# Bayesian Model Averaging: Development of an improved multi-class, gene selection and classification tool for microarray data

Ka Yee Yeung[1], Roger E. Bumgarner[1] and Adrian E. Raftery[2]

[1]Department of Microbiology, University of Washington, Seattle, WA 98195
[2]Department of Statistics, University of Washington, Seattle, WA 98195

| | | |
|---|---|---|
| **Report Documentation Page** | | *Form Approved*<br>*OMB No. 0704-0188* |

| 1. REPORT DATE<br>**08 OCT 2004** | 2. REPORT TYPE | | 3. DATES COVERED<br>**00-10-2004 to 00-10-2004** |
|---|---|---|---|
| 4. TITLE AND SUBTITLE<br>**Bayesian Model Averaging: Development of an improved multi-class, gene selection and classification tool for microarray data** | | 5a. CONTRACT NUMBER | |
| | | 5b. GRANT NUMBER | |
| | | 5c. PROGRAM ELEMENT NUMBER | |
| 6. AUTHOR(S) | | 5d. PROJECT NUMBER | |
| | | 5e. TASK NUMBER | |
| | | 5f. WORK UNIT NUMBER | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**University of Washington,Department of Statistics,Box 354322,Seattle,WA,98195-4322** | | 8. PERFORMING ORGANIZATION REPORT NUMBER | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | 10. SPONSOR/MONITOR'S ACRONYM(S) | |
| | | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) | |
| 12. DISTRIBUTION/AVAILABILITY STATEMENT<br>**Approved for public release; distribution unlimited** | | | |
| 13. SUPPLEMENTARY NOTES | | | |
| 14. ABSTRACT | | | |
| 15. SUBJECT TERMS | | | |

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | | **22** | |

# Abstract

Accurate classification of samples using gene expression profiles is critically dependent on the method used to select relevant genes. We present the Bayesian Model Averaging (BMA) method for gene selection and classification of microarray data. Typical gene selection and classification procedures ignore model uncertainty and use a single set of relevant genes (model) to predict the class. BMA accounts for the uncertainty about the best set to choose by averaging over multiple models (sets of potentially overlapping relevant genes).

We showed that BMA selects smaller numbers of relevant genes (compared to other methods) and achieves high prediction accuracy on three microarray datasets. Our BMA algorithm is applicable to microarray datasets with *any number of classes*, and outputs posterior probabilities for the selected genes and models. Our selected models typically consist of only a few genes. The combination of high accuracy, small numbers of genes and posterior probabilities for the predictions, should make BMA a powerful tool for developing diagnostics from expression data.

# Contents

## List of Tables

## List of Figures

## 1    INTRODUCTION

There has been a recent explosion in the use of microarray data for classification in a variety of diagnostic areas. The prediction of the diagnostic category of a tissue sample from its expression array phenotype given the availability of similar data from tissues in identified categories is known as *classification (*or *supervised learning*). In the context of gene expression data, the samples are usually the experiments, and the classes are usually different types of tissue samples, for example, cancer vs. non-cancer (Alon *et al.* 1999; Schummer *et al.* 1999), different tumor types (Golub *et al.* 1999; Alizadeh *et al.* 2000; Ross *et al.* 2000; Bhattacharjee *et al.* 2001; Ramaswamy *et al.* 2001), response to therapy (Shipp *et al.* 2002; van 't Veer *et al.* 2002; Nutt *et al.* 2003). A challenge in predicting the diagnostic categories using microarray data is that the number of genes is usually much greater than the number of tissue samples available, and only a subset of the genes is relevant in distinguishing different classes. Selection of relevant genes for classification is known as *feature selection*. A small set of relevant genes is essential for the development of inexpensive diagnostic tests.

Multi-class classification in which the data consist of more than two classes is rapidly gaining attention in the literature. For example, Ramaswamy *et al.* (2001) combined support vector machines, which are binary classifiers, to solve the multi-class classification problem. Nguyen and Rocke (2002a; 2002b) used partial least squares (PLS) for feature selection, together with traditional classification algorithms such as logistic discrimination and quadratic discrimination to classify multiple tumor types on microarray data. Tibshirani *et al.* (2002) developed an integrated feature selection and classification algorithm called shrunken centroid for classifying multiple cancer types in which features are selected by considering one gene at a time. Yeung and Bumgarner (2003) extended the shrunken centroid algorithm to take dependency between genes and repeated measurements into consideration. Dudoit *et al.* (2002) compared the performance of different discrimination methods, including nearest neighbor classifiers, linear discriminant analysis, and classification trees, for classifying multiple tumor types using gene expression data.

The method used for selecting relevant genes is critical to the performance of all classification algorithms. Most feature selection methods in the literature are tailored towards *binary* classification, and are *univariate* in the sense that each candidate relevant gene is considered individually. Examples of univariate methods include the signal-to-noise ratio (Golub *et al.* 1999), the t-test (Nguyen and Rocke 2002b), the ratio of between-groups to within-groups sum of squares (BSS/WSS) (Dudoit *et al.* 2002), the Significance Analysis of Microarray (SAM) statistic (Tusher *et al.* 2001), the Threshold Number of Misclassification (TNOM) score (Ben-Dor *et al.* 2000), the Wilcoxon test statistic (Dettling 2004) and many others. *Multivariate* gene selection methods consider multiple genes

simultaneously, and hence, account for dependency between genes, which hopefully will lead to a reduced number of relevant genes. Bo and Jonassen (2002) evaluated relevant genes in a pairwise fashion, while Jaeger *et al.* (2003) and Yeung and Bumgarner (2003) reduced the number of relevant genes by eliminating highly correlated ones. Recently, Lee *et al.* (2003) employed a hierarchical Bayesian model which used a Markov Chain Monte Carlo (MCMC) based stochastic search algorithm to discover relevant genes. Their multivariate gene selection algorithm is applicable to microarray data with two classes only. Sha *et al.* (2004) extended the underlying theory to multiple classes data as well, but did not give empirical results for gene selection on multi-class microarray data.

In addition, most proposed feature selection and classification algorithms ignore model uncertainty by selecting one set of relevant genes, and then by using class prediction on that set of selected genes. It is possible that there is more than one set of relevant genes that fit the model equally well, especially with microarray data in which the number of genes (variables) is much greater than the number of samples. There have been efforts to use model averaging and model ensemble approaches to classify microarray data. As an example, Li and Yang (2002) applied a model averaging approach to classify samples by averaging over multiple single-gene models to microarray data. Boosting algorithms have also been applied to microarray data (Ben-Dor *et al.* 2000; Dudoit *et al.* 2002; Dettling and Buhlmann 2003; Dettling 2004).

In this paper, we present the Bayesian Model Averaging (BMA) approach (Raftery 1995; Hoeting *et al.* 1999; Viallefont *et al.* 2001) as our multivariate feature selection method for multi-class microarray data. This is in contrast to Li and Yang (2002) in which the emphasis was on classification and genes were selected independently. Our approach also differs from Lee *et al.* (2003) and Sha *et al.* (2004) in the sense that we adopt a model averaging approach and we report empirical results on multi-class as well as binary microarray data. In addition, our algorithms are computationally efficient compared to the MCMC based algorithms in Lee *et al.* (2003) and Sha *et al.* (2004). We extended an existing BMA algorithm to be applicable to any number of input variables (genes), and to any number of classes. We show that our extended BMA algorithm generally selects fewer relevant genes and produces prediction accuracy at least comparable to that of the best existing feature selection and classification methods. We also propose to use the Brier Score (Brier 1950) and use a generalized Brier Score to assess prediction accuracy for 2-class and multi-class datasets respectively. Our approach has the additional advantage of facilitating biological interpretation by producing posterior probabilities of selected genes and models. Our BMA algorithm is a multivariate gene selection method, and our selected models are typically very simple, consisting of only a few genes. By averaging over multiple simple models and using relatively small numbers of relevant genes, we

demonstrate high prediction accuracy on both binary and multi-class microarray data. In Section 2, we review Bayesian Model Averaging (BMA) and describe our extension of existing BMA algorithms to large numbers of predictors and multi-class classification problems, and in Section 3, we give results for three gene expression datasets.

## 2   METHODS

### 2.1   Bayesian Model Averaging (BMA)

Typical model selection approaches select a model and then proceed as if the selected model has generated the data, which might lead to over-confident inferences. Bayesian Model Averaging (BMA) takes model uncertainty into consideration by averaging over the posterior distributions of multiple models, weighted by their posterior model probability (Raftery 1995; Hoeting *et al.* 1999).

For simplicity, let us first consider the binary classification problem. Let Y be the response variable (class) of a sample in the test set, where Y = 0 or 1, and let D be the *training dataset* for which the classes are known. The essence of BMA is shown in Equation 1: the posterior probability of Y=1 given the training set D is the weighted average of the posterior probability of Y=1 given the training set D and model $M_k$ multiplied by the posterior probability of model $M_k$ given training set D, summing over a set of models $M_k$ in M:

$$\Pr(Y = 1 \mid D) = \sum_{k \in M} \Pr(Y = 1 \mid D, M_k) * \Pr(M_k \mid D). \tag{1}$$

BMA presents several implementation difficulties. One of these is that the exhaustive summation of all feasible models leads to an enormous number of terms in Equation 1. Raftery (1995) used the leaps and bounds algorithm (Furnival and Wilson 1974) to efficiently identify a reduced set of good models. The leaps and bounds algorithm rapidly returns the best "nbest" models of each size (up to 30 variables). Madigan and Raftery (1994) proposed to use the Occam's window method to choose a set of parsimonious and data-supported models. Their idea is to discard models that are much less likely than the best model supported by the data (the default is 20 times less likely).

A second difficulty with BMA is that there is an implicit integral associated with the evaluation of the posterior probability for model $M_k$ given training set D. Using Bayes's Theorem, $\Pr(M_k|D)$ is proportional to $\Pr(D|M_k)*\Pr(M_k)$, where $\Pr(D|M_k)$ is the integrated likelihood of model $M_k$ in which the regression parameters for model $M_k$ are integrated over. Please refer to Raftery (1995) and Hoeting *et al.* (1999) for the mathematical details. There are many different ways to approximate this integral including MCMC approximations (Madigan and York 1995). In this paper, we use logistic regression (Hosmer and Lemeshow 2000) to predict $\Pr(Y=1|D, M_k)$ such that $\ln[\Pr(Y=1|D, M_k)/ \Pr(Y=0|D, M_k)] =$

$b_0+b_1x_1+\ldots+b_px_p$, where $x_i$'s represent the expression levels of selected genes and $b_i$'s are the regression parameters. In this case, the Bayesian Information Criterion (BIC) can be used to approximate the integral (Raftery 1995). We adopt the BMA implementation (Raftery 1995) which makes use of the BIC approximation. The source code of the BMA implementation is available at http://www.research.att.com/~volinsky/bma.html.

P-values computed from typical variable selection procedures (such as stepwise forward or backward selection) have been shown to overstate the strength of inference (Raftery 1995; Viallefont *et al.* 2001). One of the advantages of BMA is that it yields an easily interpreted summary: posterior probabilities for the selected models and the selected genes (variables). In particular, our adopted traditional BMA implementation (Raftery 1995) outputs the posterior probability that each variable is non-zero, *probne0*.

## 2.2   Our modifications to existing BMA algorihms

## Iterative BMA algorithm

With microarray data, the number of genes (variables) is typically much greater than the number of samples (responses). However, in the traditional BMA implementation (Raftery 1995), the leaps and bounds algorithm can only compute the best "nbest" models for up to 30 variables, and if the number of variables is greater than 30, backward elimination is used to reduce the number of variables to 30 before applying the leaps and bounds algorithm. However, stepwise backward elimination in which one variable is removed at a time cannot be applied in this situation in which there are more predictors (genes) than observations (samples). Instead, we developed an iterative BMA algorithm which first rank orders genes with a univariate gene selection method and then moves a 30-variable window down the ordered genes. Recall that *probne0* represents the posterior probability of a gene not being equal to zero, and hence, genes with high *probne0* are good candidates for relevant genes.

Outline of Iterative BMA Algorithm

**Input:** training set D with G genes and n samples
**Pre-processing step:** Rank all G genes using a univariate gene selection procedure. Let $x_1, x_2, \ldots, x_G$ be the ordered list of genes.
**Parameters:** nbest and p, where p is the total number of genes to be processed such that $30 < p = G$.
 1. Initially, start with the 30 top ranked genes $(x_1, x_2, \ldots, x_{30})$, and apply the traditional BMA algorithm. Let *toBeProcessed* be an ordered list of genes with ranks 31 to p. Initially, *toBeProcessed* <- $x_{31}, x_{32}, \ldots, x_p$.

2. Repeat until all p genes are processed

    a. Remove all genes with *probne0* < 1%.

    b. *Adaptive threshold step:* If all genes have *probne0* = 1%, determine the minimum probne0, *minProbne0*, among the 30 genes in the current window. Remove all genes with *probne0* < (*minProbne0* +1)%.

    c. Let *removedGenes* be the set of genes removed, and suppose q genes are removed.

    d. Replace the q removed genes with the next q genes from *toBeProcessed*. Update *toBeProcessed* <- *toBeProcessed* – *removedGenes*.

    e. Apply the traditional BMA algorithm.

**Output:** selected models and their posterior probabilities, selected genes and their corresponding probne0, maximum likelihood estimates of the regression parameters in each model

In our study, we used the ratio of between-group to within-group sum of squares (BSS/WSS) (Dudoit *et al.* 2002) to determine the initial gene order. Intuitively, genes with relatively large variation between classes and relatively small variation within classes are likely candidates as relevant genes. BSS/WSS is a univariate gene selection method in which genes with large BSS/WSS ratios are good candidate relevant genes. For a gene j, let $D_{ij}$ denote the expression level of gene j under sample i, $\overline{D}_{kj}$ denote the average expression level of gene j over samples in class k, and $\overline{D}_{.j}$ denote the average expression level of gene j over all samples. The BSS/WSS ratio for gene j is defined as

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(Y_i = k)\left(\overline{D}_{kj} - \overline{D}_{.j}\right)^2}{\sum_i \sum_k I(Y_i = k)\left(D_{ij} - \overline{D}_{kj}\right)^2}, \tag{2}$$

where $I(Y_i=k)$ is equal to 1 if sample i belongs to class k and is equal to 0 otherwise. In step 1 of the iterative BMA algorithm, we compute the BSS/WSS ratio for each of the G genes and order the genes in descending order of the BSS/WSS ratio.


**Multi-class Iterative BMA**

For multi-class microarray data, we developed an individualized regression approach in which binary logistic regressions are combined. We used the approximation of Begg and Gray (1984) (also discussed in Chapter 8 of Hosmer and Lemeshow (2000)). They studied the use of a series of individualized binary logistic regressions as an approximation for polychotomous logistic regression in

which the response variable can take more than two values. They showed that this provides a close approximation to maximum likelihood estimation of the full multinomial logistic regression model. For our purposes, it is particularly attractive because it allows us to use the well-established and computationally efficient algorithms for BMA in binary logistic regression when building BMA for multi-class classification.

Suppose there are K classes such that the response variable (class) Y takes on values 0, 1, …, or (K-1), where K = 3, and let $Y_i$ be the response variable for sample i. Our idea is to use a separate binary logistic regression to discover relevant genes for each training subset (Y=0 vs. Y=k), where k = 1, .., (K-1), and use the Begg and Gray (1984) approach to create an augmented matrix M to approximate polychotomous logistic regression using the selected genes from each training subset with binary logistic regression. Figure 1 shows a flowchart of our algorithm with an example augmented matrix M for K=3. The augmented matrix M is formed by concatenating the selected genes from each training subset and pasting the two training subsets (Y=0 vs. Y=1) and (Y=0 vs. Y=2) together. There is a column in M for the regression parameter of each gene. The first $n_1$ rows of M correspond to samples with Y=0 or Y=1 and the next $n_2$ rows of M correspond to samples with Y=0 or Y=2. Finally, we order the columns in M using BSS/WSS ratios and apply the iterative BMA algorithm to M to discover relevant genes.

<u>Outline of multi-class iterative BMA</u>

1. Using Y=0 as our baseline, create subsets of the samples from the training set for the binary classification problem in which Y=0 or Y=k, where k = 1, 2, .., (K-1), and ignore all the data from Y? 0 and Y? k. Denote the number of training samples for Y=0 vs. Y=k by $n_k$. In the training subset (Y=0 vs. Y=k), the response variable Y*=0 when Y=0, and Y*=1 when Y=k.

2. For each training sample subset (Y=0 vs. Y=k) where k=1, 2, …, (K-1), apply the iterative BMA algorithm, and let $S_k$ be the set of selected genes from this subset.

3. Merge the selected genes from each training sample subset to create an augmented design matrix with ordered columns, M, which has $\sum_{k=1}^{K} n_k$ rows and $(K - 2 + \sum_{k=1}^{K-1} |S_k|)$ columns (variables).

    a. Compute BSS/WSS ratios for each gene in $S_k$ from each training sample subset k.

    b. Sort the BSS/WSS ratios from all (K-1) training sample subsets $S_k$.

c. The first (K-2) columns of the design matrix M represent the "intercept" columns while all other columns represent genes (variables). The first $n_1$ rows of M represent the training sample subset Y=0 or Y=1, and the next $n_2$ rows of M represent Y=0 or Y=2 etc.

d. For k= 2 to (K-1), M[i, k-1]=1 for any sample i in training subset k in which $Y_i$=0 or $Y_i$=k, and M[i, k-1] =0 for all other samples.

e. For k=1 to (K-1) and each gene g in $S_k$, M[i, (K-2)+r]=$D_{ig}$ where r is the rank of gene g from step (3b) and $D_{ig}$ is the expression level of gene g under sample i in the training set D for any sample i in training subset k ($Y_i$=0 vs. $Y_i$=k), and M[i, (K-2)+r]=0 otherwise.

f. The response variable for M, $Y^M$=0 for Y=0, and $Y^M$=1 for Y=k where k=1, 2,.., (K-1).

4. Apply the iterative BMA algorithm to the augmented data matrix M.

5. Prediction step: use the regression parameters from the selected variables from Step 4.

## 2.3 Evaluation of predictive performance

In the literature, the number of classification errors is the most popular measure of predictive performance, for example, (Golub *et al.* 1999; Nguyen and Rocke 2002a; van 't Veer *et al.* 2002; Lee *et al.* 2003). However, in our case, the predicted probability for each class, Pr(Y=k|D), is available. For example, a predicted probability close to 0 or 1 is more desirable than a predicted probability around 0.5 in the binary classification case. In order to take the magnitudes of predicted probabilities into consideration, we adopted the Brier Score (Brier 1950) as our evaluation measure. For binary data, let $Y_i$ denote the response variable (class) of sample i, where $Y_i$ = 0 or 1. Denote the predicted probability that sample i belongs to class 1, Pr($Y_i$=1|D), by $p_i$. The Brier Score is defined as $\sum_{i=1}^{n}(Y_i - p_i)^2$, which is the sum of squares of the difference between the true class and the predicted probability over all samples. If the predicted probabilities, $p_i$, are constrained to equal to 0 or 1, the Brier Score is equal to the total number of classification errors. Thus the Brier Score allows us to compare the performance of the deterministic 0-1 classification methods with that of probabilistic methods such as BMA, which is an appealing feature.

We use the *generalized Brier Score* for the multi-class case, where $Y_i$ = 0, 1, …, (K-1). Let $Y_{ik}$ be an indicator variable such that $Y_{ik}$=1 if $Y_i$=k and $Y_{ik}$=0 otherwise, where k = 0, 1, …, (K-1). Let $p_{ik}$ denote

the predicted probability that $Y_i = k$. The generalized Brier Score is defined as $\frac{1}{2} \sum_{i=1}^{n} \sum_{k=0}^{K-1} (Y_{ik} - p_{ik})^2$. It

can be shown that the generalized Brier Score is reduced to the Brier Score when K=2. A high generalized Brier Score indicates poor predictive performance.

## 3   RESULTS

### 3.1   Breast cancer prognosis data (2-class)

The breast cancer prognosis dataset (van 't Veer *et al.* 2002) consists of primary breast tumor samples hybridized to cDNA arrays consisting of 24481 genes with 78 samples in the training set, and 19 samples in the test set. These samples are divided into two categories: the good prognosis group (patients who remained disease free for at least 5 years) and the poor prognosis group (patients who developed distant metastases within 5 years). We identified 4919 significantly regulated genes (at least a two-fold difference and p-value < 0.01 in least 3 samples) from the training set. We further deleted two samples with missing values from the training set. Therefore, the breast cancer prognosis training set used in our experiments consists of 76 samples and the test set consists of 19 samples (see Table 1) across 4919 genes.

We applied the iterative BMA algorithm for binary classification to the breast cancer prognosis data, and achieved a comparable number of classification errors on the test set to the reported results in van't Veer *et al.* (2002) while using significantly fewer relevant genes. We experimented with various control parameters for the iterative BMA algorithm in our study, including the number of models returned by the leaps and bounds algorithm for up to 30 variables (nbest) and the number of top genes ranked by BSS/WSS ratios (p). We observed that a large p (1000 or more genes) typically yields lower Brier Scores and classification errors, and with the exception of nbest=10, which is too small, the prediction accuracy and the number of selected genes are relatively insensitive to "nbest".

Using all 4919 genes and nbest=20, our iterative BMA algorithm produced 3 classification errors on the test set (out of 19 samples) and a Brier Score of 2.04 using 6 selected genes. van't Veer *et al.* (2002) reported 2 classification errors on the test set using 70 relevant genes. There is only one common gene between our 6 selected genes and the 70 relevant genes from van't Veer *et al.* (2002). This is probably due to the fact that 4 out of our 6 selected genes have poor univariate rankings (above 200, see Table 2). In addition, the 70 relevant genes from van't Veer *et al.* (2002) are chosen due to a high correlation (> 0.3 or <-0.3) with the response variable. Some of these high correlation genes may be correlated among themselves. For example, among the top 10 correlated genes (with

the response variable) from the 70-gene subset, four of them have correlation greater than 0.3 with the top ranking gene AL080059.

Our results demonstrate the power of our multivariate BMA gene selection procedure that explores all p given genes: genes with poor univariate rankings may be beneficial in classification when used in combination with other genes. By choosing our relevant genes from sets of genes, the iterative BMA algorithm greatly reduces the number of relevant genes needed for accurate class prediction. Furthermore, these 6 selected genes are used in 13 selected models, each of which consists of 3 to 6 genes. The predicted probabilities for the 19 test samples are illustrated in the uncertainty plot in Figure 2, in which the uncertainty $(1 - Pr(Y=1|D))$ is plotted against the test samples, sorted by increasing uncertainty (Bensmail *et al.* 1997). Figure 2 shows that two out of the three misclassified test samples have high uncertainty, indicating that our assessment of uncertainty does correspond with the errors actually made, as we would wish.

### 3.2    Leukemia data (2 and 3 classes)

The leukaemia dataset (Golub *et al.* 1999) consists of 7129 genes, 38 samples in the training set and 34 samples in the test set. We filtered out genes that do not exhibit significant variation across the training samples, leaving 3051 genes, and then performed thresholding and the logarithmic transformation. The data consist of samples from patients with either Acute Lymphoblastic Leukemia (ALL) or Acute Myeloid Leukemia (AML). However, it has also been noted that the global expression profiles also reflect two ALL sub-types (B-cell and T-cell) (Golub *et al.* 1999). Hence, this dataset can be divided into either 2 or 3 classes (see Tables 3a and 3b).

We first applied the iterative BMA algorithm to the 2-class leukemia data, and achieved a comparable number of classification errors on the test set to other reported results in the literature. Specifically, we observed a Brier Score of 1.5, with 2 classification errors on the test set (out of 34 samples) with 20 selected genes, using nbest=20 and p=1000 top ranked genes[1]. Similarly to what happened with the breast cancer prognosis data, 13 (out of 20) selected genes have poor univariate BSS/WSS rankings (above 200). This dataset is widely used in classification and feature selection papers in the literature. For example, Nguyen and Rocke (2002b) reported 1 to 3 classification errors on the test set using 50

---

[1] Using all 3051 genes yielded unstable models. We observed this unstable model phenomenon on this thresholded dataset (in which expression values are thresholded by 100 and 16000 before applying the logarithmic transformation) only, but not on other unthresholded datasets. This is probably because some genes with low BSS/WSS rankings have many identical thresholded values across the samples leading to singular matrices in our computation.

to 1500 selected genes. They also noted that test sample #66 is consistently misclassified in the microarray community and suggested that the sample might be incorrectly labelled. Sample #66 is one of the two misclassified samples in our results. Our iterative BMA algorithm consistently misclassified sample #66 in all of our experiments using different parameter values (nbest and p). Lee *et al.* (2003) reported one of the most favorable results in the literature, which is 1 classification error using 5 genes. However, it is not clear whether sample #66 was misclassified in their reported results.

Next, we applied our multi-class iterative BMA algorithm to the 3-class leukemia data (AML, ALL-B cell, ALL-T cell). This produced very encouraging results: a Brier Score of 1.5 with 1 classification error on the test set (34 samples), using 15 genes (nbest = 20, p = 1000). Figure 3 shows the uncertainty plot and Table 4 shows the selected genes and their corresponding posterior probabilities. It is interesting that we achieve a similar Brier Score in the 3-class case as in the 2-class case. Six out of the 15 relevant genes were selected from the binary classification problem comparing AML to ALL-B cell (Y=0 vs. Y=1), and nine genes were selected from comparing AML to ALL-T cell (Y=0 vs. Y=2). Nguyen and Rocke (2002a) pooled the training and test sets (38+34=72 samples) and classified each of the 72 samples in turn using the classifier built using the remaining 71 samples. They reported 4 classification errors out of 72 samples with 69 to 100 genes, using leave-one-out cross validation. We not only produced a lower error rate (3% compared to 5.5%) and good Brier Scores, but also our BMA classifier was built using only 38 training samples. Recently, Lee and Lee (2003) also applied the multicategory support vector machine to the training set (with 38 samples) of the 3-class leukaemia data. Their best result is 1 classification error on the test set (with 34 samples) using 40 relevant genes.

### 3.3 Hereditary breast cancer data (3 classes)

Hedenfalk *et al.* (2001) studied the expression patterns of hereditary breast cancer with gene mutations (BRCA1 or BRCA2 mutations). The hereditary breast cancer dataset consists of 7 samples of cancers with BRCA1 mutation, 8 samples with BRCA2 mutation, and 7 sporadic cases of primary breast cancers over 3226 genes. There is no separate test set available, so we use leave-one-out cross validation (LOOCV) in which each of the 22 samples is used in turn as the test sample and a classifier is built using the remaining 21 samples.

We applied the multi-class iterative BMA algorithm to this three-class data, and obtained encouraging results: a Brier Score of 5.5 with 6 classification errors (out of 22 samples) with 13 to 18 relevant genes, using all 3226 genes and nbest=50. Since LOOCV is used, a different classifier is built for each test sample, so the number of relevant genes may vary in each classifier. Nguyen and Rocke

(2002a) reported 6 classification errors with 343 to 438 relevant genes using their proposed partial least squares gene selection method on the same dataset.

## 4   DISCUSSION

We have proposed iterative BMA algorithms for gene selection on binary and multi-class microarray data. Both are multivariate gene selection methods in which dependency between genes is exploited. Our algorithms take advantage of model uncertainty by averaging over multiple models (sets of relevant genes). We demonstrated high prediction accuracy using smaller numbers of genes (relative to other methods) on both binary and multi-class microarray datasets. Table 5 shows an overall summary of our results. In addition, our algorithms produce posterior probabilities for both selected genes and models, and these posterior probabilities aid biological interpretation. We also observed that the selected models are generally very simple, containing only a few genes. Furthermore, we adopted the Brier Score and used the generalized Brier Score to evaluate prediction accuracy, taking the posterior probabilities for the response variables into consideration.

Unlike most feature selection algorithms, in which a pre-specified number (usually small) of top ranked genes are chosen as relevant genes and all the remaining genes are discarded, our Iterative BMA algorithm guarantees that *all p genes* are considered even though the resulting selected genes and models depend on the initial ranking. We show that genes with poor univariate scores may contribute to increased prediction accuracy, and we recommend using all available genes (i.e., $p = G$) in the iterative BMA algorithms, except in the case of thresholded data. From our experiments, nbest=20 or 50 generally yield good results.

In order to efficiently compute a reduced set of good models, we use the leaps and bounds algorithm (Furnival and Wilson 1974), which returns the best "nbest" models for each size up to 30 variables. This imposes a restriction of a 30-variable window on our iterative BMA algorithms, which in turn limits our algorithms to choosing at most 30 relevant genes. Although this restriction does not seem to hurt performance, we are currently in the process of exporting our BMA software from Splus to R, and relaxing this 30-variable limitation. Our current implementation is computationally efficient. For example, it takes under 30 minutes to run our iterative BMA algorithm on the binary breast cancer prognosis dataset (nbest=20 and $p = 4919$) on a moderate computer with a 1.4GHz AMD Athlon processor. Another future project is to study the effect of the chosen baseline (Y=0) in our multi-class iterative BMA algorithm. Our preliminary results show that changing the baseline response variable

does not affect predictive performance much. However, the number of relevant genes chosen can be different.

The combination of high accuracy, small numbers of genes and posterior probabilities for the predictions, should make BMA an attractive tool for developing diagnostics from expression data. The posterior probability of the prediction provides an estimate of the certainty of the classification, which can be useful in a diagnostic setting.

## 5   ACKNOWLEDGEMENTS

## 6   REFERENCES

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X.*, et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503-11.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* **96**: 6745-50.

Begg, C.B. and Gray, R. (1984) Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* **71**: 11-18.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000) Tissue classification with gene expression profiles. *J Comput Biol* **7**: 559-83.

Bensmail, H., Celeux, G., Raftery, A.E. and Robert, C.P. (1997) Inference in model-based cluster analysis. *Statistics and Computing* **7**: 1-10.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M.*, et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**: 13790-5.

Bo, T. and Jonassen, I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol* **3**: RESEARCH0017.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**: 1-3.

Dettling, M. (2004) BagBoosting for Tumor Classification with Gene Expression Data. ETH Zurich, Seminar for Statistics. Research Reports 122. http://stat.ethz.ch/research/research_reports/2004/122

Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* **19**: 1061-9.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97**: 77-87.

Furnival, G.M. and Wilson, R.W. (1974) Regression by Leaps and Bounds. *Technometrics* **16**: 499-511.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A.*, et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-7.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P.*, et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344**: 539-48.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinksky, C. (1999) Bayesian Model Averaging: A Tutorial. *Statistical Science* **14**: 382-417.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*, Wiley: New York.

Jaeger, J., Sengupta, R. and Ruzzo, W.L. (2003) Improved gene selection for classification of microarrays. *Pac Symp Biocomput*: 53-64.

Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M. and Mallick, B.K. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics* **19**: 90-7.

Lee, Y. and Lee, C.K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics* **19**: 1132-9.

Li, W. and Yang, Y. (2002). How many genes are needed for a discriminant microarray data analysis. *Methods of Microarray Data Analysis*. Lin, S.M. and Johnson, K.F., Kluwer Academic**:** 137-150.

Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Internat. Statist. Rev.* **63**: 215-232.

Madigan, D.M. and Raftery, A.E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**: 1335-1346.

Nguyen, D.V. and Rocke, D.M. (2002a) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **18**: 1216-26.

Nguyen, D.V. and Rocke, D.M. (2002b) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**: 39-50.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T.*, et al.* (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* **63**: 1602-7.

Raftery, A.E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology 1995*. Marsden, P.V. Cambridge, Mass., Blackwells**:** 111-196.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P.*, et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**: 15149-54.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M.*, et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**: 227-35.

Schummer, M., Ng, W.V., Bumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y. and Hood, L. (1999) Comparative hybridization of an array of 21500 ovarian cDNAs for the   discovery of genes overexpressed in ovarian carcinomas. *Genes* **238**: 375-385.

Sha, N., Vannucci, M., Tadesse, M.G., Brown, P.J., Dragoni, I., Davies, N., Roberts, T.C., Contestabile, A., Salmon, N., Buckley, C.*, et al.* (2004) Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *To appear in Biometrics.*

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S.*, et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**: 68-74.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**: 6567-72.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**: 5116-21.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T.*, et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530-6.

Viallefont, V., Raftery, A.E. and Richardson, S. (2001) Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* **20**: 3215-3230.

Yeung, K.Y. and Bumgarner, R.E. (2003) Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biol* **4**: R83.

## 7   TABLES

**Table 1 -** Prognosis groups and class sizes of the training set and test set of the breast cancer prognosis data. Y is the response (class) variable.

| Prognosis group | Y | Training set (total 76) | Test set (total 19) |
|---|---|---|---|
| Poor (develop metastases within 5 years) | 0 | 33 | 12 |
| Good (disease free for at least 5 years) | 1 | 43 | 7 |

**Table 2:** Selected genes and their corresponding posterior probabilities of not being equal to zero (probne0), BSS/WSS ranks, and membership in the 70-gene signature chosen by van't Veer *et al.* (2002) for the breast cancer prognosis data using 4919 genes and nbest=20. The genes are shown in descending order of probne0.

| selected genes | probne0(%) | BSS/WSS rank | in 70-gene signature? | gene description |
|---|---|---|---|---|
| AL080059 | 100.0 | 1 | yes | Homo sapiens mRNA; cDNA DKFZp564H142 (from clone DKFZp564H142) |
| Contig49670_RC | 80.8 | 95 | no | Homo sapiens cDNA: FLJ23228 fis, clone CAE06654 |
| NM_012214 | 70.8 | 201 | no | mannosyl (alpha-1,3-)-glycoprotein beta-1,4-N-acetylglucosaminyltransferase, isoenzyme A |
| Contig59951 | 57.3 | 793 | no | RAD21 (S. pombe) homolog |
| Contig46443_RC | 57.3 | 1349 | no | ESTs, Weakly similar to AF279265 1 putative anion transporter 1 [H.sapiens] |
| NM_003315 | 41.4 | 423 | no | tetratricopeptide repeat domain 2 |

**Table 3:** Groups and class sizes of the training and test sets of the leukemia data. Y is the response (class) variable.
a. 2-class (ALL vs. AML)

| Class | Y | Training set (total 38) | Test set (total 34) |
|---|---|---|---|
| ALL (Acute Lymphoblastic Leukemia) | 0 | 27 | 20 |
| AML (Acute Myeloid Leukemia) | 1 | 11 | 14 |

b. 3-class (AML vs. ALL-B cell vs. ALL-T cell)

| Class | Y | Training set (total 38) | Test set (total 34) |
|---|---|---|---|
| AML | 0 | 11 | 14 |
| ALL-B cell | 1 | 19 | 19 |
| ALL-T cell | 2 | 8 | 1 |

**Table 4:** Selected genes and their corresponding posterior probabilities of not being equal to zero (probne0), and BSS/WSS ranks for the 3-class leukemia data using p=1000 genes and nbest=20. The

BSS/WSS ranks represent the ranks in the binary logistic regression (Y=0 vs. Y=1) or (Y=0 vs. Y=2). If a gene is selected in only one binary logistic regression, a blank entry is shown. For example, X03934_at was ranked #1 in the binary regression between AML (Y=0) and ALL-T cell (Y=2), but X03934_at was not selected in the binary regression between AML (Y=0) and ALL-B cell (Y=1). The genes are shown in descending order of probne0.

| selected genes | probne0 (%) | BSS/WSS rank Y=0 vs. 1 | Y=0 vs. 2 | gene description |
|---|---|---|---|---|
| M27891_at | 100.0 | 1 | | CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage) |
| L28821_at | 32.6 | | 279 | MANA2 Alpha mannosidase II isozyme |
| X03934_at | 30.9 | | 1 | GB DEF = T-cell antigen receptor gene T3-delta |
| X59871_at | 30.9 | | 2 | TCF7 Transcription factor 7 (T-cell specific) |
| U02493_at | 18.7 | | 152 | 54 kDa protein mRNA |
| X05323_at | 8.1 | 213 | | OX-2 MEMBRANE GLYCOPROTEIN PRECURSOR |
| Z22551_at | 8.1 | 312 | | Kinectin gene |
| X74008_at | 8.0 | 802 | | PPP1CC Protein phosphatase 1, catalytic subunit, gamma isoform |
| U90552_s_at | 8.0 | 112 | | Butyrophilin (BTF5) mRNA |
| L33075_at | 7.9 | 354 | | Ras GTPase-activating-like protein (IQGAP1) mRNA |
| X99459_at | 6.6 | | 974 | Sigma 3B protein |
| M98539_at | 5.7 | | 523 | Prostaglandin D2 synthase gene |
| M81830_at | 5.7 | | 931 | GB DEF = Somatostatin receptor isoform 2 (SSTR2) gene |
| Y11710_rna1_at | 5.3 | | 972 | Extracellular matrix protein collagen type XIV, C-terminus |
| L32831_s_at | 5.1 | | 1000 | PROBABLE G PROTEIN-COUPLED RECEPTOR GPR3 |

**Table 5:** Summary of our results. The number of relevant genes, Brier Score and the number of classification errors on the test set obtained from our iterative BMA algorithms are shown in column 4. The number of relevant genes and number of classification errors on the test set from published results are shown in column 5. *Results from the hereditary breast cancer data were evaluated using leave-one-out cross validation (LOOCV).

| Dataset | # classes | Classes | Results from our iterative BMA algorithms | Published results |
|---|---|---|---|---|
| Breast cancer prognosis data | 2 | Poor vs. Good prognosis groups | # genes = 6<br>Brier Score = 2.04<br># errors = 3/19 | # genes = 70<br><br># errors = 2/19 |
| Leukemia data | 3 | AML vs. ALL-B cell vs. ALL-T cell | # genes = 15<br>Brier Score = 1.5<br># errors = 1/34 | # genes = 40<br><br># errors = 1/34 |
| Hereditary breast cancer data* | 3 | Sporadic vs. BRCA1 vs. BRCA2 | # genes = 13 to 18<br>Brier Score = 5.5<br># errors = 6/22 | # genes = 343 to 438<br><br># errors = 6/22 |

## 8  FIGURES

**Figure 1:** A flowchart illustrating the multi-class iterative BMA algorithm for K=3. Suppose two genes $x_1$ and $x_2$ are selected in the two binary logistic regressions (Y=0 vs. Y=1 and Y=0 vs. Y=2) from the iterative BMA algorithm. The goal of polychotomous regression is to estimate the regression parameters for $g_1(x)=\ln[P(Y=1|D)/P(Y=0|D)]=b_{10} + b_{11}x_1 + b_{12}x_2$ and $g_2(x)= \ln[P(Y=2|D)/P(Y=0|D)]=b_{20} + b_{21}x_1 + b_{22}x_2$. The augmented matrix M consists of an intercept column ($b_{20} - b_{10}$) and a column for each regression parameter $b_{11}$, $b_{12}$, $b_{21}$, and $b_{22}$.
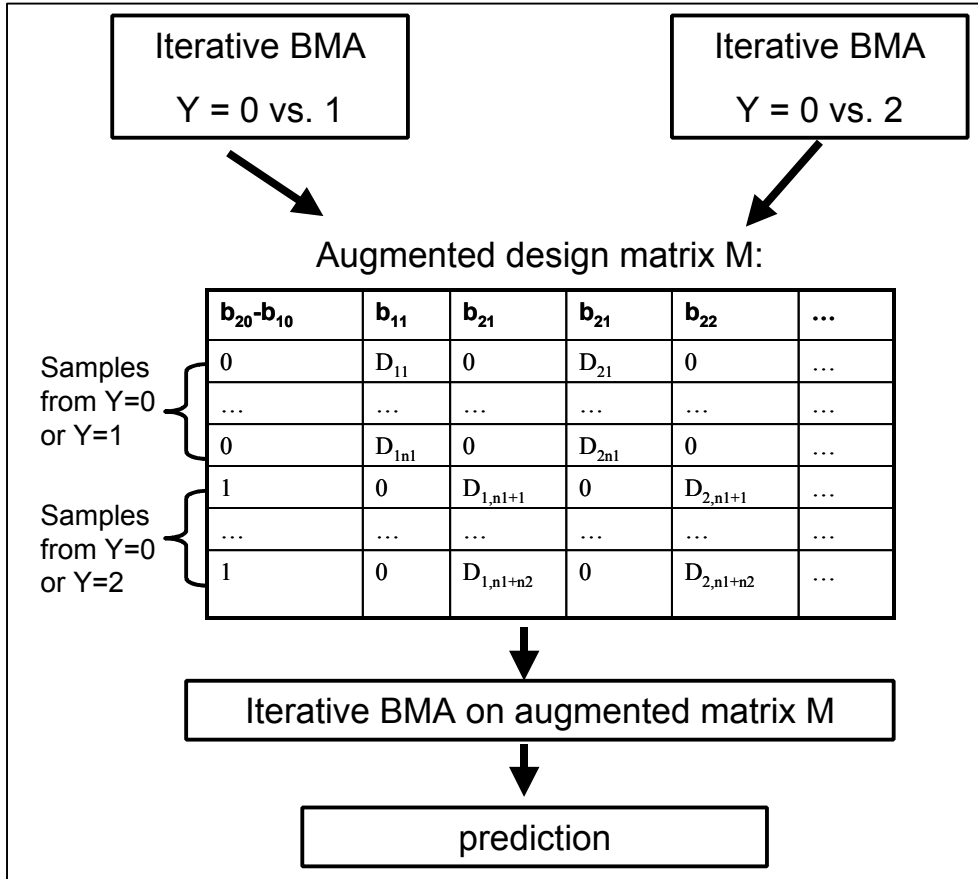
**Figure 2:** Uncertainty plot for the predicted probabilities on the test set (19 samples) of the breast cancer prognosis data. The y-axis represents the uncertainty (1 – predicted probability of Y=1), and the x-axis represents the 19 test samples sorted in increasing order of uncertainty. The follow-up time of patients is used to label the upper x-axis. The vertical bars represent classification errors. In other words, test samples # 102, 117, 109 with follow-up times 3.3, 5.3, 3.2 respectively were misclassified.
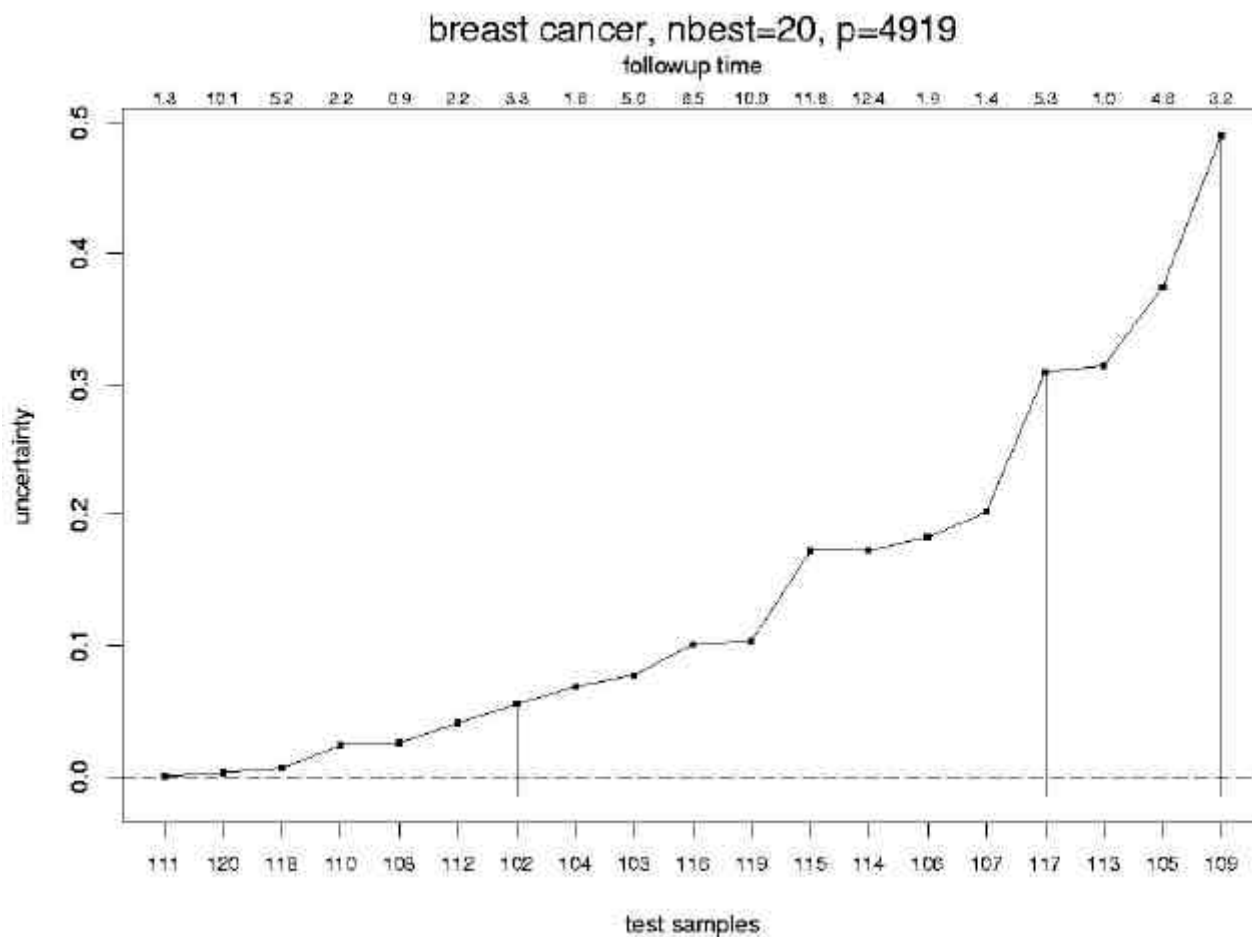


breast cancer, nbest=20, p=4919

**Figure 3:** Uncertainty plot for the predicted probabilities on the test set (34 samples) of the 3-class leukemia data. Each sample is classified as being in the class j with the maximum predicted probability Pr(Y=j|D), where j=0, 1, 2. The y-axis represents the uncertainty (1 – maximum predicted probability), and the x-axis represents the 34 test samples sorted in increasing order of uncertainty. The vertical bar represents a misclassified sample.



ALL AML 3 class, nbest=20, p=1000