Maximization of the Partial Area under the ROC curve using a Boosting Technique

Osamu Komori, Prediction and Knowledge Discovery Research Center
komori@ism.ac.jp

Introduction

- This research attempts to improve the accuracy of discrimination of the diseased group from non-diseased group. The accuracy is often measured by the partial area under the ROC curve (pAUC)[3].
- The association between markers and outcome variable (disease or non-disease) is also important in medical or biological sciences.
- We have developed a new statistical method that aims to maximize the pAUC, using a boosting technique.
- The resultant score plots show us the association in a visually-apparent way.

Status quo

- does *not* take into account the non-linear structure of the association between markers and outcome variable[5];[6].
- Moreover, Eguchi and Copas [1] and McIntosh and Pepe [4] show the optimal score function is derived from the likelihood ratio. That is, the linearity of the association is *not* sufficient in general.

ROC curve and pAUC

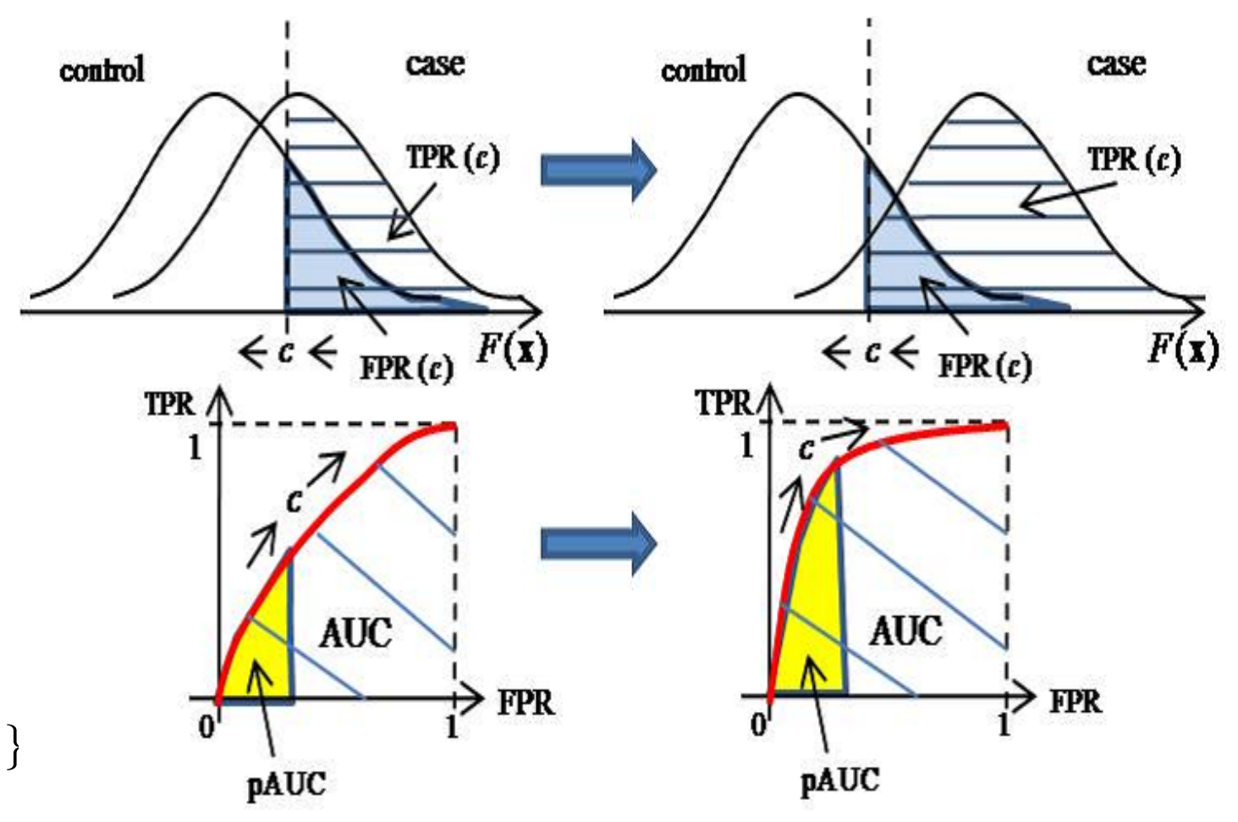
- $(\boldsymbol{x}, y)$, $\boldsymbol{x} \in \mathbf{R}^p,\ y \in \{0,\ 1\}$
- $F(\boldsymbol{x})$: score function,

  c : threshold

$$\begin{cases} F(\boldsymbol{x}) \geq c \Rightarrow \text{positive} \\ F(\boldsymbol{x}) < c \Rightarrow \text{negative} \end{cases}$$

$$\Downarrow$$

$$\mathrm{ROC}(F) = \{\mathrm{FPR}(c),\ \mathrm{TPR}(c)\}$$



⋆ The pAUC is more suitable for clinical setting in which a high true positive rate is required with a low false positive rate.

Theorem about pAUC

Theorem 1. For a pair of fixed α_1 and α_2, let

$$\Psi(\gamma) = \mathrm{pAUC}_\sigma\Big(F + \gamma\, m\big(\Lambda\big), \alpha_1, \alpha_2\Big),$$

where γ is a scalar, $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ and m is a strictly increasing function. Then, $\Psi(\gamma)$ is strictly increasing function of γ, and

$$\sup_F \mathrm{pAUC}_\sigma(F, \alpha_1, \alpha_2) = \lim_{\gamma \to \infty} \Psi(\gamma) = \mathrm{pAUC}\big(\Lambda, \alpha_1, \alpha_2\big).$$

- As seen in Theorem 1, the approximate pAUC has no maximum, but a supremum. Hence, we will consider the penalty term in the objective function in order to ensure the existence of the maximum and make the pAUCBoost algorithm numerically stable.

Objective function

- Prepare a set of weak classifiers, from which we construct a score function $F(\boldsymbol{x})$.

$$\mathcal{F} = \{f(\boldsymbol{x}) = N_{k,l}(x_k)/Z_{k,l} |\ k = 1, 2, \ldots, p,\ l = 1, 2, \ldots, m_k\}.$$

The basis functions of the natural cubic spline for x_k are defined as

$$N_{k,l}(x_k) = \begin{cases} 1,\ l = 1, \\ x_k,\ l = 2, \\ d_{l-2}(x_k) - d_{m_k-1}(x_k),\ \text{otherwise}, \end{cases}$$

where

$$d_l(x_k) = \frac{(x_k - \xi_{k,l-2})^3_+ - (x_k - \xi_{k,m_k})^3_+}{\xi_{k,m_k} - \xi_{k,l-2}},$$

and z_+ denotes the positive part of z. The standardization factor $Z_{k,l}$ for $N_{k,l}(x_k)$ is given as

$$Z_{k,l} = \begin{cases} 1,\ l = 1, \\ \xi_{k,m_k} - \xi_{k,1},\ l = 2, \\ N_{k,l}(\xi_{k,m_k}) - N_{k,l}(\xi_{k,l-2}),\ \text{otherwise}, \end{cases}$$

and $\xi_{k,l}$ is one of m_k knots ($\xi_{k,1} < \xi_{k,2} < \ldots < \xi_{k,m_k}$) for x_k.

- the objective function we propose is

$$\begin{aligned} \overline{\mathrm{pAUC}}_{\sigma,\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) &= \overline{\mathrm{pAUC}}_\sigma(F, \overline{\alpha}_1, \overline{\alpha}_2) - \lambda \sum_{k=1}^{p} \int \{F_k''(x_k)\}^2 dx_k \\ &= \frac{1}{n_0 n_1} \sum_{i \in I} \Big\{ \sum_{j \in J_{\mathrm{fan}}} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \\ &\quad + \sum_{j \in J_{\mathrm{rec}}} \mathrm{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \Big\} \\ &\quad - \lambda \sum_{k=1}^{p} \int \{F_k''(x_k)\}^2 dx_k, \end{aligned}$$

where $F_k''(x_k)$ is the second derivative of the k-th component of $F(\boldsymbol{x})$, and λ is a smoothing parameter that controls the smoothness of $F(\boldsymbol{x})$.

- Without loss of generality, we can rewrite it as

$$\begin{aligned} \overline{\mathrm{pAUC}}_{\sigma,\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) &= \overline{\mathrm{pAUC}}_{1,\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) \\ &\equiv \overline{\mathrm{pAUC}}_{\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) \end{aligned}$$

- Note that the maximizer of the objective function is shown to be the natural cubic splines [2].

pAUCBoost algorithm

```
1. Start with a score function $F_0(\boldsymbol{x}) = 0$ and set each coefficient $\beta_0(f)$ of weak classifiers to be $1$ or $-1$.
2. For $t = 1, ..., T$
   a. Calculate the values of thresholds $\bar{c}_1$ and $\bar{c}_2$ for each $F_{t-1} + \beta_{t-1}(f)f$.
   b. Update $\beta_{t-1}(f)$ to $\beta_t(f)$ with a one-step Newton-Raphson iteration.
   c. Find the best weak classifier $f_t$
        $f_t = \underset{f}{\mathrm{argmax}}\ \overline{\mathrm{pAUC}}_\lambda(F_{t-1} + \beta_t(f)f, \overline{\alpha}_1, \overline{\alpha}_2)$
   d. Update the score function as
        $F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \beta_t(f_t)f_t(\boldsymbol{x})$.
3. Finally, output a final score function $F(\boldsymbol{x}) = \sum_{t=1}^{T} \beta_t(f_t)f_t(\boldsymbol{x})$.
```

Breast cancer data

- Two types of data [7]

  clinical data: Age, Size, Grade, Angi, ERp, PRp and Lymp

  genomic data: gene expression profiles (25000 genes)
- training data: 78 patients; test data 19 patients
- We apply AUCBoost to clinical data using natural cubic splines to Age and Size (continuous markers), and decision stumps to the others (discrete markers)

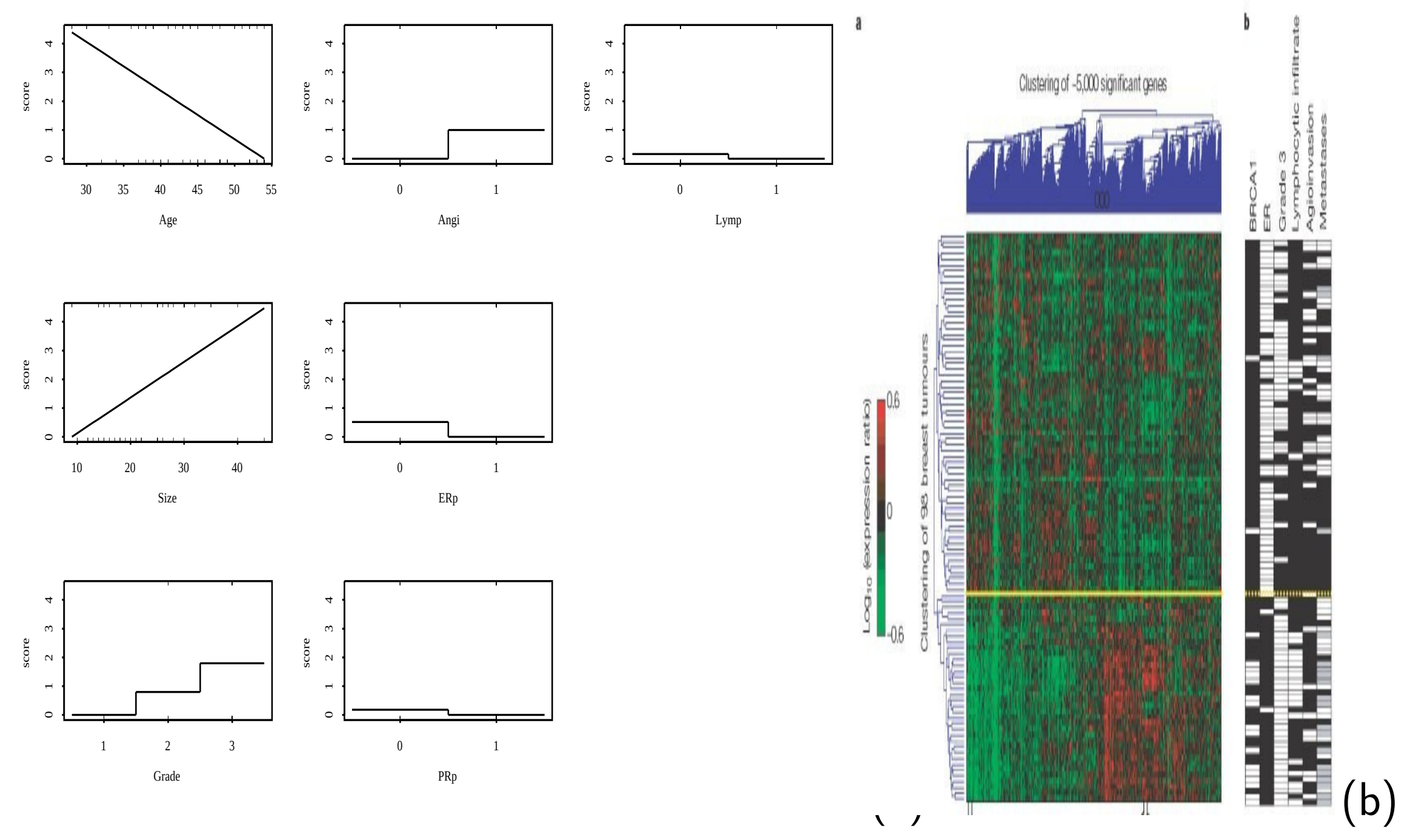


Figure 1. (a) Score plots of clinical markers; (b) Clustering cited from [7].

- The resultant AUC is 0.882 and 0.869 for training and test data, resp.
- After the pAUC-based filtering process, we apply pAUCBoost with natural cubic splines to the 11 genes.

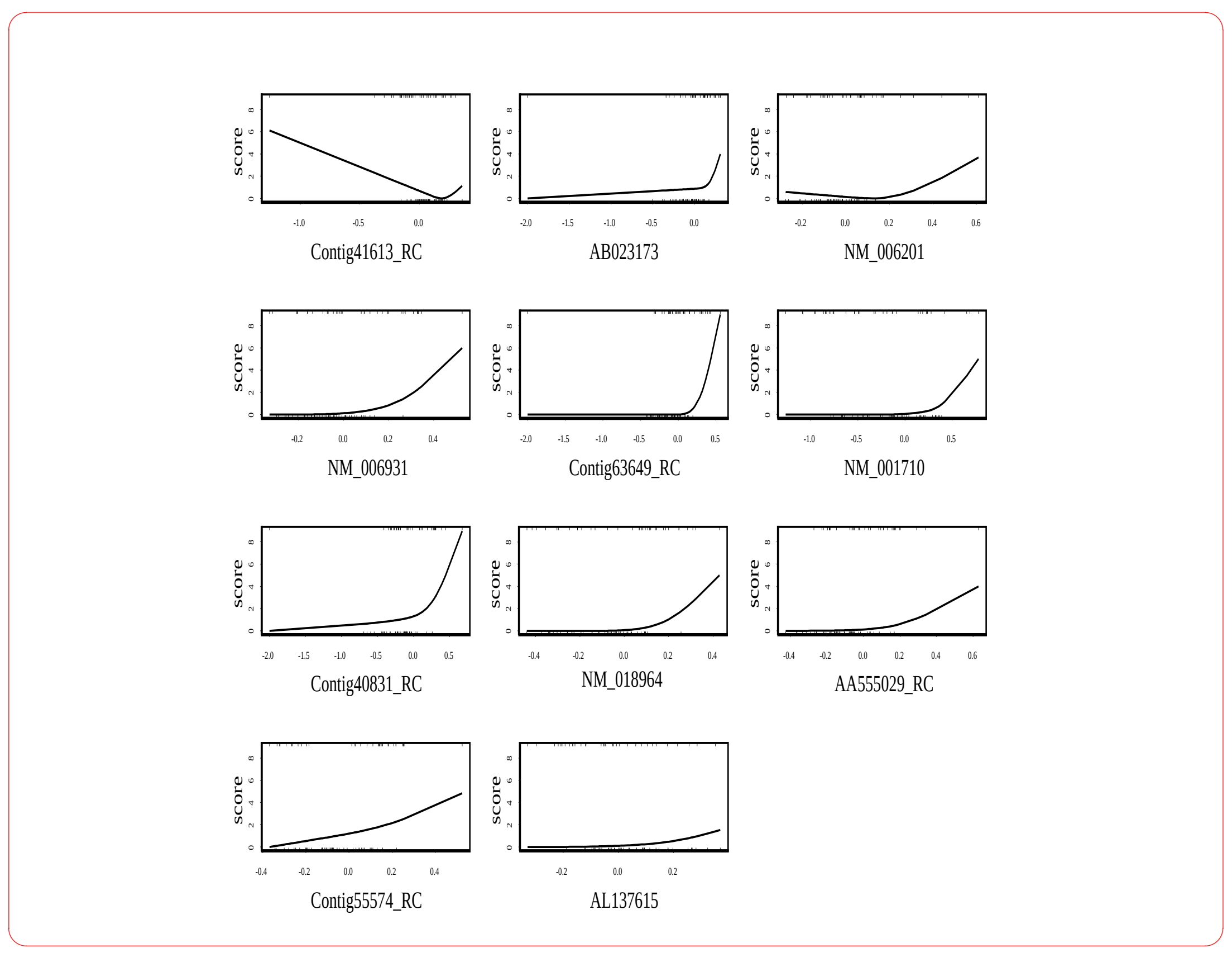
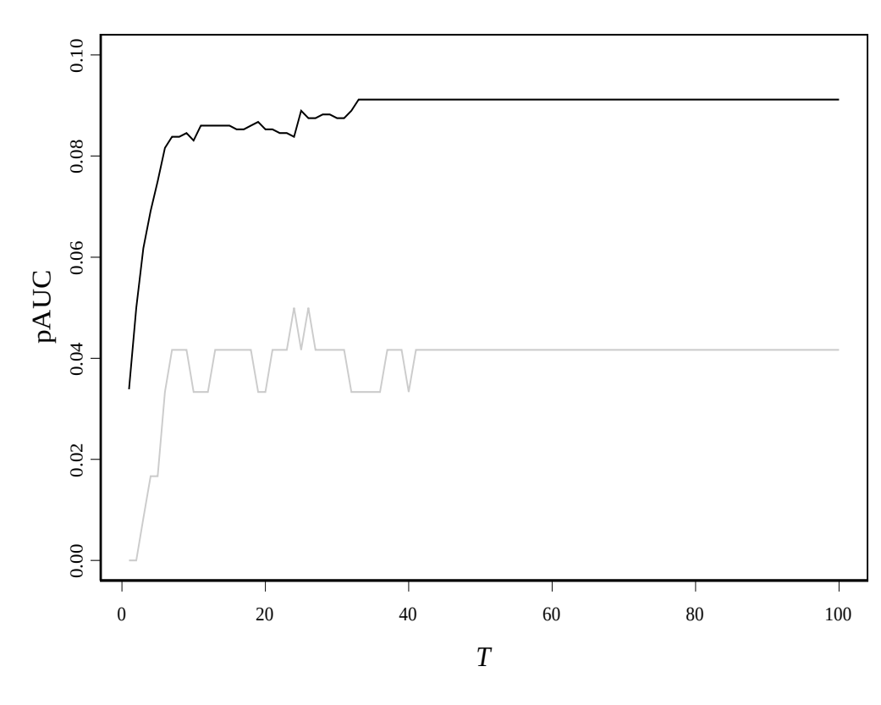


Figure 2. Score plots of the selected 11 genes.

- Results



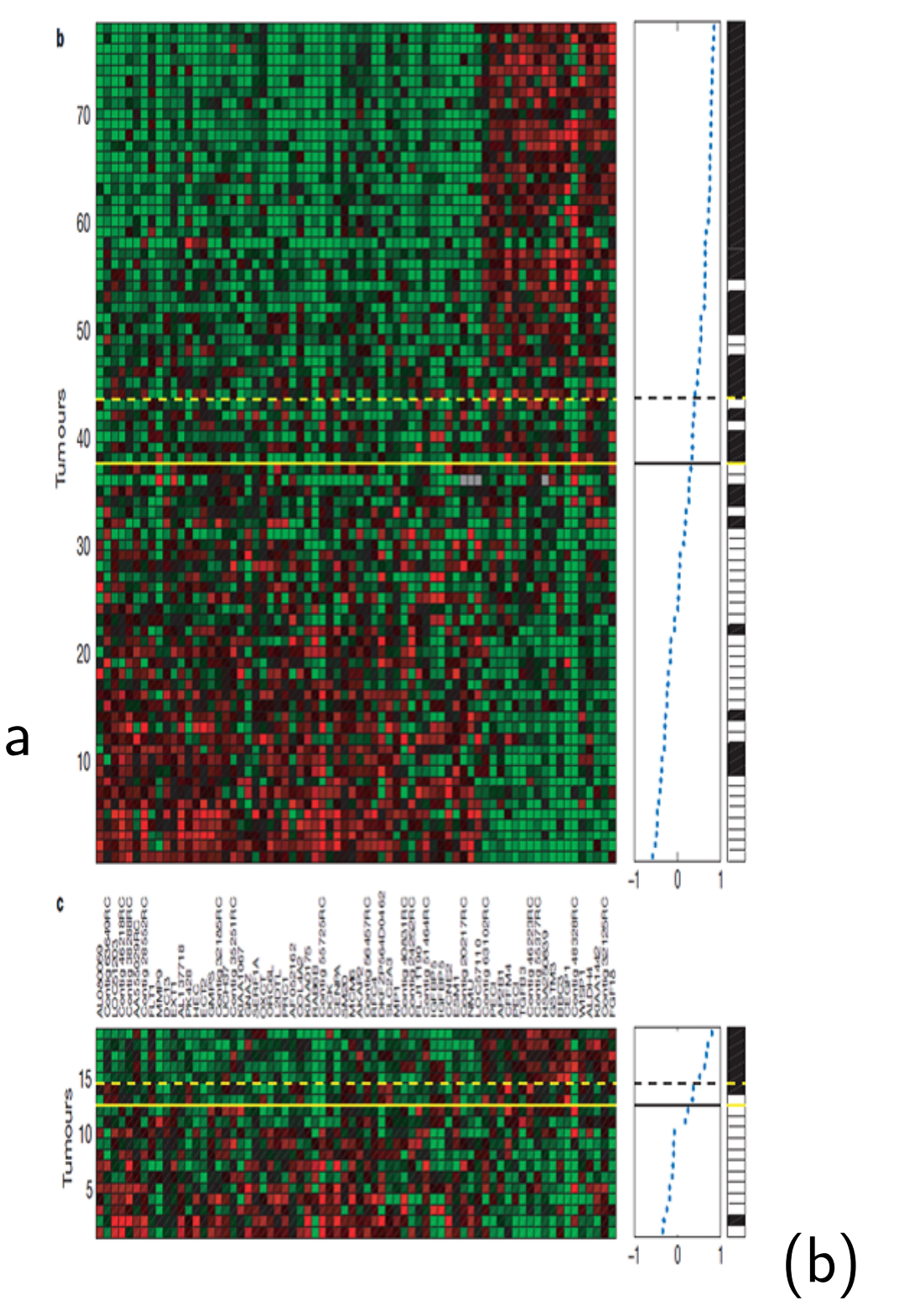


Figure 3. (a) The pAUC for training data (black) and test data (gray); (b) the results of van't Veer [7] .

- The resultant pAUCs for both training and test data are more than 3 times bigger than their results: 0.025 and 0.0008, resp. [7].

References

[1] Eguchi, S. and Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika* **89**, 1–22.

[2] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

[3] Komori, O. and Eguchi, S. (2010). A boosting method for maximizing the partial area under the ROC Curve. *BMC Bioinformatics* **11**, 314.

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

[5] Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

[6] Pepe, M. S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the Receiver Operating Characteristic curve. *Biometrics* **62**, 221–229.

[7] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

The Institute of Statistical Mathematics