Protein-based assays are a major component in developing cancer screening markers, and their expression levels can be rapidly measured via modern high-throughput technologies. Since different proteins may indicate changes in different aspects of a cancer, a composite marker developed by combining all relevant proteins can integrate all these changes and thus may improve the diagnostic accuracy. The proteomics literature has shown that a combination of multiple protein profiles tends to have a much higher sensitivity than
an individual protein profile [3,8,31]. And finding the optimal combination of different biomarkers has been a topic of interest. A widely used tool to measure the diagnostic performance of the biomarker combination is the receiving operating characteristic (ROC) curve and the area under this curve (AUC). A larger AUC usually indicates better accuracy in distinguishing cases from controls.

To combine protein profiles, both parametric and nonparametric methods have been considered. Parametric methods for the linear combination of relevant biomarkers mainly rely on the multivariate normal distribution assumption [9,17,19–21,25,27,32]. These methods are intuitive and are closely related to the linear discriminant analysis [27]. Nonparametric methods usually rely on numerical search for the best combination, and thus are computationally demanding when the data are generated from high-throughput techniques, which limits their range of application.

The existing parametric methods face two major challenges when dealing with high-dimensional data generated by modern high-throughput technologies. The first challenge is the presence of limit of detection (LOD) when measuring proteomic markers. That is, it is difficult to measure markers at a relatively low level due to the limited detecting ability of an equipment or instrument. If a marker value of a subject is below the corresponding detection range, it cannot be observed and is assigned with the measurement of NA (or zero) [22,34]. The ultrasensitive immunosensors and arrays based on nanotechnology have been developed to handle the LOD problem. However, in point-of-care applications, the existence of LOD is still a challenging issue when using the existing statistical methods to evaluate the diagnostic accuracy [24]. For example, the LOD issue exists in a protein profiling study of glioblastoma patients on their survival status at the Center for Applied Proteomics and Molecular Medicine at George Mason University. In this study, protein profiles were measured on the glioblastoma tissues of the patients, and many of them have observations below the LOD. To handle this issue, one may impute the NA values with some constants to get a complete dataset [13]. Another approach is to remove observations with NA values and only use those complete ones to conduct estimation. However, both methods fail to make the best use of the information contained in the data and may lead to biased parameter estimators [9,20].

The second challenge is encountered when the number of proteins is larger than the sample size. The traditional setting assumes that a fixed number of proteins are profiled on a large sample. To estimate the mean vectors and covariance matrices of multiple biomarkers, a maximum likelihood estimation method (MLE) can be used [9,29]. Additionally, Tomassi et al. [28] recently proposed a likelihood-based sufficient dimension reduction method for analyzing multiple correlated markers with the LOD. However, these methods cannot be applied to high-dimensional proteomic biomarkers since the sample covariance matrices are singular [5,6,15,16,33], and thus cannot be estimated by maximizing the likelihood function which involves the matrix inverse. To overcome this issue, the literature imposed additional assumptions on the covariance matrix. For example, to ensure positive definiteness, the diagonal matrix was used in the Fisher’s discriminant problem [6,10,30]. However, the assumption of independence among all the protein profiles is overly simplified and does not represent the true correlation structure. An alternative way is to add penalty terms to the likelihood function, such as the lasso-type penalty term. Besides, the lasso-type penalty imposes a sparsity constraint to the covariance/precision matrix, which is a common assumption for high-dimensional data.
Methods dealing with these two challenges in case-control studies have been scarce. A recent paper [1] considered an EM algorithm with the Gaussian graphical lasso for censored data. However, the emphasis of their method is on the relationship between gene biomarkers in an one-sample problem, and they did not consider biomarker combination for improving diagnostic accuracy. To deal with these two challenges when searching for the optimal linear combination coefficients, we propose a novel resample-replace lasso procedure to estimate the coefficients so that the AUC of the biomarker combination is maximized. In our procedure, we first impute the values below LOD using regression predictions, and then use a graphical lasso approach [11] on the fully imputed dataset to obtain mean vectors and precision matrix estimates for cases and controls. Finally, the estimates are then plugged into the expression of optimal linear combination coefficients for biomarkers [27] to get the optimal biomarker combination.

The rest of the paper is organized as follows. Section 2 gives the details of the proposed procedure, where a resample-replace imputation procedure is proposed, and the graphical lasso method is then used to estimate the mean vectors and covariance matrices. In Section 3, simulation studies are conducted to compare the performance of the proposed method with that of alternative methods that replace NA values by LOD values or remove observations with NA values. An application to a protein profiling study of glioblastoma patients on their survival status is given in Section 4. The proof is provided in Appendix.

2. Method

We introduce notations in Section 2.1 and propose a novel method to combine high-dimensional markers with the presence of the LOD in Section 2.2. The consistency of data imputation through the proposed method is given in Appendix.

2.1. Notation

The conventional assumption in this article is that after some known transformation, biomarker observations follow multivariate normal distributions. This distributional assumption has been widely used in the biostatistics literature (see, for example, [9,20,21,25,29]). Suppose that $m + n$ subjects are enrolled in a study and $p$ biomarkers are measured on each of them, where $m$ subjects are sampled from the case population and $n$ subjects are sampled from the control population. Assume that the detection limit of the $k$th biomarker is $d_k$ ($d_k > 0$), $k = 1, \ldots, p$. That is, when the value of the $k$th biomarker on a subject is larger than $d_k$, it can be detected, otherwise it is missing, in which NA is used to denote the missing value. Denote the value of the $k$th biomarker on the $i$th subject by $x_{ik}$ in the case population, and that on the $j$th individual by $y_{jk}$ in the control population. Case observations are denoted by $X = (x_{ik})_{m \times p}$ and control observations are denoted by $Y = (y_{jk})_{n \times p}$.

For the $k$th column of $X$ ($Y$), let $m_k$ ($n_k$) be the number of missing values among $m$ case ($n$ control) subjects. And denote the set that contains the index of observed values of $k$th biomarker as $M_k$ for case and $N_k$ for control, respectively. Let $\mu = (\mu_1, \ldots, \mu_p)^T$ and $\psi = (\psi_1, \ldots, \psi_p)^T$ be the population means of case and control populations, respectively, and $V = (v_{kl})_{p \times p}$ and $W = (w_{kl})_{p \times p}$ be the corresponding covariance matrices, where $\tau$
denotes the transpose of a vector or a matrix. Let $1_L$ be an $L-$dimensional column vector with all units being 1.

References [27] and [17] give the following expression for the coefficient vector of the optimal linear biomarker combination:

$$\eta = (V + W)^{-1}(\mu - \psi). \tag{1}$$

This linear biomarker combination has the maximal AUC value among all linear combinations, and the maximal AUC is

$$\Phi\left(\sqrt{(\mu - \psi)^T(V + W)^{-1}(\mu - \psi)}\right), \tag{2}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution. Denote $\zeta(\eta) = \eta^T(\mu - \psi)/(\eta^TV\eta)^{1/2}$, and $\xi(\eta) = (\eta^TW\eta)^{1/2}/(\eta^TV\eta)^{1/2}$. The partial AUC (pAUC) at a specificity $\alpha$ is

$$\text{pAUC}(\alpha) = \int_0^\alpha \Phi\left(\zeta(\eta) - \xi(\eta)\Phi^{-1}(1 - p)\right) \, dp. \tag{3}$$

2.2. The resample-replace imputation procedure with the graphical lasso

Our procedure starts by first arranging all biomarkers based on their percentages of NA values in an ascending order from left to right in the matrix. The imputation procedure is carried out from left to right as well. If the first biomarker has NA values, the observed values of this biomarker are used to estimate the mean and variance of a truncated normal distribution. Random values following the estimated normal distribution are generated, and those less than its LOD value are used to impute NA values so that the first biomarker has ‘complete’ observations. In many proteomics studies, there is at least one biomarker with complete observations. Thus the aforementioned step is not needed. For the subsequent biomarkers toward the right, each biomarker is imputed in turn via a regression-predict step. That is, its observed values are regressed on the corresponding values of all the biomarkers on its left, and its NA values are replaced by the predicted values plus random noises, which all come from the regression model. After finishing the imputation, we estimate mean vectors and precision matrices – inverse of covariance matrices, for cases and controls through the graphical lasso method. The graphical lasso [11] minimizes the negative log-likelihood with an $L_1$ penalty on the entries of the precision matrix. The estimated mean vectors and precision matrices are then used to obtain the optimal biomarker combination according to (1).

The first step of our procedure is to rearrange the columns of $X$ and $Y$ to impute missing values. Since imputation steps are carried out separately for case and control data, the biomarker sequences for the two population data after rearrangement do not need to be the same. We rearrange the biomarkers in an increasing order based on the missing data percentage, and for convenience we still use $X$ and $Y$ to denote the data after rearrangement. That is, we have $0 \leq m_1 \leq m_2 \leq \cdots \leq m_p < m$ and $0 \leq n_1 \leq n_2 \leq \cdots \leq n_p < n$. Let $a$ and $b$ be the indices such that $m_a > 0$ and $m_{a-1} = 0$, and $n_b > 0$ and $n_{b-1} = 0$. For notational convenience, always let $m_0 = 0$ and $n_0 = 0$.

Starting with the cases, if $a > 1$, we have at least one complete biomarker and we can directly carry out regression analysis to impute the missing values sequentially. If $a = 1$,
we first use the truncated normal distribution to model the \( m - m_1 \) observed values of the first biomarker

\[
L_1(\mu_1, v_1) = \prod_{i \in M_1} \frac{\phi\left(\frac{x_{i1} - \mu_1}{\sqrt{v_{11}}}\right)}{1 - \Phi\left(\frac{d_{i1} - \mu_1}{\sqrt{v_{11}}}\right)},
\]

with the log-likelihood function being

\[
l_1(\mu_1, v_{11}) = -\sum_{i \in M_1} \frac{(x_{i1} - \mu_1)^2}{2v_{11}} - (m - m_1) \ln \left[1 - \Phi\left(\frac{d_{i1} - \mu_1}{\sqrt{v_{11}}}\right)\right] - (m - m_1) \ln \sqrt{v_{11}} - \frac{m - m_1}{2} \ln(2\pi),
\]

where \( \phi(\cdot) \) denotes the probability density function of a standard normal distribution. By maximizing it, we get the estimates of the mean and variance of the first biomarker. Let \( (\hat{\mu}_1, \hat{v}_{11}) = \arg\max_{\mu_1, v_{11}} l_1(\mu_1, v_{11}) \). Then we generate \( m_1 \) random numbers from the estimated truncated distribution with the following density function

\[
f_1(x_1) = \frac{\phi\left(\frac{x_{11} - \hat{\mu}_1}{\sqrt{\hat{v}_{11}}}\right)}{\Phi\left(\frac{d_{11} - \hat{\mu}_1}{\sqrt{\hat{v}_{11}}}\right)}, x_1 \in (-\infty, d_1),
\]

denoted by \( \hat{x}_{11}, \ldots, \hat{x}_{m_1} \), to replace the NA values of the first biomarker. This estimatesample step is to ensure that there is at least one complete biomarker to initiate the regression procedure.

The regression analysis is conducted on the remaining biomarkers of \( \mathbf{X} \). We set \( k \) to start from \( a \) if \( a > 1 \), or from \( a + 1 \) if \( a = 1 \). The \( k \)th biomarker is regressed on all the biomarkers on its left. To be specific, the responses are \( \{x_{ik}, i \in M_k\} \), the predictor variables are \( \{(x_{i1}, \ldots, x_{i,k-1}), i \in M_k\} \), and the model is:

\[
x_{ik} = \beta_{k0} + \beta_{k1}x_{i1} + \cdots + \beta_{k,k-1}x_{i,k-1} + \epsilon_k, i \in M_k,
\]

where \( \beta_{k0} \) is the intercept term, \( \beta_{kl}, l = 1, \ldots, k-1 \) are coefficients and \( \epsilon_k \) is the normal error term with mean 0 and unknown variance \( \sigma_k^2 \). Denote \( \beta_k = (\beta_{k0}, \ldots, \beta_{k,k-1})^\top, \mathbf{y}_k = (x_{ik})_{(m-m_k) \times 1}, \mathbf{x}_k = (x_{i(l)}(m-m_k) \times (k-1), i \in M_k, l = 1, \ldots, k-1, \mathbf{Z}_k = [1_{m-m_k}, \mathbf{x}_k] \). When the rank of \( \mathbf{Z}_k \) is less than \( m - m_k \), apply principal component analysis to \( \mathbf{Z}_k \) and select the top ranked principal components that can explain 80% variance (80% is a hyperparameter here, the choice of which is studied in the supplemental material) of \( \mathbf{Z}_k \) and use them to replace \( \mathbf{Z}_k \) in the regression analysis. The least square estimate of \( \beta_k \) is given by \( \hat{\beta}_k = (\hat{\beta}_{k0}, \ldots, \hat{\beta}_{k,k-1})^\top = (\mathbf{Z}_k^\top \mathbf{Z}_k)^{-1}\mathbf{Z}_k^\top \mathbf{y}_k \). And the corresponding estimate of \( \sigma_k^2 \) is \( \hat{\sigma}_k^2 = ||\mathbf{y}_k - \mathbf{Z}_k \hat{\beta}_k||^2/(m - m_k) \). Thus the imputed value for \( x_{ik} \) is

\[
\hat{x}_{ik} = \hat{\beta}_{k0} + \hat{\beta}_{k1}x_{i1} + \cdots + \hat{\beta}_{k,k-1}x_{i,k-1} + \epsilon_{ik}, i \in M_k,
\]

where \( \epsilon_{ik} \) is generated from the normal distribution \( \mathcal{N}(0, \hat{\sigma}_k^2) \). The above regression process is repeated from \( k = a \) (or \( a + 1 \)) to \( k = p \). After all the missing values in \( \mathbf{X} \) are imputed, we rearrange its columns back to the original order, and denote the complete version as \( \hat{\mathbf{X}} \).

The next step of our procedure is to apply the graphical lasso method [11] to the fully imputed case data \( \hat{\mathbf{X}} \) to estimate the mean vector and the bprecision matrix. Denote
\( \hat{X} = (\hat{x}_{ij})_{m \times p} = (\hat{x}_{i1}^\top, \ldots, \hat{x}_{im}^\top)^\top \), where \( \hat{x}_i = (\hat{x}_{i1}, \ldots, \hat{x}_{ip})^\top \), representing the \( i \)th subject. And let \( \Theta = V^{-1} = (\theta_{kl})_{p \times p} \). Then up to a constant, the penalized log-likelihood is

\[
l(\mu, \Theta) = \log \det(\Theta) - \frac{1}{m} \sum_{i=1}^{m} (\hat{x}_i - \mu)^\top \Theta (\hat{x}_i - \mu) - \lambda \sum_{k \neq l} |\theta_{kl}|,
\]

where \( \lambda > 0 \) is the tuning parameter. Setting \( \partial l(\mu, \Theta)/\partial \mu = 0 \), we have \( \hat{\mu} = (\frac{1}{m} \sum_{i=1}^{m} \hat{x}_{i1}, \ldots, \frac{1}{m} \sum_{i=1}^{m} \hat{x}_{ip})^\top \). After plugging in \( \hat{\mu} \), maximizing \( l(\mu, \Theta) \) over \( \Theta \) is equivalent to maximizing

\[
\log \det(\Theta) - \text{tr}(S) - \lambda \sum_{k \neq l} |\theta_{kl}|,
\]

where \( S = \frac{1}{m} \sum_{i=1}^{m} (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^\top \). We use the R package \texttt{glasso} to obtain the estimate of \( V \), and denote it as \( \hat{V} \).

Repeating the same imputation-estimation steps on the control data \( Y \), we get the estimate of \( \psi \) and \( W \), denoted as \( \hat{\psi} \) and \( \hat{W} \), respectively. Plugging these estimate \( \hat{\mu}, \hat{\psi}, \hat{V}, \hat{U} \) into (1) gives the optimal linear coefficients. The estimates AUC and pAUC of the optimal biomarker combination can be obtained by plugging the mean and covariance matrices in (2) and (3), respectively.

Note that in many cases we adopt the assumption that the case and the control subject have the same covariance matrix, that is \( V = W \). Under this condition, the mean vectors \( \mu \) and \( \psi \), and the covariance matrix \( V \) can be estimated using the following penalized log-likelihood

\[
l(\mu, \psi, \Theta) = \log \det(\Theta) - \frac{1}{m + n} \left[ \sum_{i=1}^{m} (\hat{x}_i - \mu)^\top \Theta (\hat{x}_i - \mu) + \sum_{j=1}^{n} (\hat{y}_j - \psi)^\top \Theta (\hat{y}_j - \psi) \right] \\
- \lambda \sum_{k \neq l} |\theta_{kl}|.
\]

And the subsequent calculation is the same as described above.

The aforementioned imputation-estimation procedures are repeated \( r \) times to generate \( r \) AUC estimators, \( \text{AUC}_t \), and \( r \) pAUC estimators, \( \text{pAUC}_t \), for \( t = 1, \ldots, r \). The final AUC and pAUC estimates are then given by \( \text{AUC} = \sum_{t=1}^{r} \text{AUC}_t / r \), and \( \text{pAUC} = \sum_{t=1}^{r} \text{pAUC}_t / r \), respectively.

### 3. Numerical studies

In this section, we present a series of simulation studies to demonstrate the performance of the proposed resample-replace imputation graphical lasso method (abbreviated as RIG), and compare it with two other methods, one substituting NA values with the LOD values (abbreviated as SNL) and the other ignoring the observations with NA values (abbreviated as IGN). Let the dimension of biomarkers \( p = 120 \), the sample size combination \( (m, n) \) take on values from \{ (50, 50), (75, 75), (100, 100) \}, and the mean vectors \( \mu = (11, 11, \ldots, 11)^\top \) and \( \psi = (10, 10, \ldots, 10)^\top \). The covariance matrices \( V = W \) are both block diagonal.
matrices, with main diagonal blocks equal to $U_0$ and off-diagonal matrices as zero matrices, where $U_0$ is a $30 \times 30$ matrix with diagonal elements being 1 and off-diagonal elements being $\rho$, where $\rho$ is chosen from $\{0.2, 0.5, 0.8\}$. The detection limits are set to be 7.5 or 8.5 and the specificity $\alpha$ takes on values from $\{0.05, 0.1, 0.2, 1\}$. For each parameter setting, we generate 1000 datasets and estimated pAUCs at different specificities for each of these datasets. Three methods of interest, i.e. RIG, SNL and IGN, are compared based on the mean-squared error (MSE) of all these estimates. The true values of corresponding pAUCs are calculated via expression (3) above.

The ratios of MSE between SNL and RIG methods are shown in Tables 1 and 2. Table 1 shows the ratios calculated from estimates of pAUCs at $\alpha = 0.05$ and 0.1, and Table 2 shows the values at $\alpha = 0.2$ and 1. It can be seen that in all settings the proposed RIG method performs better than the SNL method since all the ratios are larger than 1. From both tables, we can see that 1) For the same sample size combination, the ratios increase as $\rho$ increases. This is expected since the RIG method is based on the idea that the conditional mean of a biomarker is a linear function of other biomarkers. When the correlation $\rho$ is larger, the mean of the data to be imputed can be estimated more precisely. 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) For the same correlation $\rho$, the ratio becomes larger as the sample sizes get larger. And the results in the tables indicate that the estimates of the proposed RIG method are closer to their true values. For example, when $\rho = 0.5$ and the LOD=8.5, the corresponding ratios for pAUC(0.05) are 1.003, 1.047 and 1.124, respectively, for the sample sizes $(m, n) = (50, 50), (75, 75)$ and $(100, 100)$.

The ratios of MSEs of IGN over RIG are also given in Tables 1 and 2. It can be seen that in most scenarios, the proposed RIG method performs better than the IGN method. In both tables, there are some NA values. This is because the IGN method may remove all the samples when the LOD level is relative high and the sample size is small. Thus it is not feasible to obtain parameter estimates. In that case, we record the result as NA in the table.
Table 2. Ratios of MSE for pAUC (α = 0.2) and AUC.

| (m, n) | ρ  | LOD | SNL/RIG | IGN/RIG | SNL/RIG | IGN/RIG |
|-------|----|-----|---------|---------|---------|---------|
| (50,50) | 0.2 | 7.5 | 1.000   | 0.996   | 1.000   | 0.995   |
|       |    | 8.5 | 1.000   | NA      | 1.000   | NA      |
|       | 0.5 | 7.5 | 1.002   | 0.995   | 1.002   | 0.995   |
|       |    | 8.5 | 1.002   | 0.937   | 1.002   | 0.819   |
|       | 0.8 | 7.5 | 1.041   | 1.037   | 1.033   | 1.030   |
|       |    | 8.5 | 1.038   | 0.957   | 1.030   | 0.908   |
| (75,75) | 0.2 | 7.5 | 1.001   | 1.001   | 1.001   | 1.001   |
|       |    | 8.5 | 1.001   | 0.975   | 1.001   | 0.892   |
|       | 0.5 | 7.5 | 1.034   | 1.034   | 1.030   | 1.030   |
|       |    | 8.5 | 1.032   | 1.023   | 1.028   | 1.005   |
|       | 0.8 | 7.5 | 1.156   | 1.156   | 1.127   | 1.127   |
|       |    | 8.5 | 1.158   | 1.153   | 1.129   | 1.122   |
| (100,100) | 0.2 | 7.5 | 1.011   | 1.01    | 1.009   | 1.010   |
|        |    | 8.5 | 1.010   | 1.010   | 1.008   | 1.006   |
|        | 0.5 | 7.5 | 1.091   | 1.102   | 1.081   | 1.090   |
|        |    | 8.5 | 1.090   | 1.102   | 1.081   | 1.090   |
|        | 0.8 | 7.5 | 1.286   | 1.305   | 1.238   | 1.252   |
|        |    | 8.5 | 1.279   | 1.317   | 1.233   | 1.263   |

These tables show the following patterns: (1) When the sample sizes remain the same, the larger the correlation ρ is, the better performance the proposed RIG method has. 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 increases from 0.2 to 0.8. (2) When the LOD level is high, the proposed RIG method does not always perform better. The ratios can be less than 1 when (m, n) = (50, 50). But the RIG method still outperforms the IGN method when the sample sizes are not very small. For example, when ρ = 0.2 and the LOD is 8.5, the respective ratios for pAUC(0.05) are NA, 0.993 and 1.01 for the sample size (50,50), (75,75) and (100,100). This is because the estimates of mean vectors and covariance matrices based on the IGN method are biased when sample sizes are small. The ratios get larger as the sample size gets larger. It is more reasonable to compare the IGN and proposed RIG methods under larger sample sizes, since the percentages of NA values in the 1000 estimates for pAUCs obtained from the IGN method are smaller and their influence on estimate accuracy becomes smaller as well. When the sample sizes are larger, the ratios are larger than 1 in general, indicating that the proposed RIG method has better performance.

In practice, not all of the biomarkers will provide useful information for distinguishing the case and control groups. To take this fact into consideration, we designed a new simulation setup: we set the first ten elements of μ to be 11 and the rest to be 10 and keep all the other parameter settings unchanged. The results are presented in Tables 3 and 4. The corresponding results show the same pattern as that in Tables 1 and 2. We note that some ratios are much larger than one, so we denote the ratios that are larger than 5 by more than an order of magnitude as ‘>5’ in the tables. It is noted that in most of the scenarios we considered here, the proposed RIG method performs the best among the three methods. Despite its superiority, RIG has larger MSEs than IGN does when m = n = 75, ρ = 0.2, LOD is 8.5 and α = 0.05, 0.1, 0.2. The reason is that the correlation between variables is weak (only 0.2) and the missing rate is high with a high LOD value. Under this scenario,
theresample-replaceprocedureinRIGisnotasaccurateasotherscenarios, and has larger MSEs.

When there are several biomarkers having the same missing percentage, the order of imputation for these biomarkers will affect the outcome of RIG. This problem can be solved in three ways. The first strategy is to permute the imputation order and take the average of the multiple estimates for mean vectors and for the covariance matrices, as suggested by the reviewers. The second strategy is to randomly set the imputation order and estimate the unknown parameters. Another alternative is to go through all the possible imputation orders and choose the one that produces the largest AUC as the final result. To strike a
balance between accuracy and efficiency, we employ the first strategy. We denote the variation of RIG that uses the first strategy as pRIG. As for handling data with missing values, in addition to the two imputation methods mentioned above (replacing NAs with LOD values and ignoring observations with NAs), there are two other commonly used imputation strategies: replacing NAs with LOD/2 or $\sqrt{LOD/2}$ values and imputing NAs via multiple imputation, which are denoted as SNhL, SNshL and MI, respectively. We compare the performances of these four methods pRIG, SNhL, SNshL and MI through a series of simulation studies. As for pRIG, the number of permutations for each equivalent missing rate of biomarker is 10. The sample sizes are set to be 15, 50 and 100, respectively. And the correlation $\rho$ takes values of 0.2 and 0.8, respectively.

The simulation results are provided in Table 5, in the form of MSE ratios, i.e. the ratios of MSE in pAUC estimations between SNhL and pRIG, SNshL and pRIG, and MI and pRIG. It can be seen that pRIG has significantly smaller MSEs than SNhL and MI in almost all the scenarios considered here. To be specific, when the sample sizes remain the same, the relative performance of pRIG increases as $\rho$ increase. For example, for scenarios where $n = m = 100$, LOD = 7.5, when $\rho = 0.2$, the ratios are about 1.3 in general; while when $\rho = 0.8$, the ratios are near 1.8 for SNhL/pRIG, more than 8 for SNshL/pRIG, and also near 1.8 for MI/pRIG. This is because pRIG is more accurate when variables are more correlated. When the sample sizes are very small ($n = m = 15$), the ratios are higher, which can be explained by the huge variations due to the small sample size. So compared with SNhL, SNshL and MI, pRIG tends to have smaller MSEs and thus performs better in pAUC estimations.

Summing up, the method pRIG is a variation of the proposed RIG that can deal with situations when there are several biomarkers having the same missing percentage. Based on the simulations above, we can see that RIG, together with pRIG, performs the best in pAUC estimations in most of the considered simulations.

4. A protein profiling study

The proposed RIG method is motivated by a protein profiling study of glioblastoma patients on their survival status at the Center for Applied Proteomics and Molecular Medicine at George Mason University. Glioblastoma is one of the most aggressive brain tumors in adults with poor prognosis. Even with the major advances in medical imaging techniques and cancer therapies, the prognosis has not been improved according to [12]. Although the median survival days of glioblastoma patients are less than 12 months, a small percentage of patients may survive more than 36 months [14,18,26]. Reference [2] investigated the mRNA and protein expression profiles of glioblastoma tissues from patients who survived longer than 36 months (classified by authors as long-term survival) and those who survived less than 6 months (classified as short-term survival). The authors identified a significant difference in the mRNA expression profiles of three signaling genes and two cellular acid-binding proteins between long-term and short-term survivors.

It is worth noting that glioblastoma prevalence is about 3 per 100,000 per year according to [12], and studies on this disease usually involve a small number of patients. The study in [2] only has 11 long-term survivors and 12 short-term survivors. And the age of these patients ranges from 33 to 86. All the short-term survivors were older than 50, and long-term survivors were mostly younger with some around 40. Protein profiles were measured...
Table 5. Ratios of MSE in pAUC(α) estimation for pRIG, SNhL, SNshL and MI.

| method   | n  | ρ  | LOD \(\alpha\) | 0.05 | 0.1  | 0.15 | 0.2  | 1    |
|----------|----|----|----------------|------|------|------|------|------|
| SNhL/pRIG | 15 | 0.2 | 7.5           | 10.404 | 19.805 | 28.508 | 36.455 | 81.216 |
|          | 8.5 |      | 10.871 | 19.445 | 26.469 | 32.209 | 55.250 |
|          | 0.8 |      | > 100  | > 100  | > 100  | > 100  | > 100  |
|          | 50  | 0.2 | 7.5           | 1.004 | 1.003 | 1.002 | 1.002 | 1.002 |
|          | 8.5 |      | 1.001 | 1.001  | 1.001  | 1.001  | 1.000  |
|          | 0.8 |      | 1.167 | 1.160  | 1.156  | 1.153  | 1.133  |
|          | 100 | 0.2 | 7.5           | 1.359 | 1.307 | 1.284 | 1.272 | 1.251 |
|          | 8.5 |      | 1.256 | 1.226  | 1.212  | 1.204  | 1.190  |
|          | 0.8 |      | 2.022 | 1.918  | 1.863  | 1.824  | 1.557  |
|          | 8.5 |      | > 100  | > 100  | > 100  | > 100  | > 100  |
| SNshL/pRIG | 15 | 0.2 | 7.5           | 10.477 | 19.889 | 28.551 | 36.418 | 79.872 |
|          | 8.5 |      | 9.921 | 16.878  | 22.172  | 26.251  | 40.299  |
|          | 0.8 |      | > 100  | > 100  | > 100  | > 100  | > 100  |
|          | 50  | 0.2 | 7.5           | 1.003 | 1.002 | 1.002 | 1.002 | 1.001 |
|          | 8.5 |      | 0.997 | 0.998  | 0.998  | 0.998  | 0.998  |
|          | 0.8 |      | 1.163 | 1.157  | 1.153  | 1.150  | 1.130  |
|          | 100 | 0.2 | 7.5           | 1.352 | 1.302 | 1.280 | 1.268 | 1.247 |
|          | 8.5 |      | 1.101 | 1.096  | 1.093  | 1.091  | 1.086  |
|          | 0.8 |      | 8.057 | 8.154  | 8.282  | 8.410  | 10.326  |
|          | 8.5 |      | > 100  | > 100  | > 100  | > 100  | > 100  |
| MI/pRIG  | 15 | 0.2 | 7.5           | 10.167 | 19.516 | 28.321 | 36.485 | 84.878 |
|          | 8.5 |      | 10.337 | 20.237  | 29.753  | 38.693  | 94.792  |
|          | 0.8 |      | > 100  | > 100  | > 100  | > 100  | > 100  |
|          | 50  | 0.2 | 7.5           | 1.004 | 1.003 | 1.003 | 1.003 | 1.002 |
|          | 8.5 |      | 1.001 | 1.001  | 1.001  | 1.001  | 1.000  |
|          | 0.8 |      | 1.167 | 1.161  | 1.157  | 1.153  | 1.134  |
|          | 100 | 0.2 | 7.5           | 1.363 | 1.310 | 1.287 | 1.275 | 1.253 |
|          | 8.5 |      | 1.330 | 1.285  | 1.265  | 1.254  | 1.234  |
|          | 0.8 |      | 2.020 | 1.917  | 1.862  | 1.824  | 1.559  |
|          | 8.5 |      | 32.105 | 34.300  | 35.613  | 36.635  | 47.885  |

on the glioblastoma tissues of these patients. The main goal of the study is to identify the optimal linear combination of these protein profiles that distinguish long-term survival versus short-term survival status. Forty-five biomarker profiles with LOD are of particular interest for the combination. Each of these profiles has less than 30% of observations that are below LOD. We first estimated the mean vectors and covariance matrices. The corresponding results are shown in Figure 1. Using these estimates, we obtained the coefficients for the optimal linear combination based on (1).

Since investigators are particularly interested in partial AUC values at low specificities, we calculated the respective partial AUCs when the specificity is 0.05, 0.1, and 0.2. The estimates of these pAUCs based on the proposed method are 0.05, 0.1, and 0.2, respectively. However, when combining the biomarkers after replacing the NA values with zeros, the corresponding pAUCs estimates are 0.03989, 0.08372, and 0.17503, respectively. This indicates that simply replacing NA values with the LOD values may underestimate the diagnostic accuracy of the combined protein biomarkers for distinguishing short-term and long-term survival status in glioblastoma patients.
Figure 1. The estimate for the real data using RIG. (a) mean estimate for the short-term survival (STS) group; (b) mean estimate for the long-term survival (LTS) group; (c) variance estimate for the STS; (d) variance estimate for the LTS; (e) variance-covariance matrix estimate for the STS; (f) variance-covariance matrix estimate for the LTS.

5. Conclusion

We propose a novel method to combine biomarkers with the presence of limit of detection. The method is motivated by a proteomic study, in which part of the data is missing due to the existence of LOD and the sample sizes are smaller than the number of biomarkers of interest. Since existing methods either replace missing data with constants or remove observations that have missing values, they all lead to biased parameter estimates. Besides, because in this case covariance matrices are essentially singular, traditional methods for
estimating covariance matrices that rely on the inverse of sample covariance matrices are not applicable. To deal with these two challenges, we first propose a resample-replace imputation procedure to impute missing data based on the multivariate normal model and linear regression technique, and then we employ the graphical lasso method to estimate mean vectors and covariance matrices. The simulation studies show that the proposed method outperforms another two widely used methods in terms of estimating partial AUC values for a biomarker combination. Furthermore, a data analysis based on a glioblastoma study conducted by the Center for Applied Proteomics and Molecular Medicine shows that the biomarker combination obtained via the proposed method has higher diagnostic accuracy rate.

Our method rearranges the biomarkers according to their missing percentage in an increasing order, and each biomarker is imputed by being regressed on those biomarkers that have a lower missing ratio. It requires that at least one biomarker has complete observations. This requirement is satisfied in many proteomics studies including the one used in this paper. Although rearranging the biomarkers may use part of the available biomarkers for the imputation, the estimated covariance matrices are still consistent after the graphical lasso. It is feasible to impute the values below LOD without first rearranging biomarkers, but the variation of parameter estimates will increase.

In the imputation procedure, RIG applies principle component analysis to the regression model, where the top ranked principal components that can explain 80% variance are chosen. 80% here is a hyper-parameter (denoted as $q$) and may affect the performance of RIG. To study the sensitivity of RIG to the choice of this hyper-parameter, we compare its MSEs under different $q$ and come to the conclusion that the other commonly used $q$'s in practice make little difference regarding the performance of RIG. Details are provided in the supplemental material.

Alternative versions of the sequential linear regression used in this article have been used to impute missing values in microarray data and survey data [7, 23]. The authors mentioned that the results are satisfactory for a missing rate as large as 30% for a predictor variable. A thorough review of imputation using sequential regression is provided in [4]. However, our paper is the first to show the statistical property of such imputation methods.

A recent paper by Augugliaro et al. [1] also considered using graphical lasso in high-dimensional censored data. Although both their work and ours can handle the LOD in the high-dimensional setting, their focus is different from ours in several aspects. First, they only consider the one-sample problem, while ours is for the two-sample problem. Second, they are interested in estimating relationship between gene biomarkers, while we focus on the classification problem for distinguishing cases from controls.

A similar data configuration to the data with LOD is the mixture distribution of a mass at 0 and the remaining larger than 0. The Tobit model is traditionally used to deal with this kind of data. Under certain conditions, 0’s in the data can be treated as a proxy for values smaller than 0. It is worth noting that the proposed method can also handle this type of data.

Simulations in the main text and in the supplemental material show that the proposed resample-replace imputation strategy produces consistent estimates for pAUCs. We prove that asymptotically the imputed data have the same distribution as a sample from the original corresponding unknown population when $p$ is fixed, and the details are given in the supplemental material. However, proving the consistency in the high-dimensional settings
is non-trivial, and the theory will be developed as a future topic. Moreover, when the number of imputed values grows as sample size increases, the correlation structure between imputed values becomes complicated and is non-trivial to estimate. Therefore, more future work needs to be done to prove consistency in this case. Also, as a referee pointed out, it might be possible to use lasso rather than the PCA method for the imputation. It will be a future topic to appropriately implement lasso and compare the performance with the PCA described in the paper.

**Acknowledgments**

We thank the editor, the associate editor, and the three anonymous referees for the valuable comments on our previous manuscript. We also thank Emanuel Petricoin at George Mason University, Helen Fillmore and William Broaddus at Virginia Commonwealth University for access to the data, generated the proteomic data and the clinical specimens that underpinned the generation of the data.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work is supported in part by the Intramural Research Program of the National Institutes of Health and the US Social Security Administration. The opinions expressed in this article are the author’s own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, US Social Security Administration, or the United States government. Besides, this work is partially supported by Beijing Natural Science Foundation [grant number Z180006], National Nature Science Foundation of China [grant number 11722113].

**References**

[1] L. Augugliaro, A. Abbruzzo, and V. Vinciotti, *L1-penalized censored Gaussian graphical model*, Biostatistics 2018. Available at https://doi.org/10.1093/biostatistics/kxy043.

[2] S. Barbus, B. Tews, D. Karra, M. Hahn, B. Radlwimmer, N. Delhomme, C. Hartmann, J. Felsberg, D. Krex, and G. Schackert, *Differential retinoic acid signaling in tumors of long-and short-term glioblastoma survivors*, J. Natl. Cancer. Inst. 103 (2011), pp. 598–601.

[3] R.C. Bast, *Perspectives on the future of cancer markers*, Clin. Chem. 39 (1993), pp. 2444–2451.

[4] D. Bertsimas, C. Pawlowski, and Y.D. Zhuo, *From predictive methods to missing data imputation: an optimization approach*, J. Mach. Learn. Res. 18 (2017), pp. 7133–7171.

[5] P.J. Bickel, and E. Levina, *Regularized estimation of large covariance matrices*, Ann. Stat. 36 (2008), pp. 199–227.

[6] P.J. Bickel, and E. Levina, *Some theory for Fisher’s linear discriminant function, ‘naïve Bayes’, and some alternatives when there are many more variables than observations*, Bernoulli 10 (2004), pp. 989–1010.

[7] T.H. Bø, B. Dysvik, and I. Jonassen, *Lsimpute: accurate estimation of missing values in microarray data with least squares methods*, Nucleic. Acids. Res. 32 (2004), pp. e34–e34.

[8] J.B. Chook, Y.F. Ngeow, S.F. Yap, T.C. Tan, and R. Mohamed, *Combined use of wild-type hbv precore and high serum iron marker as a potential tool for the prediction of cirrhosis in chronic hepatitis b infection*, J. Med. Virol. 83 (2011), pp. 594–601.

[9] T. Dong, C.C. Liu, E.F. Petricoin, and L.L. Tang, *Combining markers with and without the limit of detection*, Stat. Med. 33 (2014), pp. 1307–1320.
[10] S. Dudoit, J. Fridlyand, and T.P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, J. Am. Stat. Assoc. 97 (2002), pp. 77–87.

[11] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics 9 (2008), pp. 432–441.

[12] O. Gallego, *Nonsurgical treatment of recurrent glioblastoma*, Current Oncol. 22 (2015), pp. e273.

[13] H. Hoffman, and R. Johnson, *Estimation of multiple trace metal water contaminants in the presence of left-censored and missing data*, J. Environ. Stat. 2 (2011), pp. 1–15.

[14] D. Krex, B. Klink, C. Hartmann, A. von Deimling, T. Pietsch, M. Sabel, J. P. Steinbach, O. Heese, G. Reifenberger, M. Weller, and G. Schackert, *Long-term survival with glioblastoma multiforme*, Brain 130 (2007), pp. 2596–2606.

[15] W. Krzanowski, P. Jonathan, W. McCarthy, and M. Thomas, *Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data*, J. Royal Stat. Soc. Series C (Appl. Stat.) 44 (1995), pp. 101–115.

[16] O. Ledoit, and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivar. Anal. 88 (2004), pp. 365–411.

[17] A. Liu, E.F. Schisterman, and Y. Zhu, *On linear combinations of biomarkers to improve diagnostic accuracy*, Stat. Med. 24 (2005), pp. 37–47.

[18] D.N. Louis, H. Ohgaki, O.D. Wiestler, W.K. Cavenee, P.C. Burger, A. Jouvet, B.W. Scheithauer, and P. Kleihues, *The 2007 who classification of tumours of the central nervous system*, Acta Neuropathol. 114 (2007), pp. 97–109.

[19] M.W. McIntosh, and M.S. Pepe, *Combining several screening tests: optimality of the risk score*, Biometrics 58 (2002), pp. 657–664.

[20] N.J. Perkins, E.F. Schisterman, and A. Vexler, *Receiver operating characteristic curve inference from a sample with a limit of detection*, Am. J. Epidemiol. 165 (2006), pp. 325–333.

[21] N.J. Perkins, E.F. Schisterman, and A. Vexler, *Roc curve inference for best linear combination of two biomarkers subject to limits of detection*, Biometrical J. 53 (2011), pp. 464–476.

[22] E.F. Petricoin, D.K. Ornstein, and L.A. Liotta, *Clinical proteomics: applications for prostate cancer biomarker discovery and detection*, in *Urologic Oncology: Seminars and Original Investigations*, Vol. 22, Elsevier, Amsterdam, 2004, pp. 322–328.

[23] T.E. Raghunathan, J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger, *A multivariate technique for multiply imputing missing values using a sequence of regression models*, Surv. Methodol. 27 (2001), pp. 85–96.

[24] J.F. Rusling, C.V. Kumar, J.S. Gutkind, and V. Patel, *Measurement of biomarker proteins for point-of-care early detection and monitoring of cancer*, Analyst 135 (2010), pp. 2496–2511.

[25] E.F. Schisterman, A. Vexler, A. Ye, and N.J. Perkins, *A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity*, Ann. Appl. Stat. 5 (2011), pp. 2651–2667.

[26] Stupp, M.E. Hegi, W.P. Mason, M.J. van den Bent, M.JB. Taphoorn, R.C. Janzer, S.K. Ludwin, A. Allgeier, B. Fisher, K. Belanger, P. Hau, A.A. Brandes, J. Gijtenbeek, C. Marosi, C.J. Vecht, K. Mokhtari, P. Wesseling, S. Villa, E. Eisenhauer, T. Gorlia, M. Weller, D. Lacombe, J.G. Cairncross and R. Mirimanoff. *Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-n cic trial*, Lancet Oncol. 10 (2009), pp. 459–466.

[27] J.Q. Su, and J.S. Liu, *Linear combinations of multiple diagnostic markers*, J. Am. Stat. Assoc. 88 (1993), pp. 1350–1355.

[28] D. Tomassi, L. Forzani, E. Bura, and R. Pfeiffer, *Sufficient dimension reduction for censored predictors*, Biometrics 73 (2017), pp. 220–231.

[29] A. Vexler, A. Liu, E. Eliseeva, and E.F. Schisterman, *Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection*, Biometrics 64 (2008), pp. 895–903.

[30] D.M. Witten, and R. Tibshirani, *Penalized classification using fisher’s linear discriminant*, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 73 (2011), pp. 753–772.
[31] Y. Yasui, M. Pepe, M.L. Thompson, B.L. Adam, G.L. Wright Jr, Y. Qu, J.D. Potter, M. Winget, M. Thornquist, and Z. Feng, *A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection*, Biostatistics 4 (2003), pp. 449–463.

[32] Z. Yuan, and D. Ghosh, *Combining multiple biomarker models in logistic regression*, Biometrics 64 (2008), pp. 431–439.

[33] H. Zou, T. Hastie, and R. Tibshirani, *Sparse principal component analysis*, J. Comput. Graph. Stat. 15 (2006), pp. 265–286.

[34] A. Zupa, G. Improtta, J. Deng, M. Aieta, P. Musto, L. Liotta, C. Belluco, E. Mammano, J. Wulfkuhle, and E. Petricoin III, *Use of protein pathway activation mapping of nsclc to identify distinct molecular subtypes and a prognostic signature for aggressive node-negative tumors.*, J. Clin. Oncol. 28 (2010), pp. 10594–10594.