Yu and Park *BMC Genomics* 2014, **15**(Suppl 10):S1
http://www.biomedcentral.com/1471-2164/15/S10/S1

BMC Genomics

RESEARCH Open Access

AucPR: An AUC-based approach using penalized regression for disease prediction with high-dimensional omics data

Wenbao Yu[1], Taesung Park[1,2*]

From The 25th International Conference on Genome Informatics (GIW/ISCB-Asia)
Tokyo, Japan. 15-17 December 2014

Abstract

Motivation: It is common to get an optimal combination of markers for disease classification and prediction when multiple markers are available. Many approaches based on the area under the receiver operating characteristic curve (AUC) have been proposed. Existing works based on AUC in a high-dimensional context depend mainly on a non-parametric, smooth approximation of AUC, with no work using a parametric AUC-based approach, for high-dimensional data.

Results: We propose an AUC-based approach using penalized regression (AucPR), which is a parametric method used for obtaining a linear combination for maximizing the AUC. To obtain the AUC maximizer in a high-dimensional context, we transform a classical parametric AUC maximizer, which is used in a low-dimensional context, into a regression framework and thus, apply the penalization regression approach directly. Two kinds of penalization, lasso and elastic net, are considered. The parametric approach can avoid some of the difficulties of a conventional non-parametric AUC-based approach, such as the lack of an appropriate concave objective function and a prudent choice of the smoothing parameter. We apply the proposed AucPR for gene selection and classification using four real microarray and synthetic data. Through numerical studies, AucPR is shown to perform better than the penalized logistic regression and the nonparametric AUC-based method, in the sense of AUC and sensitivity for a given specificity, particularly when there are many correlated genes.

Conclusion: We propose a powerful parametric and easily-implementable linear classifier AucPR, for gene selection and disease prediction for high-dimensional data. AucPR is recommended for its good prediction performance. Beside gene expression microarray data, AucPR can be applied to other types of high-dimensional omics data, such as miRNA and protein data.

Background

Nowadays, it is easy and common to measure thousands of markers simultaneously through high-throughput technologies, for example, the microarray study. A disease is usually related to several markers and the combination of multiple makers for classifying a subject into different statuses of a specific disease is widely studied. The performance of a combination of markers is frequently measured by indices related to the Receiver Operating Characteristic (ROC) curve: sensitivity, specificity, or the area under the ROC curve (AUC). Sensitivity (specificity) is defined as the probability of success in classifying a diseased (non-diseased) individual accurately. By varying the decision rules (thresholds), different sensitivities and specificities are obtained. The ROC curve plots all possible sensitivities against 1-specificities and expresses the trade-off between sensitivity and specificity visually. AUC is the most popular summary index for the curve; it has been shown to be the probability that the score of a randomly chosen diseased

* Correspondence: tspark@stats.snu.ac.kr
[1]Department of statistics, Seoul National University, Shilim-dong, Kwanak-gu 151-742, Seoul, Korea
Full list of author information is available at the end of the article

BioMed Central

© 2014 Yu and Park; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

individual exceeds that of a randomly chosen non-diseased subject [1].

Therefore, it is natural to construct a combination of markers in order to maximize the ROC-based metrics. A number of combinations based on ROC indices have been suggested by [2-9]. Among these, [3] and [5] developed distribution-free methods to achieve the best linear combination for maximizing the smoothed AUC for high-dimensional situations. They developed algorithms based on optimizing a sigmoid approximation of AUC. The sigmoid approximation of AUC relies on a smoothing parameter, which should be carefully chosen, though there are no theoretical guidelines for choosing this parameter. The rule of thumb for the choice of the smoothing parameter may reduce the power of the method. Moreover, the sigmoid approximation of AUC is not a concave function and multiple local maxima may exist. The global maximum is not guaranteed to be attained through commonly used numeric algorithms. For example, the performance of the linear combination decided by [5] is very poor for microarray data [9]. To avoid the difficulties of maximizing a non-parametric approximation of AUC, we can use a parametric method. To our knowledge, there is no published parametric method for maximizing the AUC under a high-dimensional context. This paper tries to fill this gap.

We suggest an AUC-based approach using penalized regression (AucPR), based on a classical parametric linear combination derived by [2] in a low-dimensional context. The problem is then transformed into a linear regression framework, and the existing software for solving linear regression with penalization can be used directly, which facilitates the implementation of the proposed method. There are many penalty functions available, for example, the elastic net criterion [10], which is a mixture of penalties of L_1 and L_2 norms of the linear coefficients. The lasso penalty [11] is a special form of elastic net. Both the lasso and the elastic net have been widely used for marker selection and disease classification for high-dimensional data [3,5,9,10,12,13]. In this work, we maximize AUC through elastic net or lasso penalty. We compare the proposed AucPR to a logistic regression with elastic net or lasso penalty and the AUC-based non-parametric method proposed by [3], through four microarray data sets and synthetic data. The performance is gauged on the AUC and sensitivity given specificity equals to 0.95 on testing samples. AucPR achieves better prediction performance.

Methods

AucPR: An AUC-based approach using penalized regression

Suppose non-diseased samples $\{\boldsymbol{X_i}; 1 \le i \le m\}$ and diseased samples $\{\boldsymbol{Y_j}; 1 \le j \le n\}$ are independent and identically distributed (i.i.d.) from multivariate normal distributions $N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and $N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, respectively, where $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are p-dimension mean vectors, and $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ are$p \times p$ covariance matrices.

Under the multivariate normal distribution assumption, [2] showed that, among all possible linear combination of markers, without a positive constant multiplier, the combination with the coefficient vector

$$\beta = (\sum\nolimits_x + \sum\nolimits_y)^{-1}(\mu_y - \mu_x) \tag{1}$$

is optimum for maximizing the AUC. Furthermore, they also proved that if $\boldsymbol{\Sigma}_x$ is proportional to $\boldsymbol{\Sigma}_y$, $\boldsymbol{\beta}$ is uniformly optimum, that is, it achieves the highest ROC curve among all linear combinations for all possible values of specificity.

Although this approach has been widely used in disease classification [14-17], it cannot be applied directly to high-dimensional problems, where the number of markers (p) are larger than the number of observations in the sample. Penalized regression methods such as lasso [11] and elastic net [10], are effective tools for variable selection in high-dimensional problems. We thus try to restate our problem in a regression framework.

Note that from Equation (1), $\boldsymbol{\mu}_y - \boldsymbol{\mu}_x = (\boldsymbol{\Sigma}_y + \boldsymbol{\Sigma}_x)\boldsymbol{\beta}$ holds. Instead of solving this equation, we suggest approximating β by solving the following linear regression problem:

$$\mu_y - \mu_x = (\sum\nolimits_y + \sum\nolimits_x)\beta + \in, \quad \in \sim N(0, \sigma^2 I), \tag{2}$$

where $\boldsymbol{I}$ is a $p \times p$ identity matrix. By this transformation, we can avoid calculating the inverse of a large covariance matrix in (1), which is intractable due to lack of samples.

We then propose using a regularized linear regression method to obtain $\boldsymbol{\beta}$. Let $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y + \boldsymbol{\Sigma}_x = ((\sigma_{ij}))$, $1 \le i, j \le p$, and $\boldsymbol{\mu} = \boldsymbol{\mu}_y - \boldsymbol{\mu}_x = (\mu_1, \ldots, \mu_p)'$. Then, using the elastic net, we have

$$\beta = \arg\min_{\beta \in R^p} \sum_{j=1}^{p} (\mu_i - \sum_{j=1}^{p} \sigma_{ij}\beta_j)^2 + \lambda(\alpha \sum_{i=1}^{p} |\beta_i| + \frac{1-\alpha}{2} \sum_{i=1}^{p} \beta_i^2), \tag{3}$$

where λ is a parameter controlling the strength of the penalty and α is a mixing parameter that determines the relative strength of the L_1 norm to the L_2 norm, with $0 \le \alpha \le 1$. When $\alpha = 1$, the elastic net reduces to lasso. The elastic net encourages a group of highly correlated markers to enter the model together, while lasso is quite parsimonious in selecting correlated markers. Under some conditions, both penalties were shown to have consistency in model selection [18,19], or in other words, the selected model includes the true model with a high probability.

In practice, we replace the covariance matrices with the sample covariance matrices, and the mean vectors with the sample mean vectors. Formally,

$$\widehat{\mu}_x = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \widehat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$\widehat{\sum}_x = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \widehat{\mu}_x)(X_i - \widehat{\mu}_x)',$$

$$\widehat{\sum}_y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_y)(Y_i - \widehat{\mu}_y)',$$

and

$$\widehat{\mu} = \widehat{\mu}_y - \widehat{\mu}_x, \quad \widehat{\sum} = \widehat{\sum}_y + \widehat{\sum}_x.$$

The idea of the proposed AucPR is similar to a procedure proposed by [20] for the sparse linear discriminant analysis (LDA), where they restrict L_∞ error and obtain the combination by linear programming. When Σ_x and Σ_y are proportional to each other, $\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ is proportional to the coefficient vector of LDA. In this sense, AucPR also provides a solution for sparse LDA.

There are several computationally efficient algorithms to implement penalized linear regression for high-dimensional data, for example, program *lars* by [21] and *glmnet* by [22]. In this paper, we use *glmnet* to solve Equation (3), since it is more efficient than *lars* [22].

Remark 1: We use sample mean vectors and sample covariance matrices, which are quite sensitive to the outlier observations. Therefore, intuitively, it may lead to the proposed method being inefficient under a general mean and a covariance structure without any restriction, especially when the sample sizes are small. However, AucPR can be powerful for some structures of $\mathbf{\Sigma}$ and $\boldsymbol{\mu}$, for example, when $\mathbf{\Sigma}$ or $\boldsymbol{\mu}$ are sparse, which is common in high-dimensional data. We illustrate this with numerical studies in the Result and discussion Section.

Choice of tuning parameter

The tuning parameter λ controls the trade-off between data fitting and model complexity. Given a larger λ, fewer markers are selected and the data may not be well fitted, while for a smaller λ, a larger number of markers are chosen and overfitting may occur. We tune λ in our numeric studies by a three-fold cross-validation (CV) method. Note that when the sample sizes are large, we can use a K-fold CV with $K > 3$.

For the K-fold CV, we randomly divide the samples into K subsets of equal size. We select λ that maximizes the following CV score:

$$CV_\lambda = \sum_{i=1}^{k} \widehat{AUC}_\lambda^{(i)}(\hat{\beta}_\lambda^{(-i)}), \tag{4}$$

where $\hat{\beta}_\lambda^{(-i)}$ is the coefficient vector estimated without the samples in the i-th fold, and $\widehat{AUC}_\lambda^{(i)}$ is the empirical AUC estimator with the data in the i -th fold, for a given λ, $i = 1, \ldots, K$. The empirical AUC estimator for a given $\boldsymbol{\beta}$ is defined as $\sum\sum I(\beta'(Y_i - X_j) > 0)/nm$, with $I(\cdot)$ being the indicator function.

For the elastic net, α is fixed at 0.5 in this investigation. We note that although α can be tuned in the same fashion as λ, a simple, fixed α still captures the characteristics of the elastic net and is widely used in the literature as well [13,23].

Another practical issue about tuning the parameter λ is how to provide the candidates of λ for CV, as it has not been specified clearly in the literature. We propose finding the range of λ using the whole data, and then generating a fixed number of candidates within that range such that they are evenly distributed in the log scale. Denoting the range of candidates for λ as $[\lambda_l, \lambda_u]$, where λ_l corresponds to the most complex model (for example, 100 markers are selected) while λ_u corresponds to the least complex model (for example, 1 marker is chosen). It is easy to use the bisection method [24] to fix $\lambda = \lambda_k$, such that there are exactly k non-zero coefficients ($k = 1,\ldots, p$). To do this, we first have an initial guess at the value of λ. Let $r(\lambda)$ be the number of non-zero coefficients of the tuning parameter λ. If $r(\lambda) = k$, we are done. If $r(\lambda) < k$, we let $\lambda = \lambda/2$, continuing this until $r(\lambda) \geq k$. Once we have an interval $[\lambda_1, \lambda_2]$, we employ the bisection method. We test the middle point $\lambda_m = (\lambda_1 + \lambda_2)/2$, and if $r(\lambda_m) = k$, we are done. If $r(\lambda_m) < k$, set $\lambda_2 = \lambda_m$; otherwise, set $\lambda_1 = \lambda_m$. Repeat until $r(\lambda_m) = k$.

Results and discussion

Application to gene selection and cancer classification

In this section, we apply our proposed AucPR, the penalized logistic regressions, and the AUC based non-parametric method proposed by [3], which maximizes a sigmoid approximation of AUC, to four microarray datasets for gene selection and cancer classification. We refer to our approaches to AucPR with elastic net and lasso penalty as *AucEN* and *AucL*, respectively, the logistic regression approaches with elastic net and lasso penalties as *LogEN* and *LogL*, respectively, and maximizing the sigmoid approximation of AUC as *MSauc* in the following content. The four microarray data sets are:

Brain cancer data: The original data have five different types of tumors, and 42 samples with 5597 expressions. This data set was also studied by [25], and we use their preprocessed data and denote the first two types as the control group and the other three as the case group.

It can be downloaded from http://stat.ethz.ch/~dettling/bagboost.html.

- *Colon* cancer data: Expression levels of 40 tumors and 22 normal colon tissues for 2000 human genes, with the highest minimal intensity from 62 subjects measured [26]. The data can be downloaded from *colonCA* package on the Bioconductor website (http://www.bioconductor.org).
- *Leukemia* data: We consider two types of leukemia cancer: acute myeloid leukemia (AML) and acute lymphoblast leukemia (ALL). Samples used by [27] were derived from 47 patients with ALL and 25 patients with AML, with 7129 genes. The data set is available in the *golubEsets* package on the Bioconductor website (http://www.bioconductor.org).
- *DLBCL* data: The diffused large B-cell lymphoma (DLBCL) data set contains 58 DLBCL patients and 19 follicular lymphoma patients from a related germinal center B-cell lymphoma [28]. The data are available from the Broad Institute website (http://www.genome.wi.mit.edu/MPR/lymphoma).

All data sets are further processed using quantile normalization and logarithm transformation (except the *Brain* cancer data, since it has been preprocessed). To save computation time, we screen the genes such that the 1000 genes with the largest absolute moderated t-statistics [29] are kept. Filtering genes by t-typed statistics has been widely used in the literature, for example, [3,5,20] among others. Our empirical study shows that including more than 1000 genes does not significantly change the patterns found. *LogEN* and *LogL* are also implemented by R package *glmnet* and the tuning parameter is chosen by a three-fold CV, using the CV score defined in Equation (4).

Then, we randomly split the data into training and testing sets, comprising 2/3 and 1/3 of the sample, respectively. The AUC value and sensitivity when specificity = 0.95 are evaluated based on the testing set. This procedure is repeated 100 times and the box-plots of the two comparison metrics are plotted in Figures 1 - 4.

We can see that the proposed AucPR outperforms the other approaches for all the four datasets,. The *AucEN* has the best prediction performance. The *AucL* is slightly less powerful than the *AucEN*, but better than

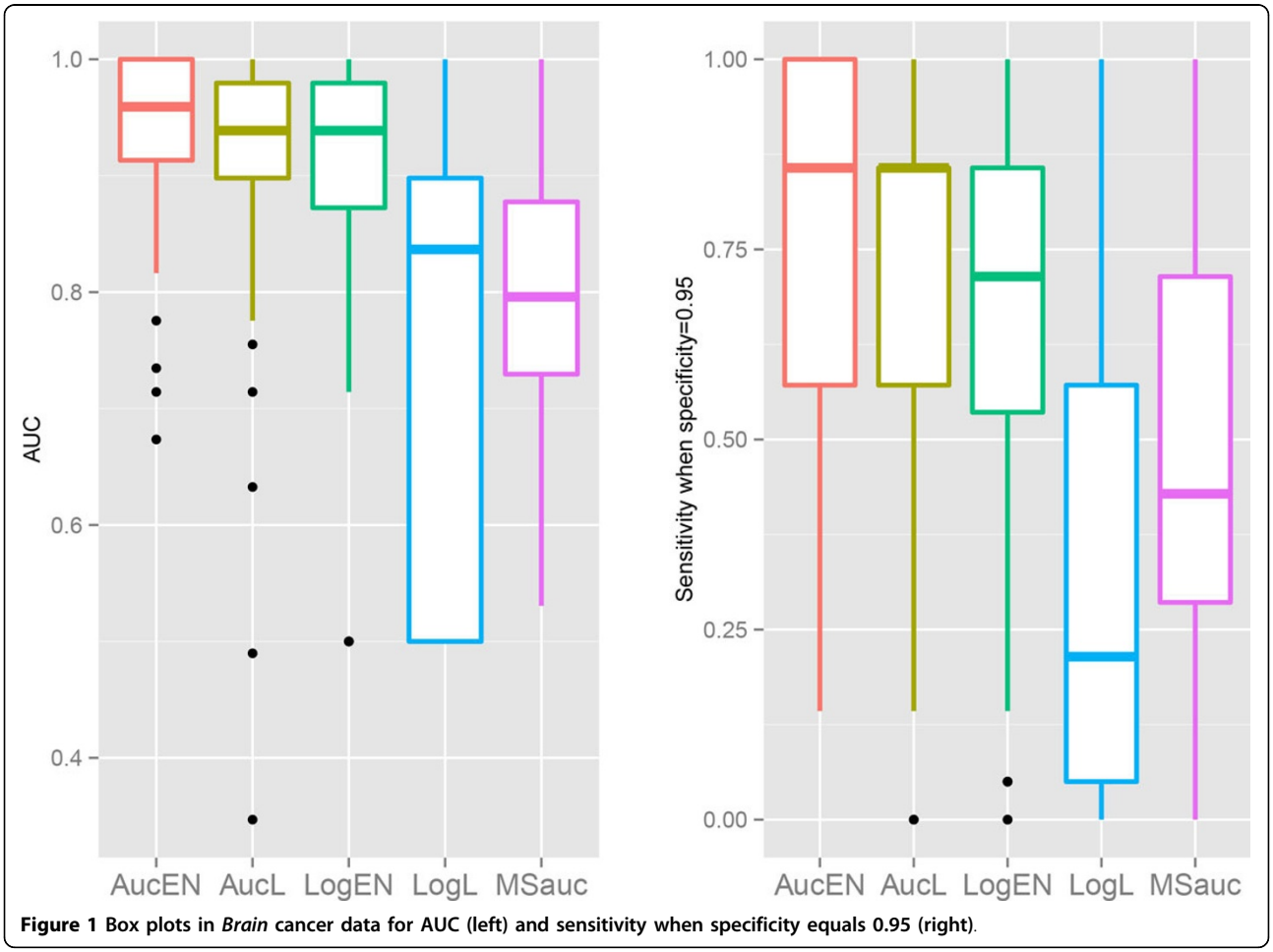

Figure 1 Box plots in *Brain* cancer data for AUC (left) and sensitivity when specificity equals 0.95 (right).

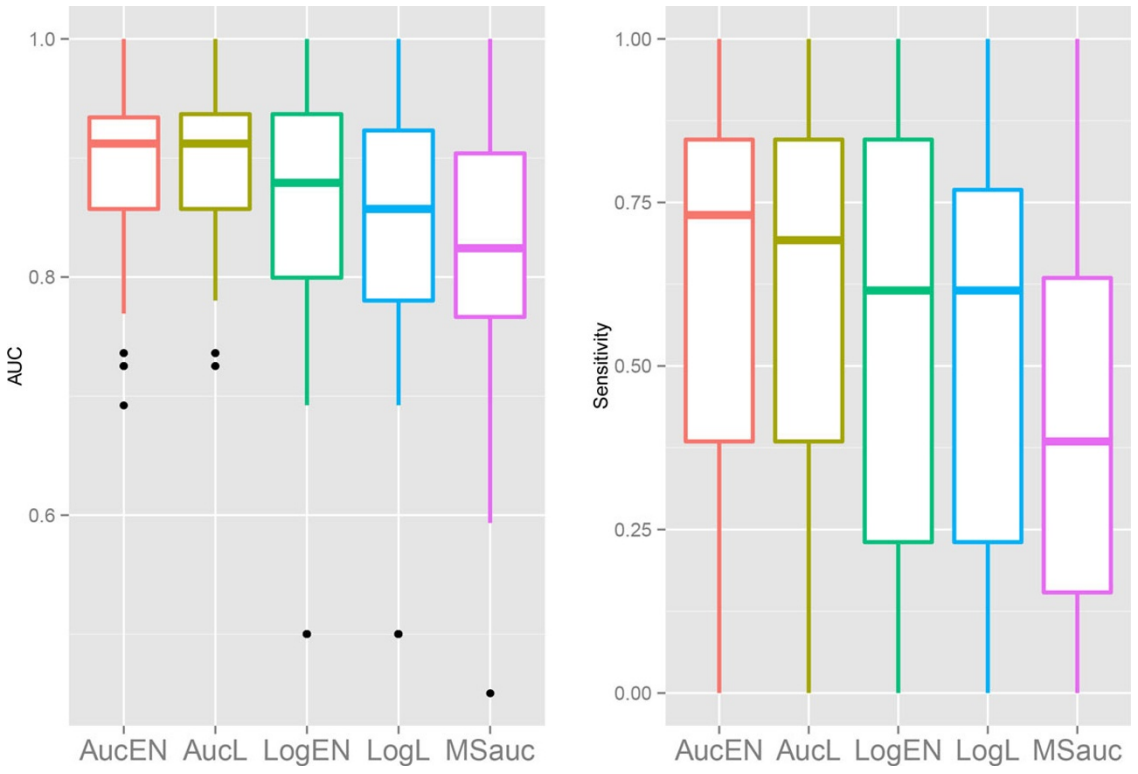

Figure 2 Box plots in *Colon* cancer data for AUC (left) and sensitivity when specificity equals 0.95 (right).

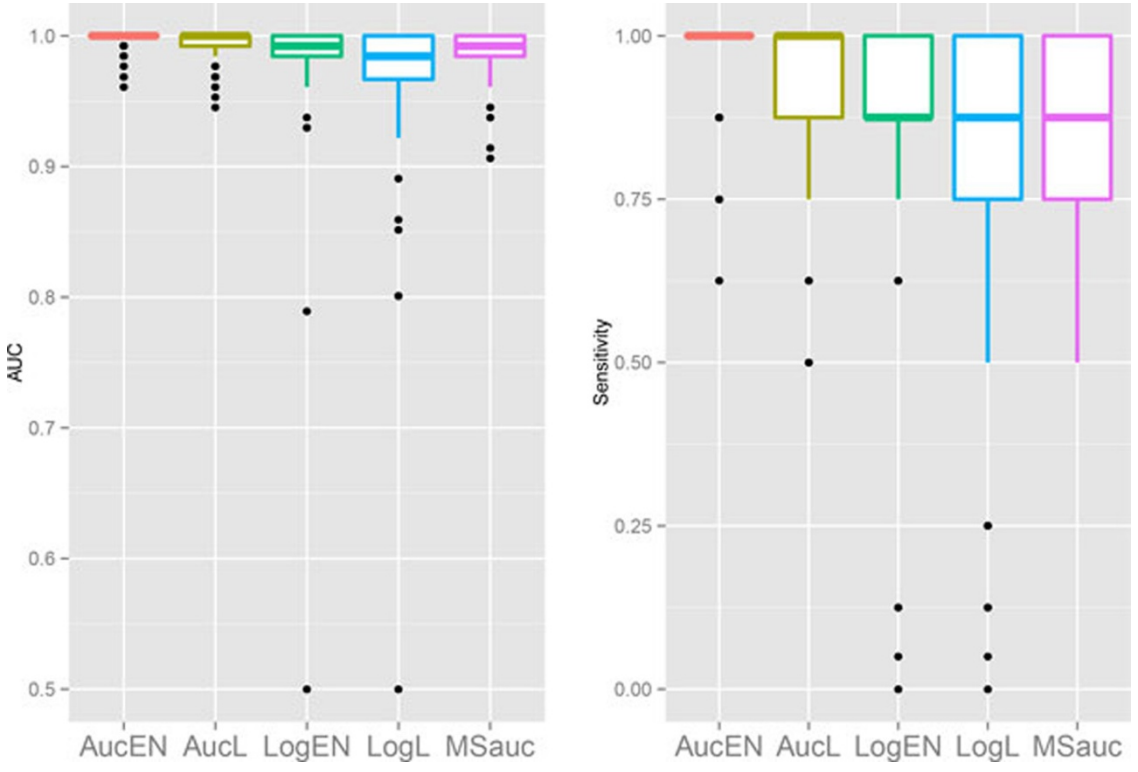

Figure 3 Box plots in *Leukemia* cancer data for AUC (left) and sensitivity when specificity equals 0.95 (right).

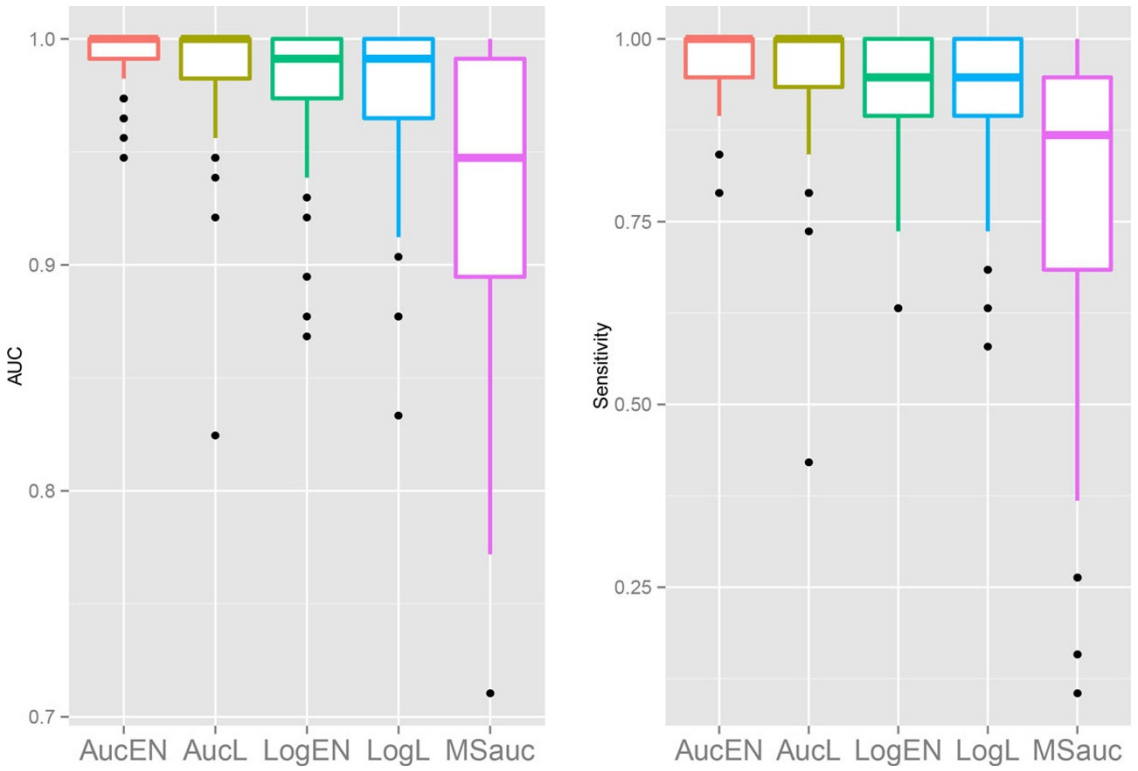

Figure 4 Box plots in *DLBCL* data for AUC (left) and sensitivity when specificity equals 0.95 (right).

the other three methods. The penalized logistic regression and *MSauc* perform poorly for the *Brain* and *Colon* cancer data. Even though the differences of AUC between these approaches are small for *Leukemia* and *DLBCL* data, the superiority of the proposed AUC-based methods becomes larger in sensitivity when specificity is as high as 0.95. This finding is very meaningful, since high sensitivity and high specificity are greatly appreciated for real cancer studies.

For the four data sets, Table 1 shows the median number of non-zero coefficients for each method. The AucPR selects more markers than the others. The approaches with the elastic net provide more genes than the lasso penalized approaches, which is consistent with the literature [10]. *MSauc* generally selects more genes than the penalized logistic regressions, but does not always give a better prediction performance than the penalized logistic regressions (Figures 1 - 4). The averaged ROC curves are plotted in Figure 5, showing that the ROC curve of the proposed AucPR lies above the curves of others, especially in *Brain* and *Colon* cancer data.

Table 1 The median number of genes being selected in four microarray studies.

	AucEN	*AucL*	*LogEN*	*LogL*	*MSauc*
Brain	42	30	37	3	16
Colon	30	22	3	2	22
Leukemia	51	36	7	5	10
DLBCL	51	35	26	9	17

In summary, the proposed AucPR selects more genes than the other three competing approaches and also have better prediction performance. Although a sparse model is good for interpretation, a better prediction performance is the primary objective and more appealing in many real world applications. As pointed out by [20], sparse models commonly ignore the correlations between the variables, which are generally inefficient even when the zero markers (or "unimportant" markers) are known in advance and all the important markers are selected correctly. It was demonstrated that those unimportant markers are in fact useful and even potentially important for classification because of the correlations. In addition, including a sufficient number of genes to the model has a practical implication; more potential important genes may be incorporated and these genes would have a higher chance for further investigation. In Table 2 the top 10 frequently selected genes by *AucEN* are listed for *Colon, Leukaemia* and *DLBCL* data sets (The gene information is not included in the preprocessed *Brain* cancer data, so we omit those results). The genes which are commonly selected by other approaches are marked.

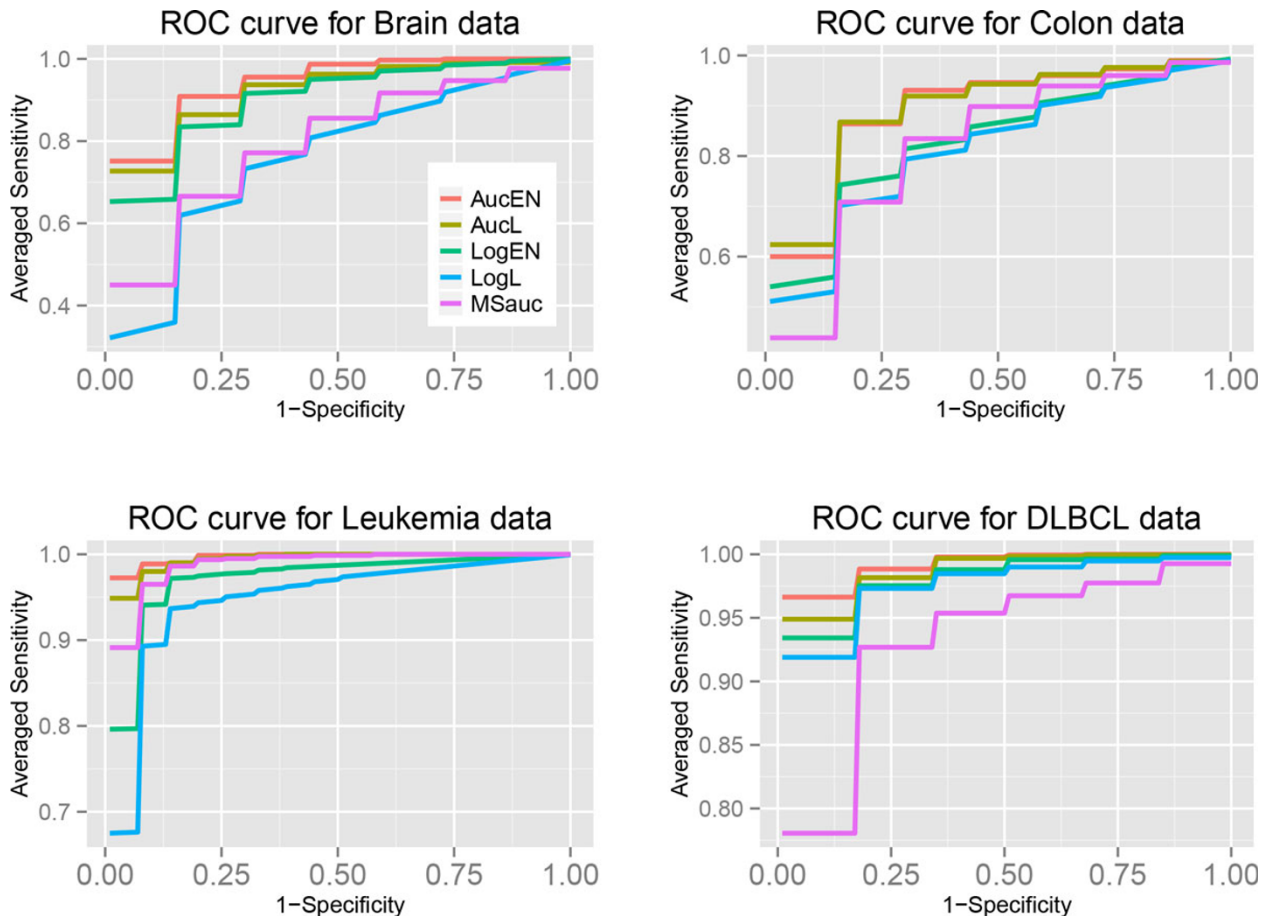

Figure 5 ROC curves for four real data sets.

Table 2 Top 10 frequently selected genes by AucEN.

data	gene id	gene symbol	description	Coverage
Colon	Hsa.2097	R14852	Human vasoactive intestinal peptide (vip) mrna, complete cds	AucL, LogL, LogEN, [35]
	Hsa.3331	T86473	Nucleoside diphosphate kinase a (Human)	AucL [36]
	Hsa.37937	R87126	Myosin heavy chain, nonmuscle (Gallus gallus)	AucL, LogL, LogEN, Msauc, [3,4]
	Hsa.601	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds	AucL, LogEN, [4]
	Hsa.36952	H43887	Complement factor d precursor (Homo sapiens)	AucL,[37]
	Hsa.8125	T71025	Human (HUMAN)	AucL, [38]
	Hsa.8147	M63391	Human desmin gene, complete cds	AucL, LogL, LogEN, Msauc, [3,4,39]
	Hsa.3306	X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1	LogL, LogEN, [4]
	Hsa.26673	R76825	RNA-specific gtpase-activating protein (Homo sapiens)	AucL, [40]
	Hsa.14069	T67077	Sodium/potassium-transporting atpase gamma chain (Ovis aries)	[41]
Leukaemia	X59711 at	NFYA	NFYA Nuclear transcription factor Y, alpha	[42]
	M30938 at	XRCC5	ATP-DEPENDENT DNA HELICASE II, 86 KD SUBUNIT	[43]
	U57721 at	Kynu	L-kynurenine hydrolase mRNA	[44]
	X07834 at	Sod2	SOD2 Superoxide dismutase 2, mitochondrial	[45]
	U37408 at	Ctbp1	CtBP mRNA	[46]
	M98539 at	ptgds	Prostaglandin D2 synthase gene	[47]

Table 2 Top 10 frequently selected genes by AucEN. *(Continued)*

	U35113 at	Mta1	Metastasis-associated mta1 mRNA	[48]
	X13973 at	rnh1	RNH Ribonuclease/angiogenin inhibitor	[49]
	D49817 at	pfkfb3	Fructose 6-phosphate,2-kinase/fructose 2,6-bisphosphatase	[50]
	M83233 at	TCF12	TCF12 Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)	LogL, LogEN, [51]
DLBCL	U96113 at	WWP1	Nedd-4-like ubiquitin-protein ligase WWP1 mRNA, partial cds	AucL, [52]
	U46006 s at	CSRP2	Smooth muscle LIM protein (h-SmLIM) mRNA	AucL, LogL, LogEN, [53]
	M35878 at	igfbp3	INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 3 PRECURSOR	AucL, [54]
	U77949 at	cdc6	Cdc6-related protein (HsCDC6) mRNA	AucL, [55]
	L41067 at	Nfatc3	Transcription factor NFATx mRNA	AucL, [56]
	U95006 at	STRA13	D9 splice variant A mRNA	[57]
	U64863 at	Pdcd1	HPD-1 (hPD-1) mRNA	AucL, [58]
	AB002409 at	ccl21	SLC	AucL, MSauc, [28]
	HG2279-HT2375 at	TPI1	Triosephosphate Isomerase	AucL
	U17969 at	eif5a	EIF5A Eukaryotic translation initiation factor 5A	[59]

The "Coverage" column shows the genes frequently selected by other methods or reported in the literature.

Gene description and related studies in the literature are shown too. The top frequently selected genes by *AucEN* were also reported in the literature.

There are several other popular approaches available for classification in high-dimensional situation. For example, the "SIS" function in package *SIS* (http://cran.r-project.org/web/packages/SIS/index.html), which first implements the Iterative Sure Independence Screening [30], and then fits the final model by penalized regression; tree-based method "randomForest" in randomForest package [31]; *LogEN* with alternative CV score ("type.measure = deviance" in *glmnet* package). As a demonstration, we implemented the third approach on the *Brain* cancer data. The result is improved and has been updated. However, Our method still outperforms others (see Figure 1).

Remark 2: Prediction accuracy and interpretation are two major concerns for microarray cancer classification study. A sparse model is generally easier to interpret but may not reflect the true biological phenomena or have poor prediction. For example, many genes are highly correlated in microarray data, and these genes may work together. Therefore, it is worthwhile to identify these genes jointly to increase prediction performance and to provide a sufficient number of potential risks for a further validation study. Note that the lasso penalized logistic regression is too parsimonious, as it cannot select a sufficient number of genes in a highly correlated group and thus, has poor prediction performance, while our *AucL* method, although with the lasso penalty, seems to be able to alleviate this problem by selecting more genes.

Simulation

In this section, we demonstrate our approaches using synthetic data under two scenarios; genes are generated from a normal distribution or a mixture of normal distributions.

We first simulate gene expressions following a setting similar to [32], where they mimicked the real microarray data. We generate data under a different number of independent blocks (*block* = 1, 2, 3), and the number of genes per block (*size*=5, 20, 40). The data are simulated from multivariate normal distributions $N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x(\rho))$ and $N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y(\rho))$ for diseased and non-diseased classes, respectively. All genes have a variance of 1, and the correlation between genes within a block is ρ (ρ = 0.3, 0.6, 0.9), whereas the correlation between genes among blocks is 0. In other words, the covariance matrix is a block-diagonal matrix

$$\sum\nolimits_x(\rho) = \sum\nolimits_y(\rho) \begin{pmatrix} A(\rho) & 0 & \cdots & 0 \\ - & A(\rho) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & A(\rho) \end{pmatrix},$$

where

$$A(\rho) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

The mean vectors are set as $\boldsymbol{\mu}_y$ = (0.6, 0.6,... , 0.6) and $\boldsymbol{\mu}_x$ = (−0.6, −0.6, ..., −0.6). Here the mean vectors are selected such that the AUC of each single gene is 0.8.

After the informative genes described above are generated, we evenly add a type of non-informative "genes" from N (0, 1) and another type of non-informative "genes" from U [−1, 1], for both diseased and non-diseased observations, and make 1000 markers in total.

We generate $n = m = 40$ i.i.d. individuals as a training set for diseased and non-diseased samples, respectively, from the above distributions. Under the same structure as the training set, another $n = m = 20$ samples are simulated independently as a testing set. Each method is applied to the training set and the prediction performance is measured on the testing set. We repeat this procedure 100 times, as we have done in the examples with real data.

For the synthetic data, the AucPR shows better prediction accuracy than the other three approaches in most scenarios. The median values of AUC, sensitivity when specificity = 0.95, the number of true informative markers being selected (*nIMS*), and the number of total markers being selected (*nTMS*), are summarized in Tables 3, 4, 5. There are some facts we can state, based on the simulation results:

1 Given ρ and the number of blocks (*block*), as the block *size* increases, our AucPR dominates the other approaches. We summarize the results when *size* = 5 and *size* = 40 in Table 3.

2 Given block *size* and the number of blocks, as ρ becomes larger, the performance of our methods do not vary much, while those of the other three methods become worse. Specifically, the sensitivities of our methods are getting larger than others when ρ is getting larger. The results for $\rho = 0.3$ and $\rho = 0.9$ are given in Table 4.

3 As the number of independent blocks increases, all methods have improved performances. When the number of blocks is 3, except *LogL* and *MSauc*, the other three methods seem to be similar in each case with *AucEN* performing slightly better (Table 5).

4 Penalized logistic regression performs better only when ρ is small (for example, 0.3) and the number of the informative genes is small. Approaches with elastic net penalty always lead to better results than the approaches with lasso penalty (Tables 3, 4, 5).

5 Generally, our AucPR approaches select more informative genes, and the approaches with elastic net penalty incorporate more informative genes than the approaches with lasso penalty (Tables 3, 4, 5). Note that as *block* and/or *size* increase (or equivalently, as the number of informative genes increases), the number of selected informative genes for our AUC-based methods increase faster, but logistics regression based approaches and *MSauc* do not. This fact may be interpreted as that our approaches show better prediction accuracy.

Table 3 Summary of simulation results for different sizes of each block, when $\rho = 0.6$ and *block* = 1 under a normal scenario. *nIMS* and *nTMS* stand for the number of the true informative markers selected and the total number of markers selected, respectively.

Size	*Method*	*Auc*	*Sensitivity*	*nIMS*	*nTMS*
5	*AucEN*	0.84	0.50	3	4
	AucL	0.82	0.45	3	4
	LogEN	0.82	0.45	2	2
	LogL	0.80	0.40	1	1
	MSauc	0.81	0.40	1	4
40	*AucEN*	0.86	0.55	20	24
	AucL	0.86	0.55	13	18
	LogEN	0.85	0.50	4	4
	LogL	0.82	0.45	2	2
	MSauc	0.81	0.45	1	3

Table 4 Summary of simulation results for different ρ, when *size* = 5 and *block* = 1, under a normal scenario.

ρ	*Method*	*Auc*	*Sensitivity*	*nIMS*	*nTMS*
0.3	*AucEN*	0.81	0.45	3	12
	AucL	0.81	0.45	3	6
	LogEN	0.85	0.50	3	3
	LogL	0.85	0.50	2	2
	MSauc	0.82	0.45	2	6
0.9	*AucEN*	0.81	0.45	3	4
	AucL	0.81	0.45	2	2
	LogEN	0.80	0.40	2	2
	LogL	0.80	0.40	1	1
	MSauc	0.79	0.40	1	3

Next, we also study the scenario where the genes are generated from a non-Gaussian setting. We simulate 50 informative genes from $0.8N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y(0.8)) + 0.2N(\mathbf{0}, \boldsymbol{I})$ and $0.8N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x(0.8)) + 0.2N(\mathbf{0}, \boldsymbol{I})$ for diseased and non-diseased groups, respectively. The non-informative genes are generated in the same way as in the first scenario. Similar patterns can be found as in the normal distribution scenario (data are not shown).

Table 5 Summary of simulation results for different *block*, when *size* = 20 and $\rho = 0.6$, under a normal scenario.

block	*Method*	*Auc*	*Sensitivity*	*nIMS*	*nTMS*
1	*AucEN*	0.86	0.55	12	14
	AucL	0.85	0.55	10	12
	LogEN	0.83	0.47	4	4
	LogL	0.81	0.40	2	2
	MSauc	0.81	0.40	1	4
3	*AucEN*	0.96	0.85	25	40
	AucL	0.95	0.85	20	32
	LogEN	0.95	0.85	16	16
	LogL	0.94	0.75	8	8
	MSauc	0.92	0.70	10	31

In summary, through selecting more genes, the proposed AucPR performs better when there are a lot of informative genes or the correlations between them are high (larger than 0.6 for example).

Remark 3: Note that the penalized logistic regressions are very powerful for marker selection in the sense that all the selected genes are the true informative genes, that is, *nIMS* = *nTMS*. For AucPR, the *nTMS* is larger than the *nIMS*, that is, there are some noisy genes selected. If the sample size increases, this phenomenon can be avoided or become negligible. Figure 6 shows that when the sample size is larger than 100, the number of noisy genes selected by *AucL* becomes very small.

Discussion

Note that, in our comparison study, the tuning parameters for all methods are tuned with an empirical (non-parametric) AUC estimator as the CV score. When sample size is very small, some difficulties may occur for calculating such AUC estimators as we did in the *brain* cancer study. Alternatively, parametric AUC estimators or the deviance from a distribution model can be used as the CV score. Different CV scores may lead to different results, especially when the sample sizes are small. It is worthy of investigating this issue as a future research topic.

Although we only use gene expression microarray data, AucPR can also be applied to other types of high-throughput omics data, such as miRNA and protein data.

AucPR methods rely on sample mean vectors and sample covariance matrices, which may not be stable enough, specifically when only a small number of samples are available. An improvement may exist in practice by replacing them with, for example, sample median and the positive-definite estimator of a large covariance matrix proposed by [33]. This can be a topic of future research.

Note that after the transformation, we try to solve a regression problem with p "samples" and p "predictors." Thus, the computation cost would grow quickly as p increases. Although screening the original p genes to a smaller number (1000 in our numerical studies) of genes is widely used and does not affect the prediction performance, as seen from our empirical study and the relevant literature [3,5,13,20], it is still worthwhile to develop fast algorithms for large scale and high-dimensional regression problem. This, too, needs further investigation.

Conclusions

We propose a powerful parametric and easily-implementable linear classifier AucPR, for gene selection and disease prediction for high-dimensional data. We transform a classical parametric AUC estimator into a linear regression and thus, the existing packages for regularized linear regression can be used directly. This novelty makes the implementation of the proposed methods very easy and efficient, since the regularized regression has been well studied. The proposed parametric method also avoids maximizing a non-concave objective function and elaborately choosing the smoothing parameter in a conventional non-parametric method. Comparisons among the AucPR, the penalized logistic regression, and a non-parametric AUC-based approach shows that our methods lead to better classifiers in the sense of predictive

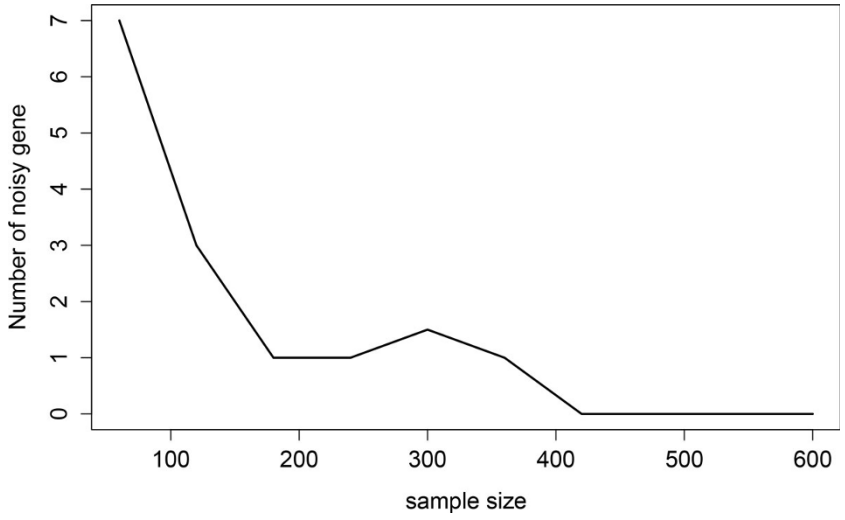

Figure 6 The number of noisy genes selected by *AucL* vs. the sample size of the simulation study, with ρ = 0.6, *block* = 2, and *size* = 20.

performance, through application to real microarray and synthetic data. In addition, the proposed AucPR selects more markers than the others and thus, could include more potential important markers for further investigation.

In addition, [34] demonstrated that the linear combination of multiple markers based on maximizing AUC generally performs better than logistic regression when the logistic model does not hold, and the two methods are comparable when the logistic model is satisfied, but their analysis was done under the condition that a very limited number of markers would be considered. This paper states that the AUC-based approach could also be advocated in high-dimensional setting, since it achieves better prediction ability than the penalized logistic regression.

Availability and supporting data

This work was implemented in R software. The R source codes are freely available at http://bibs.snu.ac.kr/software/aucpr.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WY and TP designed the method. WY performed the analyses. WY and TP interpreted the results and wrote the manuscript.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government(MSIP) (No. 2012R1A3A2026438 and 2008-0062618).

Declarations

Publication charges for this work was funded by the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIP) (No. 2012R1A3A2026438 and 2008-0062618).
This article has been published as part of *BMC Genomics* Volume 15 Supplement 10, 2014: Proceedings of the 25th International Conference on Genome Informatics (GIW/ISCB-Asia): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S10.

Authors' details

[1]Department of statistics, Seoul National University, Shilim-dong, Kwanak-gu 151-742, Seoul, Korea. [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea.

Published: 12 December 2014

References

1. Bamber D: **The area above the ordinal dominance graph and the area below the receiver operating characteristic graph.** *Journal of mathematical psychology* 1975, **12(4)**:387-415.
2. Su JQ, Liu JS: **Linear combinations of multiple diagnostic markers.** *Journal of the American Statistical Association* 1993, **88(424)**:1350-1355.
3. Ma S, Huang J: **Regularized ROC method for disease classification and biomarker selection with microarray data.** *Bioinformatics* 2005, **21(24)**:4356-4362.
4. Ma S, Song X, Huang J: **Supervised group lasso with applications to microarray data analysis.** *BMC bioinformatics* 2007, **8(1)**:60.
5. Wang Z, Yuan-chin IC, Ying Z, Zhu L, Yang Y: **A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve.** *Bioinformatics* 2007, **23(20)**:2788-2794.
6. Osamu K, Shinto E: **A boosting method for maximizing the partial area under the ROC curve.** *BMC Bioinformatics* 2010, **11.**
7. Wang Z, Chang YCI: **Marker selection via maximizing the partial area under the ROC curve of linear risk scores.** *Biostatistics* 2011, **12(2)**:369-385.
8. Hsu MJ, Hsueh HM: **The linear combinations of biomarkers which maximize the partial area under the ROC curves.** *Computational Statistics* 2013, 1-20.
9. Yu W, Chang Ycl, Park E: **A modified area under the roc curve and its application to marker selection and classification.** *Journal of the Korean Statistical Society* 2014, **43(2)**:161-175.
10. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67(2)**:301-320.
11. Tibshirani R: **Regression shrinkage and selection via the lasso.** *Journal of the Royal Statistical Society Series B (Methodological)* 1996, 267-288.
12. Ghosh D, Chinnaiyan AM: **Classification and selection of biomarkers in genomic data using lasso.** *BioMed Research International* 2005, **2005(2)**:147-154.
13. Liu Z, Jiang F, Tian G, Wang S, Sato F, Meltzer SJ, Tan M: **Sparse logistic regression with lp penalty for biomarker identification.** *Statistical Applications in Genetics and Molecular Biology* 2007, **6(1)**.
14. Schisterman E, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Armstrong D, Reiser B, Trevisan M: **Minimal and best linear combination of oxidative stress and antioxidant biomarkers to discriminate cardiovascular disease.** *Nutrition, metabolism, and cardiovascular diseases: NMCD* 2002, **12(5)**:259-266.
15. Weber F, Shen L, Aldred MA, Morrison CD, Frilling A, Saji M, Schuppert F, Broelsch CE, Ringel MD, Eng C: **Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination.** *Journal of Clinical Endocrinology & Metabolism* 2005, **90(5)**:2512-2521.
16. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M: **Assessing the limits of genomic data integration for predicting protein networks.** *Genome research* 2005, **15(7)**:945-953.
17. Attallah AM, Mosa TE, Omran MM, Abo-Zeid MM, El-Dosoky I, Shaker YM: **Immunodetection of collagen types i, ii, iii, and iv for differentiation of liver fibrosis stages in patients with chronic hcv.** *Journal of Immunoassay & Immunochemistry* 2007, **28(2)**:155-168.
18. Zhao P, Yu B: **On model selection consistency of lasso.** *The Journal of Machine Learning Research* 2006, **7**:2541-2563.
19. Jia J, Yu B: **On model selection consistency of the elastic net when p¿¿ n.** *Technical report, DTIC Document* 2008.
20. Cai T, Liu W: **A direct estimation approach to sparse linear discriminant analysis.** *Journal of the American Statistical Association* 2011, **106(496)**.
21. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *The Annals of statistics* 2004, **32(2)**:407-499.
22. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *Journal of statistical software* 2010, **33(1)**:1.
23. Ayers KL, Cordell HJ: **Snp selection in genome-wide and candidate gene studies via penalized logistic regression.** *Genetic epidemiology* 2010, **34(8)**:879-891.
24. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression.** *Bioinformatics* 2009, **25(6)**:714-721.
25. Dettling M: **Bagboosting for tumor classification with gene expression data.** *Bioinformatics* 2004, **20(18)**:3583-3593.
26. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proceedings of the National Academy of Sciences* 1999, **96(12)**:6745-6750.
27. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *science* 1999, **286(5439)**:531-537.
28. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, *et al*: **Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nature medicine* 2002, **8(1)**:68-74.

29. Smyth GK, *et al*: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Statistical applications in genetics and molecular biology* 2004, **3(1)**:3.
30. Fan J, Lv J: **Sure independence screening for ultrahigh dimensional feature space.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2008, **70(5)**:849-911.
31. Liaw A, Wiener M: **Classification and regression by randomforest.** *R news* 2002, **2(3)**:18-22.
32. Díaz-Uriarte R, De Andres SA: **Gene selection and classification of microarray data using random forest.** *BMC bioinformatics* 2006, **7(1)**:3.
33. Xue L, Ma S, Zou H: **Positive-definite 1-penalized estimation of large covariance matrices.** *Journal of the American Statistical Association* 2012, **107(500)**:1480-1491.
34. Pepe MS, Cai T, Longton G: **Combining predictors for classification using the area under the receiver operating characteristic curve.** *Biometrics* 2006, **62(1)**:221-229.
35. Jabari S, da Silveira AB, de Oliveira EC, Quint K, Wirries A, Neuhuber W, Brehmer A: **Mucosal layers and related nerve fibres in non-chagasic and chagasic human colona quantitative immunohistochemical study.** *Cell and tissue research* 2014, 1-9.
36. Álvarez-Chaver P, Rodríguez-Piñeiro AM, Rodríguez-Berrocal FJ, García-Lorenzo A, Páez de la Cadena M, Martínez-Zorzano VS: **Selection of putative colorectal cancer markers by applying pca on the soluble proteome of tumors: Ndk a as a promising candidate.** *Journal of proteomics* 2011, **74(6)**:874-886.
37. Nambiar PR, Gupta RR, Misra V: **An omics based survey of human colon cancer.** *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2010, **693(1)**:3-18.
38. Xq Z, Zhang F, Tao Y, Cm W, Sz L, Fl H, *et al*: **Expression profiling based on graph-clustering approach to determine colon cancer pathway.** *Journal of cancer research and therapeutics* 2013, **9(3)**:467.
39. Jiang W, Li X, Rao S, Wang L, Du L, Li C, Wu C, Wang H, Wang Y, Yang B: **Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements.** *BMC systems biology* 2008, **2(1)**:72.
40. Tabuchi Y, Takasaki I, Doi T, Ishii Y, Sakai H, Kondo T: **Genetic networks responsive to sodium butyrate in colonic epithelial cells.** *FEBS letters* 2006, **580(13)**:3035-3041.
41. Floyd RV, Wray S, Martín-Vasallo P, Mobasheri A: **Differential cellular expression of fxyd1 (phospholemman) and fxyd2 (gamma subunit of na, k-atpase) in normal human tissues: a study using high density human tissue microarrays.** *Annals of Anatomy-Anatomischer Anzeiger* 2010, **192(1)**:7-16.
42. Samet I, Han J, Jlaiel L, Sayadi S, Isoda H: **Olive (olea europaea) leaf extract induces apoptosis and monocyte/macrophage differentiation in human chronic myelogenous leukemia k562 cells: Insight into the underlying mechanism.** *Oxidative medicine and cellular longevity* 2014, **2014.**
43. Cierniewski CS, Papiewska-Pajak I, Malinowski M, Sacewicz-Hofman I, Wiktorska M, Kryczka J, Wysocki T, Niewiarowska J, Bednarek R: **Thymosin β4 regulates migration of colon cancer cells by a pathway involving interaction with ku80.** *Annals of the New York Academy of Sciences* 2010, **1194(1)**:60-71.
44. Damm F, Thol F, Hollink I, Zimmermann M, Reinhardt K, van den Heuvel-Eibrink M, Zwaan CM, de Haas V, Creutzig U, Klusmann J: **Prevalence and prognostic value of idh1 and idh2 mutations in childhood aml: a study of the aml-bfm and dcog study groups.** *Leukemia* 2011, **25(11)**:1704-1710.
45. Zgheib C, Zouein FA, Kurdi M, Booz GW: **Chronic treatment of mice with leukemia inhibitory factor does not cause adverse cardiac remodeling but improves heart function.** *European cytokine network* 2012, **23(4)**:191-197.
46. Perry C, Pick M, Podoly E, Gilboa-Geffen A, Zimmerman G, Sklan E, Ben-Shaul Y, Diamant S, Soreq H: **Acetylcholinesterase/c terminal binding protein interactions modify ikaros functions, causing t lymphopenia.** *Leukemia* 2007, **21(7)**:1472-1480.
47. Sasaki H, Nishikata I, Shiraga T, Akamatsu E, Fukami T, Hidaka T, Kubuki Y, Okayama A, Hamada K, Okabe H: **Overexpression of a cell adhesion molecule, tslc1, as a possible molecular marker for acute-type adult t-cell leukemia.** *Blood* 2005, **105(3)**:1204-1213.
48. Toh Y, Nicolson GL: **The role of the mta family and their encoded proteins in human cancers: molecular functions and clinical implications.** *Clinical & experimental metastasis* 2009, **26(3)**:215-227.
49. Guan X, Yang J, Zhu N, Wang Y, Li R, Zheng Z: **[gene expression differences between high and low metastatic cells of adenoid cystic carcinoma].** *Zhonghua kou qiang yi xue za zhi= Zhonghua kouqiang yixue zazhi= Chinese journal of stomatology* 2004, **39(2)**:118-121.
50. Carlet M, Janjetovic K, Rainer J, Schmidt S, Panzer-Grümayer R, Mann G, Prelog M, Meister B, Ploner C, Kofler R: **Expression, regulation and function of phosphofructo-kinase/fructose-biphosphatases (pfkfbs) in glucocorticoid-induced apoptosis of acute lymphoblastic leukemia cells.** *BMC cancer* 2010, **10(1)**:638.
51. Meyer C, Kowarz E, Yip SF, Wan TSK, Chan TK, Dingermann T, Chan LC, Marschalek R: **A complex¡ i¿ mll¡/i¿ rearrangement identified five years after initial mds diagnosis results in out-of-frame fusions without progression to acute leukemia.** *Cancer genetics* 2011, **204(10)**:557-562.
52. Chen C, Zhou Z, Ross JS, Zhou W, Dong JT: **The amplified wwp1 gene is a potential molecular target in breast cancer.** *International journal of cancer* 2007, **121(1)**:80-87.
53. Zangrando A, Dell'Orto MC, te Kronnie G, Basso G: **Mll rearrangements in pediatric acute lymphoblastic and myeloblastic leukemias: Mll specific and lineage specific signatures.** *BMC medical genomics* 2009, **2(1)**:36.
54. Sung CO, Kim SC, Karnan S, Karube K, Shin HJ, Nam DH, Suh YL, Kim SH, Kim JY, Kim SJ, *et al*: **Genomic profiling combined with gene expression profiling in primary central nervous system lymphoma.** *Blood* 2011, **117(4)**:1291-1300.
55. Delmolino LM, Saha P, Dutta A: **Multiple mechanisms regulate subcellular localization of human cdc6.** *Journal of Biological Chemistry* 2001, **276(29)**:26947-26954.
56. Glud SZ, Sørensen AB, Andrulis M, Wang B, Kondo E, Jessen R, Krenacs L, Stelkovics E, Wabl M, Sering E, *et al*: **A tumor-suppressor function for nfatc3 in t-cell lymphomagenesis by murine leukemia virus.** *Blood* 2005, **106(10)**:3546-3552.
57. Seimiya M, Bahar R, Wang Y, Kawamura K, Tada Y, Okada S, Hatano M, Tokuhisa T, Saisho H, Watanabe T, *et al*: **Clast5/stra13 is a negative regulator of b lymphocyte activation.** *Biochemical and biophysical research communications* 2002, **292(1)**:121-127.
58. de Leval L, Rickman DS, Thielen C, de Reynies A, Huang YL, Delsol G, Lamant L, Leroy K, Brière J, Molina T, *et al*: **The gene expression profile of nodal peripheral t-cell lymphoma demonstrates a molecular link between angioimmunoblastic t-cell lymphoma (aitl) and follicular helper t (tfh) cells.** *Blood* 2007, **109(11)**:4952-4963.
59. Lin YW, Aplan PD: **Gene expression profiling of precursor t-cell lymphoblastic leukemia/lymphoma identifies oncogenic pathways that are potential therapeutic targets.** *Leukemia* 2007, **21(6)**:1276-1284.

doi:10.1186/1471-2164-15-S10-S1
Cite this article as: Yu and Park: AucPR: An AUC-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC Genomics* 2014 **15**(Suppl 10):S1.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central