Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1643–1655
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data

Yongdai Kim[a,*], Sunghoon Kwon[a], Seuck Heun Song[b]

[a] *Seoul National University, Korea*
[b] *Korea University, Korea*

Received 22 March 2006; received in revised form 23 May 2006; accepted 5 June 2006
Available online 30 June 2006

**Abstract**

Monitoring gene expression profiles is a novel approach to cancer diagnosis. Several studies have showed that the sparse logistic regression is a useful classification method for gene expression data. Not only does it give a sparse solution with high accuracy, it provides the user with explicit probabilities of classification apart from the class information. However, its optimal extension to more than two classes is not obvious. In this paper, we propose a multiclass extension of sparse logistic regression. Analysis of five publicly available gene expression data sets shows that the proposed method outperforms the standard multinomial logistic model in prediction accuracy as well as gene selectivity.


*Keywords:* Classification; Gene expression data; Multinomial logit model; One-against-all; Sparse logistic regression

## 1. Introduction

Constructing a classification rule for tissue samples based on gene expression profiles has received much attention recently due to emerging microarray technology. A new challenge is that the number of genes (i.e. the dimension of inputs) is much larger than the number of tissue samples, in which case standard classification methods either are not applicable or perform badly. Also, identifying a small subset of informative genes, called marker genes, which discriminate types of tumors or tumor versus normal tissues, has become an important subject. Hence, good learning algorithms with gene expression data should provide a classification rule which not only yields high accuracy but also has the ability to identify marker genes. In related literature, Guyon et al. (2002) proposed a recursive feature elimination technique with support vector machines, Li et al. (2002) introduced two Bayesian approaches with the technique of automatic relevance determination, and Shevade and Keerthi (2003) and Roth (2002) applied the sparse logistic regression, to name just a few.

* Corresponding author.
*E-mail address:* ydkim@stats.snu.ac.kr (Y. Kim).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.007

Among these tools, sparse logistic regression is a useful classification method for gene expression data. It gives a sparse solution with high accuracy and also it provides the user with explicit probabilities of classification apart from the class information. However, its optimal extension to more than two classes is not obvious. A standard multiclass extension of sparse logistic regression might be sparse multinomial logistic (SML) regression (Krishnapuram et al., 2004), which is a sparse version of the multinomial logit model—a popular multiclass formulation in statistics (see, for example, Agresti, 1990). SML, however, has a problem in gene selection. Since the estimates of the regression coefficients depend on the choice of the baseline class (see Section 2 for definition), and so do the selected genes. Hence, some important genes are dropped in the final model, which in turn degrades the prediction accuracies. Empirical results in Section 4 confirms this observation.

In this paper, we propose a new multiclass extension of sparse logistic regression called *sparse one-against-all logistic* (*SOVAL*) *regression*, whose main idea is to reduce a multiclass problem to multiple binary problems and to construct a classifier using the reduced multiple binary problems simultaneously. By analyzing five real data sets of gene expressions, we show that SOVAL outperforms SML in prediction accuracy as well as gene selectivity.

The paper is organized as follows. In Section 2, SOVAL as well as SML are presented. A computational algorithm based on the gradient LASSO algorithm of Kim et al. (2005) is given in Section 3. Results of numerical experiments are presented in Section 4 and concluding remarks follow in Section 5.

## 2. Models

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be input–output pairs of a given data set where $\mathbf{x}_i \in R^p$ is a gene expression level and $y_i \in \{1, 2, \ldots, J\}$ is a type of cancer of the $i$th tissue sample. Here, $n$ is the number of tissues, $p$ the number of genes and $J$ the number of classes (i.e. tumor types). We first present SML and then propose SOVAL.

### 2.1. SML regression

SML starts with the multinomial logit model

$$\Pr(y_i = j|\mathbf{x}_i) = \frac{\exp\left(f_j(\mathbf{x}_i)\right)}{\sum_{m=1}^{J} \exp(f_m(\mathbf{x}_i))}$$

for $j = 1, \ldots, J$ where

$$f_j(\mathbf{x}_i) = \beta_0^{(j)} + \beta_1^{(j)} x_{i1} + \cdots + \beta_p^{(j)} x_{ip}.$$

For identifiability, we let $\beta_k^{(J)} = 0$ for $k = 0, 1, \ldots, p$.

Let $\boldsymbol{\beta}_0 = \left(\beta_0^{(1)}, \ldots, \beta_0^{(J-1)}\right)$, $\boldsymbol{\beta}_j = \left(\beta_1^{(j)}, \ldots, \beta_p^{(j)}\right)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1})$. For the sparse model, we estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ by maximizing the log-likelihood

$$\mathscr{L}_1(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[ \sum_{j=1}^{J} I(y_i = j) f_j(\mathbf{x}_i) - \log\left( \sum_{m=1}^{J} \exp(f_m(\mathbf{x}_i)) \right) \right] \tag{1}$$

with the constraint $\sum_{j=1}^{J-1} \sum_{k=1}^{p} \left|\beta_k^{(j)}\right| \leqslant \lambda$. Here, $\lambda > 0$ is a regularization parameter, which should be selected in advance using cross validation or any other method.

Once the regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ are estimated, the classifier is constructed as follows. Let $c(i|j)$ be the cost of classifying an observation to the $i$th class when the true class is $j$. Then, a new tissue sample with gene expression $\mathbf{x}$ is classified into class $C(\mathbf{x})$ where

$$C(\mathbf{x}) = \arg\min_j \sum_{i=1}^{J} c(i|j) \Pr(y = j|\mathbf{x}).$$

If $c(i|j)$ are all equal, which is most frequent in practice, $C(\mathbf{x})$ becomes $\arg\max_j \Pr(y = j|\mathbf{x})$.

The importance of the $k$th gene for classification of tumor types is measured by $\rho_k$ where

$$\rho_k = \sum_{j=1}^{J-1} \left|\beta_k^{(j)}\right|.$$

The larger $\rho_k$ is, the more important the $k$th gene is for classifying the tumor type and so genes with sufficiently large $\rho_k$ can be considered as marker genes. Using $\rho_k$, we can reformulate SML as

$$f_j(\mathbf{x}_i) = \theta_0^{(j)} + \rho_1\theta_1^{(j)}x_{i1} + \cdots + \rho_p\theta_p^{(j)}x_{ip}$$

with $\sum_{j=1}^{J-1}\left|\theta_k^{(j)}\right| = 1$, $\rho_k \geqslant 0$ for $k = 1, \ldots, p$ and $\sum_{k=1}^{p}\rho_k \leqslant \lambda$. Hence, SML can be considered as a garrot type estimate (Breiman, 1995) for $\rho_k$, and so we expect that the solution of $\rho_k$ is sparse.

In SML, we set $\beta_k^{(J)} = 0$ for $k = 1, \ldots, p$ for identifiability of the model, and the regression coefficient $\beta_k^{(j)}$, $j \neq J$ can be interpreted as the log odds ratio of the $j$th group versus the $J$th group for the $k$th gene. In this sense, we call the $J$th class the baseline class. This convention has a problem that the estimates depends on the choice of the baseline class. For an example, consider the following simple situation. Let $p = 1$, $J = 3$ and $\lambda = 1$. Suppose $\mathbf{x}_1$ is binary (i.e. $\mathbf{x}_1 \in \{0, 1\}$). Let $Odd(k, j)$ be the odds ratio of the $k$th group versus the $j$th group. That is,

$$Odd(k, j) = \frac{\sum_{i=1}^{n} I(y_i = k, x_{1i} = 1)\sum_{i=1}^{n} I(y_i = j, x_{1i} = 0)}{\sum_{i=1}^{n} I(y_i = k, x_{1i} = 0)\sum_{i=1}^{n} I(y_i = j, x_{1i} = 1)}.$$

Suppose $\log Odd(1|3) = 0.5$ and $\log Odd(2|3) = -0.5$. Then, the estimates of the regression coefficients from SML become $\beta_1^{(1)} = 0.5$ and $\beta_1^{(2)} = -0.5$ if we choose the third class as the baseline class. Now, suppose we change the baseline class to the second class. Then since $\log Odd(1|2) = 1.0$ and $\log Odd(3|2) = 0.5$, in order for the class probabilities to remain the same, the estimates of $\beta_1^{(1)}$ and $\beta_1^{(3)}$ have to be 1.0 and 0.5, respectively, which is impossible since it violates the constraint (i.e. $\left|\beta_1^{(1)}\right| + \left|\beta_1^{(3)}\right| > 1$). Hence, there is a danger that some important genes may be dropped in the final model due to the choice of the baseline class, which results in poor prediction accuracy. Empirical results in Section 4 confirms this observation.

Instead of choosing the baseline class, there are other ways to resolve the identification problem. An example is to let

$$\sum_{j=1}^{J} \beta_k^{(j)} = 0 \tag{2}$$

for all $k$. This constraint, however, makes the computation harder. A main technical difficulty of sparse logistic regression is that computation is relatively demanding. This is mainly because the objective function to be optimized is not differentiable due to $L_1$ constraint, and hence special optimization techniques are required. Within the authors' knowledge, there is no special optimization algorithm for sparse logistic regression which can deal with the constraint (2), in particular for large number of genes.

### 2.2. Sparse one-against-all logistic regression

For given $y_i$, the standard one-against-all (OVA) approach makes $J$ many binary outputs $y_i^{(1)}, \ldots, y_i^{(J)}$ via $y_i^{(j)} = I(y_i = j)$, and assumes

$$\Pr\left(y_i^{(j)} = 1 \middle| \mathbf{x}_i\right) = \frac{\exp\left(f_j(\mathbf{x}_i)\right)}{1 + \exp\left(f_j(\mathbf{x}_i)\right)}$$

for $j = 1, \ldots, J$ where

$$f_j(\mathbf{x}) = \beta_0^{(j)} + \beta_1^{(j)} x_1 + \cdots + \beta_p^{(j)} x_p.$$

Let $\boldsymbol{\beta}_0 = \left(\beta_0^{(1)}, \ldots, \beta_0^{(J)}\right)$, $\boldsymbol{\beta}_j = \left(\beta_1^{(j)}, \ldots, \beta_p^{(j)}\right)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$. Then, it estimates $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ by estimating $\beta_0^{(j)}$ and $\boldsymbol{\beta}_j$ for $j = 1, \ldots, J$ via maximizing the log-likelihood of $y_i^{(j)}$ given by

$$\sum_{i=1}^{n} \left[ y_i^{(j)} f_j(\mathbf{x}_i) - \log\left(\exp\left(f_j(\mathbf{x}_i)\right) + 1\right)\right]$$

subject to $\sum_{k=1}^{p} \left|\beta_k^{(j)}\right| \leqslant \lambda_j$. There are multiple regularization parameters $\lambda_1, \ldots, \lambda_J$, which should be selected simultaneously in advance using cross validation or any other method. Note that selecting multiple regularization parameters is computationally very hard since computational complexity is exponentially proportional to the number of regularization parameters.

To resolve this problem, SOVAL estimates $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ by maximizing the following (pseudo) log-likelihood

$$\mathscr{L}_2(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left[ y_i^{(j)} f_j(\mathbf{x}_i) - \log\left(\exp\left(f_j(\mathbf{x}_i)\right) + 1\right)\right] \tag{3}$$

subject to $\sum_{k=1}^{p} \sum_{j=1}^{J} \left|\beta_k^{(j)}\right| \leqslant \lambda$. Note that there is a single regularization parameter $\lambda$. Moreover, SOVAL is equally flexible to the standard OVA approach in the sense that if the optimal model is constructed using the standard OVA approach with the regularization parameters $\lambda_1, \ldots, \lambda_J$, the same model can be constructed using SOVAL with the regularization parameter $\lambda = \sum_{j=1}^{J} \lambda_j$.

Once the regression coefficients are estimated, the class probabilities are estimated by

$$\Pr(y = j \,|\, \mathbf{x}) = \frac{1}{C(\mathbf{x})} \Pr\left(y^{(j)} = 1 \,\middle|\, \mathbf{x}\right),$$

where $C(\mathbf{x}) = \sum_{m=1}^{J} \Pr\left(y^{(m)} = 1 \,\middle|\, \mathbf{x}\right)$. And the corresponding classifier can be constructed similarly to the SML case. Also, the gene importance measure is defined similarly (that is, $\rho_k = \sum_{j=1}^{J} \left|\beta_k^{(j)}\right|$).

## 3. A computational algorithm

We first present a general version of the gradient LASSO algorithm developed by Kim et al. (2005), and explain how to modify it for SOVAL as well as SML. Let $\mathbf{z} \in R^q$ and $\mathscr{L}(\mathbf{z})$ be a convex function defined on $R^q$. The objective of the gradient LASSO is to find the minimizer of $\mathscr{L}(\mathbf{z})$ over $\mathbf{z} \in D$ where $D$ is the subset of $R^q$ defined by $D = \left\{\mathbf{z} \in R^q : \sum_{k=1}^{q} |z_k| \leqslant 1\right\}$. Let $\mathbf{e}_k$ be the vector in $R^q$ with the $k$th component equal 1 and the others 0. Fig. 1 is the gradient LASSO algorithm for this problem.

The hardest part of the gradient LASSO is the step (a)(ii) and (b)(iv) for obtaining $\hat{\alpha}$ and $\hat{\delta}$, but it can be done using standard optimization techniques such as the Newton–Raphson algorithm. That is, the gradient LASSO algorithm does not require any special non-linear optimization algorithms. Also, Kim et al. (2005) proved that the convergence rate of the gradient LASSO is $1/m$ where $m$ is the number of iterations under some regularity conditions. A surprising result is that this convergence rate does not depend on the dimension of inputs which is very large for gene expression data. This feature makes the gradient LASSO algorithm well suited for analyzing gene expression data.

In SOVAL as well as SML, the intercept term $\boldsymbol{\beta}_0$ is not constrained, and hence the gradient LASSO algorithm cannot be applied directly. For this, we propose to estimate the intercept term $\boldsymbol{\beta}_0$ by letting $\boldsymbol{\beta} = 0$, and maximize

1. **Initialization:** $\mathbf{z} = \mathbf{0}$.

2. **Repeat until convergence** ;

   (a) Addition:

      i. Find
      $$\hat{k} = \text{argmax}_{k=1,\ldots,q} |\partial \mathcal{L}(\mathbf{z})/\partial z_k|$$
      and let $s = \text{sign}\ \partial L(\mathbf{z})/\partial z_{\hat{k}}$.

      ii. Find
      $$\hat{\alpha} = \text{argmax}_{\alpha \in [0,1]} \mathcal{L}((1-\alpha)\mathbf{z} + \alpha s \lambda \mathbf{e}_{\hat{k}}).$$

      iii. $\mathbf{z} = (1-\hat{\alpha})\mathbf{z} + \hat{\alpha} s \mathbf{e}_{\hat{k}}$

   (b) Deletion:

      i. Let $\nabla(\mathbf{z}) = (\nabla_1(\mathbf{z}), \ldots, \nabla_q(\mathbf{z}))$ and $\theta(\mathbf{z}) = (\theta_1(\mathbf{z}), \ldots, \theta_q(\mathbf{z}))$ where $\nabla_k(\mathbf{z})$ is $\partial \mathcal{L}(\mathbf{z})/\partial z_k$ for $z_k \neq 0$ and 0 for $z_k = 0$ and $\theta_k(\mathbf{z}) = \text{sign}(z_k)$.

      ii. If $< \nabla(\mathbf{z}), \theta(\mathbf{z}) >< 0$ and $< \theta(\mathbf{z}), \mathbf{z} >= 1$, set
      $$\mathbf{h} = -\nabla(\mathbf{z}) + \frac{< \nabla(\mathbf{z}), \theta(\mathbf{z}) >}{\sigma(\mathbf{z})} \theta(\mathbf{z})$$
      where $\sigma(\mathbf{z})$ is the number of nonzero elements in $\mathbf{z}$.

      iii. Else, set $\mathbf{h} = -\nabla(\mathbf{z})$.

      iv. Compute $\hat{\delta} = \text{argmin}_{0<\delta<L} \mathcal{L}(\mathbf{z} + \delta \mathbf{h})$ where $L = \min\{-z_k/h_k : z_k h_k < 0\}$.

      v. Update $\mathbf{z} = \mathbf{z} + \hat{\delta}\mathbf{h}$.

Fig. 1. Gradient LASSO algorithm.

the log-likelihood functions $\mathscr{L}_1$ and $\mathscr{L}_2$ with respect to $\boldsymbol{\beta}$ only. For SML, $\boldsymbol{\beta}_0$ becomes

$$\beta_0^{(j)} = \log \frac{\bar{y}^{(j)}}{\bar{y}^{(J)}}$$

for $j = 1, \ldots, J-1$ where $\bar{y}^{(j)} = \sum_{i=1}^{n} I\,(y_i = j)\,/n$. Similarly, for SOVAL, we have

$$\beta_0^{(j)} = \log \frac{\bar{y}^{(j)}}{1 - \bar{y}^{(j)}}$$

for $j = 1, \ldots, J$. The gradient LASSO algorithm can be modified for two multiclass sparse logistic regressions by letting $\mathbf{z} = \boldsymbol{\beta}/\lambda$ and replacing $\mathscr{L}$ by either $\mathscr{L}_1$ or $\mathscr{L}_2$.

**Remark.** The gradient LASSO algorithm presented here is a simpler version of the original gradient LASSO algorithm of Kim et al. (2005). In fact, using a more complicated version of the gradient LASSO algorithm, we can estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ simultaneously. But, the algorithm for this is much more involved, and the results from estimating $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ sequentially as is done here are not much different from those that result from estimating $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ simultaneously.

## 4. Numerical experiments

We compare the two multiclass extensions of sparse logistic regressions on five publicly available data sets.

### 4.1. Data description

*Leukemia*: The data set for this project is the gene expression data from leukemia patients used in Golub et al. (1999). This data set comes from a study of gene expressions in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). There are two key subclasses of ALL, those arising from T-cells and those arising from B-cells. This data set is composed of 38 samples classified as ALL T cell or ALL B cell or AML in the training set and an independent test set of 34 samples. The training set contains 8 ALL T-cell and 19 ALL B-cell samples and 11 AML samples. The independent test set consist of 1 ALL T cell and 19 ALL B cell samples and 14 AML samples. Each sample contains 7129 gene expression values obtained from Affymetrix oligonucleotide microarrays. In this paper, we combine the training and test samples and analyze them together. This data set can be downloaded at http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html.

*Lymphoma*: This data set is available at http://llmpp.nih.gov/lymphoma/ and contains gene expression levels of the 3 most prevalent adult lymphoid malignancies: 42 samples of diffuse large Bcell lymphoma (DLBCL, class 0), 9 observations of follicular lymphoma (FL, class 1), and 11 cases of chronic lymphocytic leukemia (CLL, class 2). The total sample size is $n = 62$, and the expression of $p = 4026$ well-measured genes, preferentially expressed in lymphoid cells or with known immunological or oncological importance, are documented. More information on these data can be found in Alizadeh et al. (2000). We imputed missing values and standardized the data as described in Dudoit et al. (2002).

*Small, round blue-cell tumors*: This data set about the small, round blue cell tumors (SRBCTs) of childhood includes 63 samples classified as neuroblastoma, rhabdomyosarcoma, non-Hodgkin lymphoma and the Ewing family of tumors. Gene-expression data from the cDNA microarray experiment contains 6567 genes. For data preprocessing, we followed the protocol detailed in the supplementary information to Khan et al. (2001). This data set can be downloaded at http://research.nhgri.nih.gov/microarray/Supplement/Images/supplemental_data.

*Brain cancer*: This data set, presented in Pomeroy et al. (2002), contains $n = 42$ microarray gene expression profiles from five different tumors of the central nervous system, that is, 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors (AT/RTs), 8 primitive neuro-ectodermal tumors (PNETs) and 4 human cerebella. The raw data were originated using the Affymetrix technology and are publicly available at http://www-genome.wi.mit.edu/cancer. For data preprocessing, we followed the protocol described in the supplementary information to Pomeroy et al. (2002). After thresholding, filtering, applying a logarithmic transformation and standardizing each expression profile to zero mean and unit variance, a data set comprising $p = 5597$ genes remained.

*NCI60*: NCI60 is a data set of gene expression profiles of 60 National Cancer Institute (NCI) cell lines. These 60 human tumor cell lines are derived from patients with leukemia, melanoma, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers. The data set is comprised of gene-expression levels of $p = 7129$ genes for $n = 60$ human tumor cell lines which can be divided into 8 classes: eight breast, six CNS, seven colon, six leukemia, eight melanoma, nine non-small-cell lung carcinoma, six ovarian and eight renal tumors. A more detailed description of the data can be found at Staunton et al. (2001). This data set can be downloaded at http://discover.nci.nih.gov/datasetsPnas2001.jsp.

### 4.2. Prediction accuracy

We evaluated the prediction accuracy of the two sparse multiclass logistic regression models using random partition. This means that we divided the data set at random such that 70% of the data set becomes training samples and the other 30% test samples. We repeated this procedure 100 times and the averaged misclassification errors were reported. For selecting $\lambda$, we used the five-fold cross validation.

We used a number of preprocessing steps as was done by Guyon et al. (2001) that included: taking the logarithm of all values, normalizing sample vectors, normalizing feature vectors, and passing the results through a squashing function of the type $f(x) = c \arctan(x/c)$ to diminish the importance of outliers.

Along with the prediction errors, we investigated the effect of prescreening of genes to the prediction accuracy. One of the standard approaches for analyzing gene expression data is to pick out relevant genes using simple prescreening

Table 1
Average test errors

| Data | Method | Number of covariates | | | | | |
|---|---|---|---|---|---|---|---|
| (The number of classes) | | $p = 10$ | $p = 50$ | $p = 100$ | $p = 500$ | $p = 1000$ | Full |
| Leukemia | SML | 0.040 | 0.044 | 0.053 | 0.066 | 0.065 | 0.067 |
| (3) | SOVAL | 0.044 | 0.046 | 0.041 | 0.041 | 0.043 | 0.043 |
| Lymphoma | SML | 0.070 | 0.054 | 0.023 | 0.020 | 0.022 | 0.029 |
| (3) | SOVAL | 0.073 | 0.033 | 0.022 | 0.009 | 0.013 | 0.020 |
| Small, round blue-cell | SML | 0.029 | 0.024 | 0.032 | 0.043 | 0.043 | 0.044 |
| (4) | SOVAL | 0.007 | 0.017 | 0.017 | 0.020 | 0.020 | 0.021 |
| Brain | SML | 0.293 | 0.209 | 0.191 | 0.224 | 0.276 | 0.392 |
| (5) | SOVAL | 0.231 | 0.121 | 0.103 | 0.113 | 0.149 | 0.279 |
| NCI60 | SML | 0.500 | 0.476 | 0.373 | 0.373 | 0.429 | 0.543 |
| (8) | SOVAL | 0.481 | 0.426 | 0.334 | 0.256 | 0.286 | 0.477 |

measures to reduce computational costs as well as to improve prediction accuracy (see for example, Golub et al., 1999; Dudoit et al., 2002). Since multiclass problems are of current concern in this paper, we used the $F$-ratio of between class sum of squares to within class sum of squares for each gene, following Dudoit et al. (2002). For gene $l$, the $F$-ratio is defined as

$$\frac{BSS(l)}{WSS(l)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J} I\,(y_i = j)\left(\bar{x}_{.l}^{(j)} - \bar{x}_{.l}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{J} I\,(y_i = j)\left(x_{il} - \bar{x}_{.l}^{(j)}\right)^2},$$

where $\bar{x}_{.l}^{(j)}$ indicates the average expression level of gene $l$ for class $j$ samples, and $\bar{x}_{.l}$ is the overall mean expression level of gene $l$ in the training set. We use the $F$-ratio for its simplicity, and there are different types of the $F$-ratio.

Table 1 and Fig. 2 reports the test errors with different gene subset sizes obtained by the prescreening with the $F$-ratio, which shows that SOVAL is more accurate in most cases than SML. In some cases, the improvements are larger than 50%.

Second, we can see from Tables 1 and 2 that the prescreening affects the accuracy significantly. The optimum test errors are achieved around $p = 100$ or $p = 500$ (except for the data set small, round blue-cell where the optimum error is achieved when $p = 10$). From this finding, we may conclude that the purpose of prescreening is not to select relevant genes but to eliminate irrelevant genes. This result somehow contrasts with that of Dudoit et al. (2002) where finding small numbers of relevant genes by prescreening affects prediction accuracies significantly in some cases. A reason for this difference would be that we use sparse methods while Dudoit et al. (2002) do not. For non-sparse methods, the classifier depends on all genes used as inputs and so prescreening would be important. However, sparse methods automatically select genes while they construct a classifier, and so prescreening is not necessary. Moreover, the prescreening may drop some informative genes in an early stage, and the resulting model would be suboptimal. In this view, for sparse methods, efficient computational algorithms for dealing with large dimensional inputs without prescreening are necessary, and our algorithm is such an algorithm.

### *4.3. Performance of gene selection*

Table 2 presents the average number of genes selected from the two sparse methods. It shows that SML tends to yield more sparse models than SOVAL, in particular when the number of classes is large. Along with the error rates in Table 1, we can conclude that SML fails to detect some important genes, which results in higher error rates.

To confirm our conclusion, we did the following experiment. The effectiveness of gene identification was tested on miniature data sets synthesized from the original data. The miniature data sets of 100 genes were constructed as follows. First, using the $F$-ratio as a measure of marginal association between each gene and the tumor type, we ranked the genes and selected the top 20 genes as variables truly associated with the class. As irrelevant variables, we included the bottom 80 genes with the class label corresponding to each covariate vector of 80 genes randomly mixed together, so that they

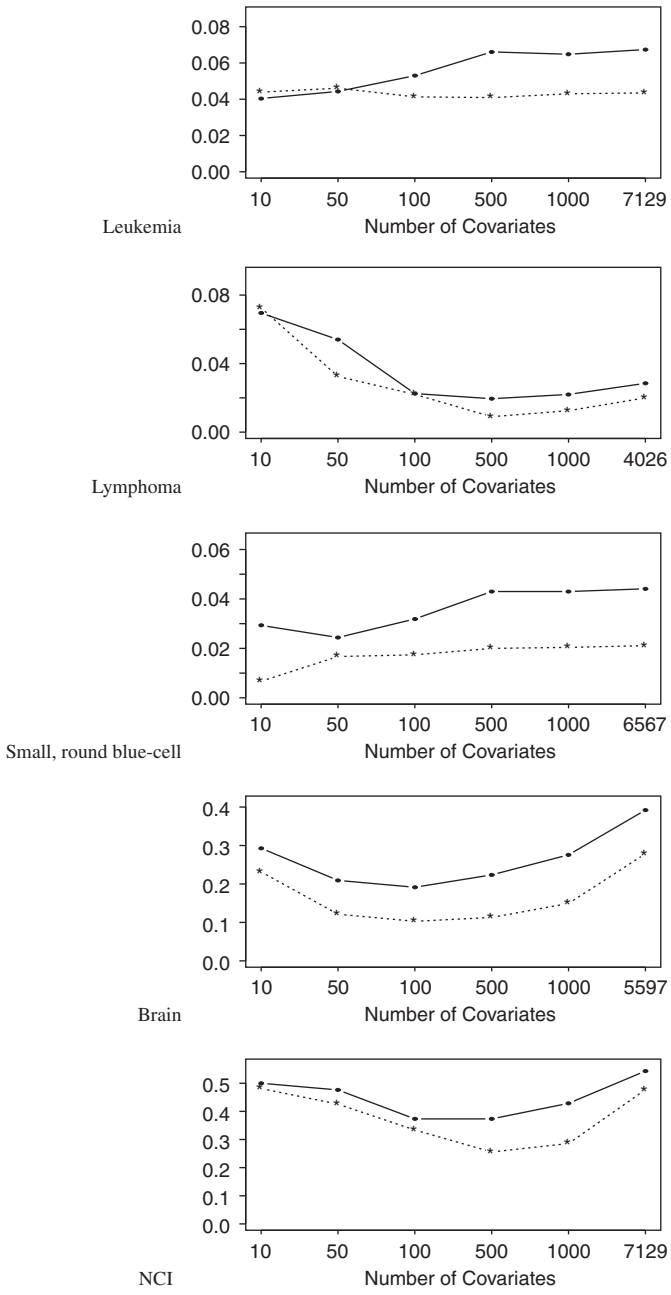


Fig. 2. Average test errors.

were genuinely unrelated to the class, but the potential correlations between those genes were intact. Ten replicates of synthetic training data were obtained by the 10-fold cross validation from these miniature data sets, keeping the class proportions in each sample the same as these in the original data. See Lin (2005) and Jung and Jang (2006) for similar experiments.

We applied the two sparse multiclass logistic regression models to these 10 replicates, and the optimal regularization parameters were selected with the 10 test data sets constructed from the 10 fold cross validation. Fig. 3 is the boxplot of the number of selected genes and the number of the selected genes among the 20 informative genes from the 10 replicates of the miniature data sets by the 10-fold cross validation of the original data sets. It shows that SOVAL includes more informative genes than SML when the number of classes is large.

Table 2
The averaged numbers of genes selected

| Data | Method | Number of covariates | | | | | |
|---|---|---|---|---|---|---|---|
| (The number of classes) | | $p = 10$ | $p = 50$ | $p = 100$ | $p = 500$ | $p = 1000$ | Full |
| Leukemia | SML | 4.78 | 12.46 | 14.23 | 17.50 | 18.80 | 21.62 |
| (3) | SOVAL | 5.42 | 17.03 | 22.31 | 32.82 | 33.96 | 34.54 |
| Lymphoma | SML | 8.55 | 14.55 | 13.76 | 16.09 | 17.65 | 18.19 |
| (3) | SOVAL | 9.12 | 18.77 | 22.65 | 36.14 | 39.33 | 46.57 |
| Small, round blue-cell | SML | 9.59 | 22.02 | 27.67 | 30.74 | 32.73 | 33.44 |
| (4) | SOVAL | 9.80 | 24.26 | 31.02 | 33.86 | 35.29 | 35.16 |
| Brain | SML | 8.72 | 20.10 | 22.89 | 25.66 | 26.81 | 32.32 |
| (5) | SOVAL | 9.54 | 28.20 | 38.22 | 55.12 | 59.04 | 63.27 |
| NCI60 | SML | 9.06 | 27.18 | 37.97 | 46.01 | 45.85 | 60.52 |
| (8) | SOVAL | 9.65 | 29.69 | 49.06 | 89.14 | 102.65 | 118.58 |

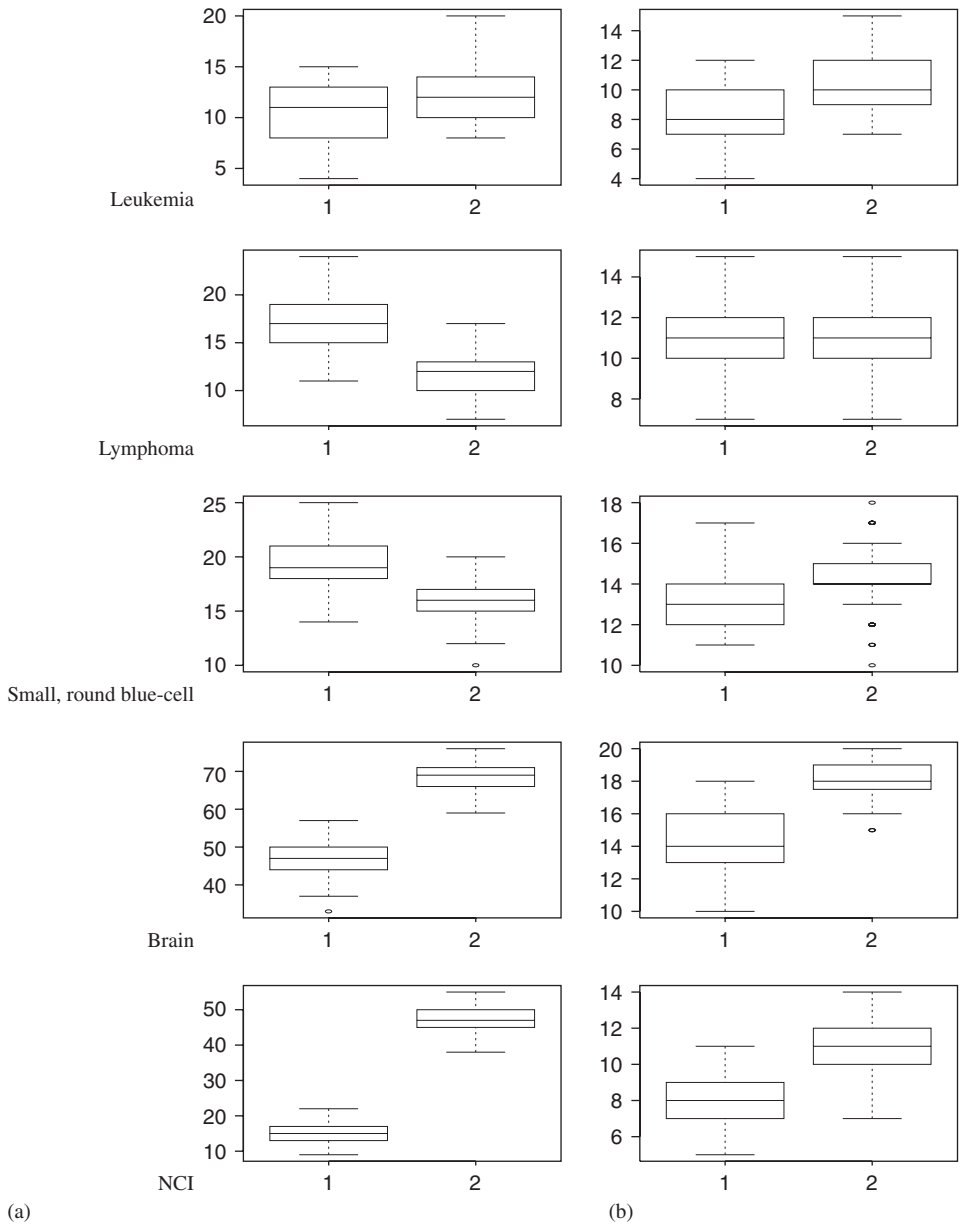


Fig. 3. The boxplots of: (a) the total number of genes selected and (b) the number of genes selected among the top 20 informative genes.

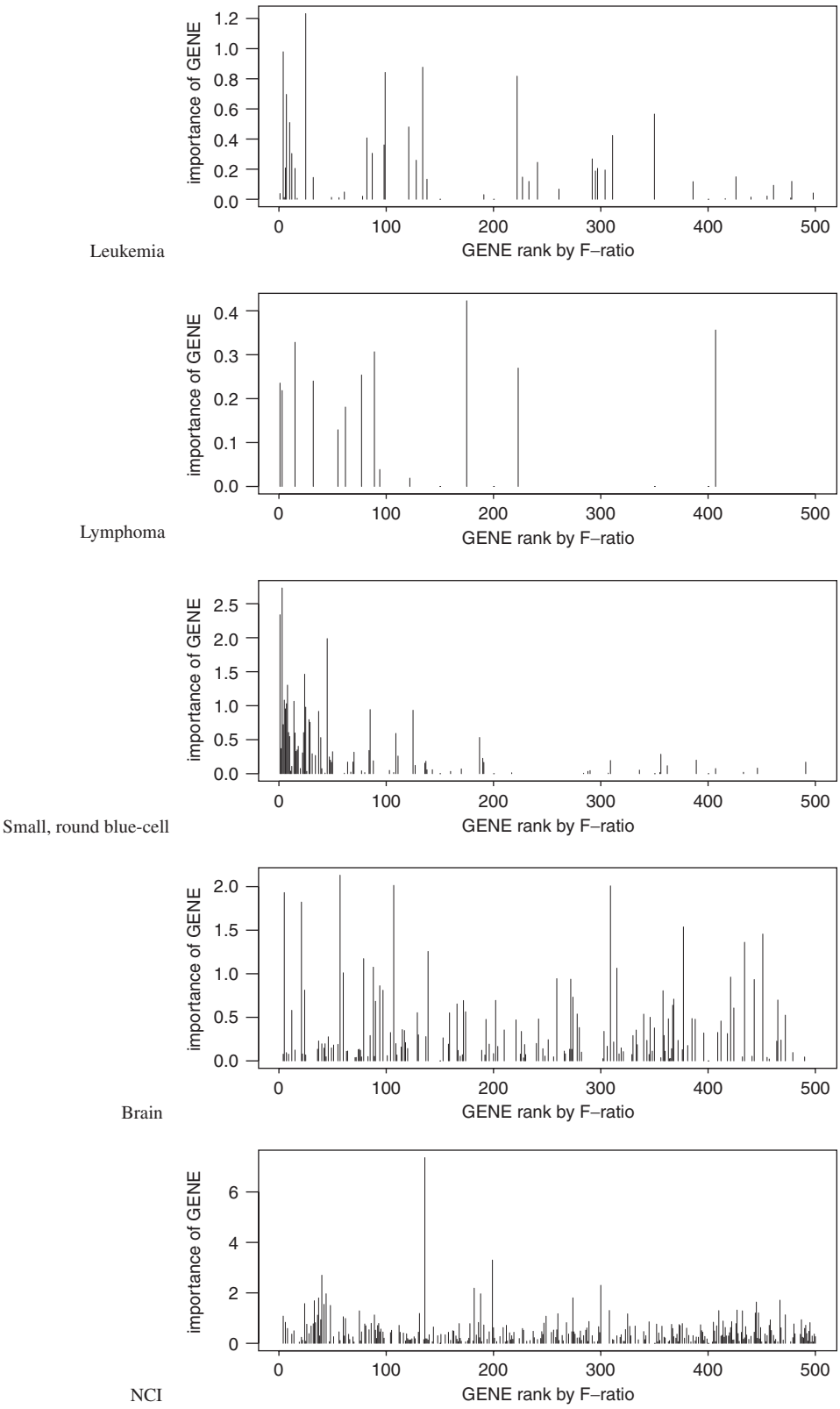


Fig. 4. The plots of the importance versus gene rank by *F*-ratio.

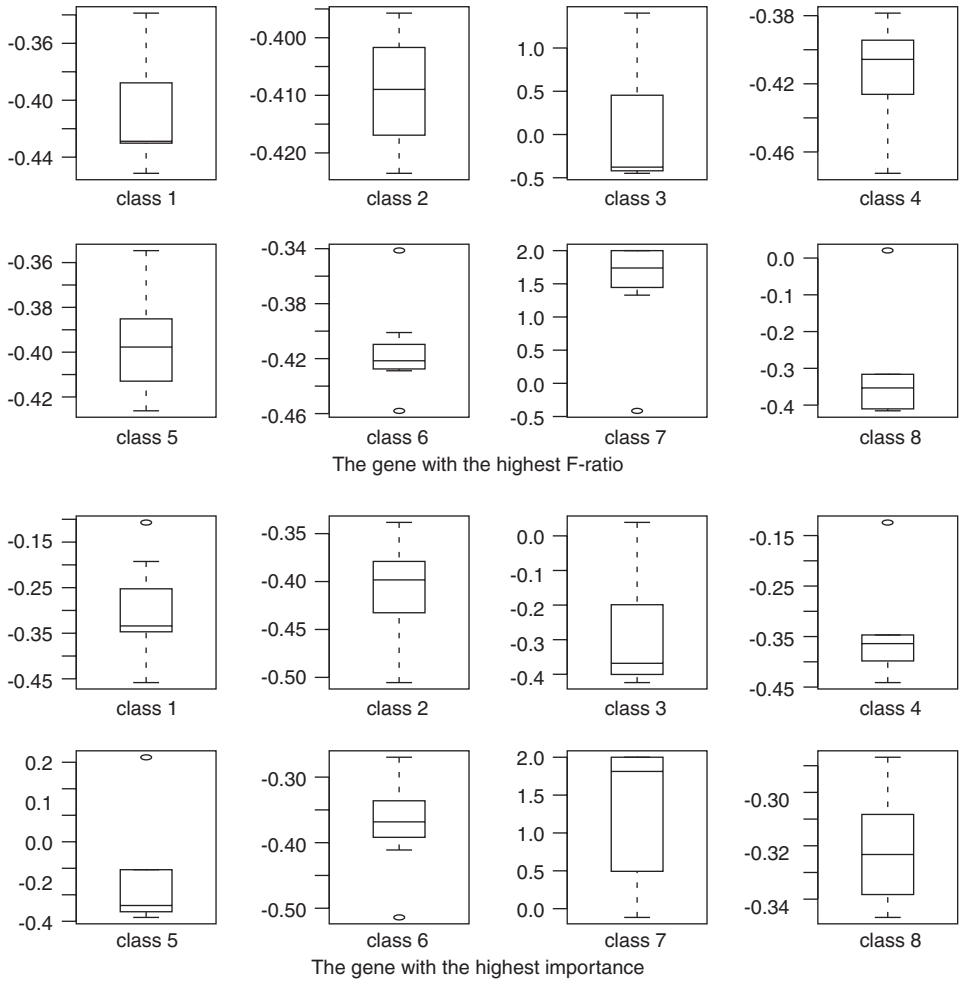


Fig. 5. The boxplots of the expression levels of the two genes having the highest $F$-ratio and highest importance according to the class labels in the NCI data set.

Finally, we compared genes selected from SOVAL and genes selected from the marginal $F$-ratio. Fig. 4 shows the plots where the $x$-axis displays the gene ranks obtained by the marginal $F$-ratio and the $y$-axis is gene importance measured by the SOVAL. The results are striking, in particular when the number of classes is large (i.e. in the data sets Brain and NCI). There are many genes having simultaneously lower ranks of the marginal $F$-ratio but having larger importance.

To understand why this happens, we select the two genes from the NCI data sets, one which have the largest $F$-ratio and the other which has the largest importance. The rank of the $F$-ratio of the gene with the highest importance is 132, and the importance of the gene with the highest $F$-ratio is 0. That is, these two genes have significantly different $F$-ratio and gene importance values. Fig. 5 presents the boxplot of the gene expression levels of these two genes according to the class labels. First of all, the distributions of the expression levels of the two genes are similar. They have large positive expression levels at the seventh class and negative expression levels for the other classes. An exception is the third class, where the gene with the highest $F$-ratio has expression levels around 0 while the gene with the highest importance has negative expression levels. This difference partially explains why the ranks from the $F$-ratio and from gene importance are quite different. The $F$-ratio measures the variation of the mean expression levels of the classes, and so the gene with the highest $F$-ratio has additional variation due to the third class compared to the gene with the highest importance. In contrast, SOVAL basically measures the difference of the means from one class to the other classes. For the seventh class, this difference is larger for the gene with the highest importance than for the gene with

the highest $F$-ratio. So, we conclude that if we want to detect genes which affect all the classes, the $F$-ratio would be more appropriate. However, if we want to detect genes which affect a certain class, sparse logistic regression would be more appealing.

## 5. Concluding remarks

In this paper, we proposed a multiclass extension of sparse logistic regression, so called SOVAL, compared it with SML, and developed the efficient computational algorithm suitable for gene expression data. The numerical experiments showed that SOVAL outperforms SML in many aspects. The former: (i) gives better accuracies in particular; (ii) has higher power of detecting important genes and (iii) does not require the choice of a baseline class.

The main idea of SOVAL is somehow related to the Scott's method of estimating a mixture model (Scott, 2001, 2004). The Scott's method relaxed a constraint of the density function and focused on a particular component rather than all components. SOVAL also relaxed a constraint that the sum of the probabilities of the classes is 1 and implicitly found genes important for a specific class rather than all classes. This similarity would partially explain the good prediction performance of SOVAL. We leave this conjecture as a future work.

We have seen that the selected genes by SOVAL are much different from those selected by the marginal $F$-ratio. This is partly because SOVAL measures the classification power of genes for a specific class while the marginal $F$-ratio measures the overall effect of genes on all classes. Hence, if one wants to detect genes which affect a specific class, SOVAL is more suitable. In this view, SOVAL can be considered as a new way of detecting relevant genes and can be used as a preprocessing procedure for more complicated non-linear classification methods such as the support vector machine or boosting. For this purpose, however, efficient computational algorithms are required since we should work with large numbers of genes without prescreening, and the algorithm proposed in this paper can serve for this purpose.

## Acknowledgments

The first author and second author were supported in part by KOSEF through the Statistical Research Center for Complex Systems at Seoul National University. The third author was supported in part by KOSEF (R14-2003-002-01000).

## References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37 (4), 373–384.

Dudoit, S., Fridlyand, J., Speed, T., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Jung, S.H., Jang, W., 2006. How accurately can we control the FDR in analyzing microarray data? Bioinformatics, to appear. http://bioinformatics.oxfordjournals.org/cgi/reprint/btl161?

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Atonescu, C., Peterson, C., Meltzer, P., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med. 7, 673–679.

Kim, J., Kim, Y., Kim, Y., 2005. A gradient descent algorithm for generalized LASSO. Technical Report, Department of Statistics, Seoul National University, Korea. ⟨http://idea.snu.ac.kr/ResearchTechnical.html⟩.

Krishnapuram, B., Carlin, L., Figueiredo, M., Hartemink, A., 2004. Learning sparse classifier: multi-class formulation, fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. 27, 957–968.

Li, Y., Campbell, C., Tipping, M., 2002. Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18, 1332–1339.

Lin, D.Y., 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 43, 274–285.

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., et al., 2002. Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature 415, 436–442.

Roth, V., 2002. The generalized LASSO: a wrapper approach to gene selection for microarray data. Technical Report, University of Bonn, Computer Science III.

Scott, D.W., 2001. Parametric statistical modeling by minimum integrated square error. Technometrics 43, 274–285.

Scott, D.W., 2004. Partial mixture estimation and outlier detection in data and regression. In: Theory and Applications of Recent Robust Methods.Birkhäuser, Basel, pp. 274–285.

Shevade, K., Keerthi, S., 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19, 2246–2253.

Staunton, J., Slonim, D., Coller, H., Tamayo, P., Angelo, M., Park, J., Scherf, U., Lee, J., Reinhold, W., Weinstein, J., Mesirov, J., Lander, E., Golub, T., 2001. Chemosensitivity prediction by transcriptional profiling. Proc. Nat. Acad. Sci. 98 (19), 10787–10792.