

Supplementary Materials:

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

Ka Yee Yeung¹, Roger E. Bumgarner¹ and Adrian E. Raftery²

¹Department of Microbiology, University of Washington, Seattle, WA 98195

²Department of Statistics, University of Washington, Seattle, WA 98195

Table of Contents

A. Results on the Breast Cancer Prognosis data

B. Results on the 2-class Leukemia data

C. Results on the 3-class Leukemia data

D. Results on the BRCA Hereditary Breast Cancer data

E. Algorithmic and implementational details

1. BMA

2. iterative BMA

3. Additional modifications to the Iterative BMA algorithm: interaction terms, wrap around

4. multi-class iterative BMA

A. Results on the Breast Cancer Prognosis data

1. Brier Score and # classification errors on the test set over different p and different nbest

Table A.1: The Brier Score, number of classification errors, number of selected genes and number of selected models are shown on the test set (19 samples) of the breast cancer prognosis data. "nbest" is the number of best models for each size (up to 30 variables) returned by the leaps and bounds algorithm.

nbest	probne0	p	# errors	Brier Score	# genes	# models
10	1	10	6	4.27	10	36
10	1	30	9	3.43	3	4
10	1	50	6	4.53	5	9
10	1	100	4	3.00	4	2
10	1	500	7	4.51	7	12
10	1	1000	5	3.77	2	2
10	1	2000	4	2.53	3	4
10	1	4919		NA		
20	1	10	6	4.27	10	36
20	1	30	8	3.78	6	15
20	1	50	6	4.01	2	1
20	1	100	4	2.02	8	11
20	1	500	4	2.48	7	10
20	1	1000	2	1.77	6	11
20	1	2000	2	2.12	3	4
20	1	4919	3	2.04	6	13
50	1	10	6	4.27	10	36
50	1	30	5	4.20	7	18
50	1	50	5	4.35	4	4
50	1	100	4	2.80	6	8
50	1	500	4	3.03	9	8
50	1	1000	3	2.22	7	3
50	1	4919	3	2.23	5	2
150	1	10	6	4.27	10	36
150	1	30	5	4.35	4	4
150	1	50	5	4.35	4	4
150	1	100	4	2.80	6	8
150	1	500	4	3.03	9	8
150	1	1000	3	2.22	7	3
150	1	4919	3	2.22	7	3

Observations:

- The number of classification errors and Brier Scores are usually improved (reduced) when p (the number of top genes considered by BMA) is increased.
- The results (in terms of Brier Scores and number of errors on the test set) are relatively insensitive to nbest (nbest = 20, 50 or 150 yield similar results).

- The number of genes selected is relatively few. For example, only 6 genes are selected for $n_{best}=20$, $p=4919$ and $prob_{ne0}=1\%$.

2. Detailed results for $n_{best}=20$ and $p=G=4919$

When $n_{best}=20$ and $p=4919$, Brier Score = 2.04 and the number of classification errors = 3 on the test set (19 samples).

Table A.2.a: Selected genes and their corresponding BSS/WSS ranks, posterior probabilities and membership in the 70-gene signature chosen by [van't Veer et al. 2002]

selected genes	prob _{ne0} (%)	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 A.2.b: Models selected (total 13 models using 6 genes) and their corresponding posterior probabilities sorted in descending order of model posterior probabilities.

Model #	Genes in the model (genes are represented using Affy Accession numbers)	Model posterior probability
1	AL080059,Contig59951,NM_003315,Contig46443_RC	26.6%
2	AL080059,NM_012214,Contig49670_RC	13%
3	AL080059,Contig59951,NM_003315,Contig46443_RC,NM_012214	9.4%
4	AL080059,Contig59951,NM_012214,Contig49670_RC	9.1%
5	AL080059,Contig59951,NM_003315,NM_012214,Contig49670_RC	7.1%
6	AL080059,Contig59951,NM_003315,Contig46443_RC,Contig49670_RC	6.8%
7	AL080059,Contig59951,NM_003315,Contig46443_RC,NM_012214,Contig49670_RC	6.4%
8	AL080059,Contig59951,Contig46443_RC,NM_012214,Contig49670_RC	5.2%
9	AL080059,NM_003315,NM_012214,Contig49670_RC	5%
10	AL080059,Contig46443_RC,NM_012214,Contig49670_RC	4.5%
11	AL080059,Contig59951,NM_003315,NM_012214	2.8%
12	AL080059,NM_003315,Contig46443_RC,NM_012214,Contig49670_RC	2.2%
13	AL080059,Contig59951,NM_003315	1.9%

Figure A.2.c: Uncertainty plot for the predicted probabilities on the test set (19 samples). The y-axis represents the uncertainty (1 – predicted probability of $Y=1$), and the x-axis represents the 19 test samples sorted in increasing uncertainties. The followup 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 mis-classified.

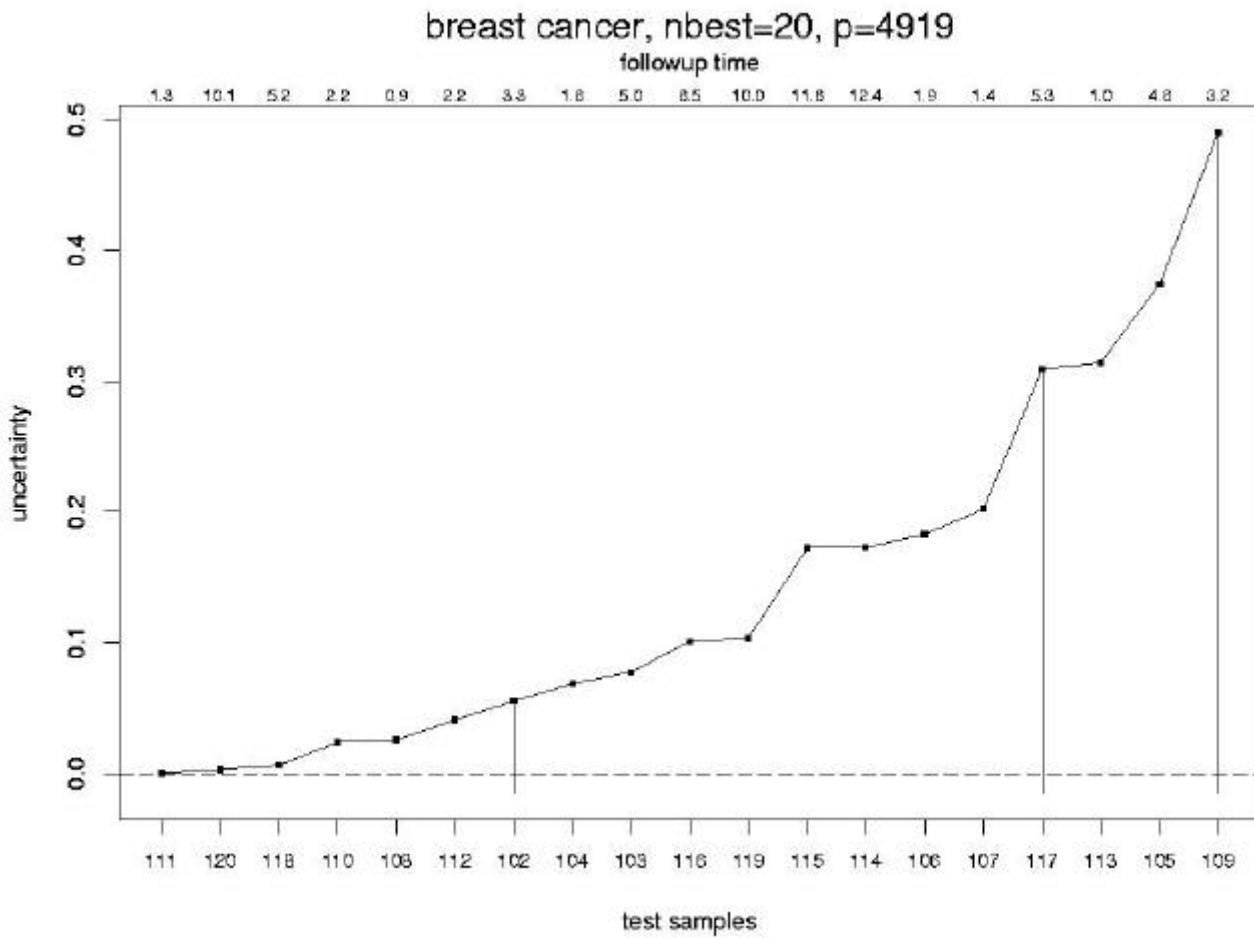


Table A.2.d: Predicted probabilities of Y=1 when nbest=20, p=1000 or 4919.

sample no.	truth	p=1000	p=4919	followup time(yr)
Poor111	0	0.001	0.000	1.27
Poor112	0	0.046	0.041	2.25
Poor113	0	0.209	0.314	1.00
Good114	1	0.763	0.827	12.43
Good115	1	0.926	0.827	11.59
Good116	1	0.896	0.899	8.54
Good117	1	0.435	0.309	5.30
Good118	1	0.991	0.994	5.23
Good119	1	0.788	0.896	9.98
Good120	1	0.997	0.997	10.10
Poor102	0	0.963	0.945	3.29
Poor103	0	0.073	0.078	4.95
Poor104	0	0.141	0.069	1.84
Poor105	0	0.478	0.375	4.77
Poor106	0	0.194	0.183	1.92
Poor107	0	0.036	0.202	1.37
Poor108	0	0.098	0.026	0.89
Poor109	0	0.242	0.510	3.20
Poor110	0	0.027	0.024	2.17
Brier Score		1.77	2.04	

Table A.2.e: Comparing $\Pr(Y=1|D)$ using all 13 selected models and the highest posterior probability model (26.6% using AL080059, Contig59951, NM_003315, Contig46443_RC) when nbest=20, p=4919. The mis-classifications on the test set (19 samples) are highlighted in yellow. Samples #113 and #107 were mis-classified using the top model with predicted posterior probabilities 0.566 and 0.540 respectively, but were correctly classified using all 13 selected models.

class	all models	top model	survival time	
0	0.000	0.000	1.27	Poor111
0	0.041	0.003	2.25	Poor112
0	0.314	0.566	1.00	Poor113
1	0.827	0.995	12.43	Good114
1	0.827	0.691	11.59	Good115
1	0.899	0.840	8.54	Good116
1	0.309	0.428	5.30	Good117
1	0.994	0.999	5.23	Good118
1	0.896	0.910	9.98	Good119
1	0.997	0.995	10.10	Good120
0	0.945	0.958	3.29	Poor102
0	0.078	0.050	4.95	Poor103
0	0.069	0.067	1.84	Poor104
0	0.375	0.169	4.77	Poor105
0	0.183	0.173	1.92	Poor106
0	0.202	0.540	1.37	Poor107
0	0.026	0.004	0.89	Poor108
0	0.510	0.918	3.20	Poor109
0	0.024	0.034	2.17	Poor110
# errors	3	5		
Brier score	2.040	2.895		

3. Detailed results for nbest=20 and p=1000

When nbest=20 and p=1000, Brier Score = 1.77 and the number of classification errors = 2 on the test set (19 samples).

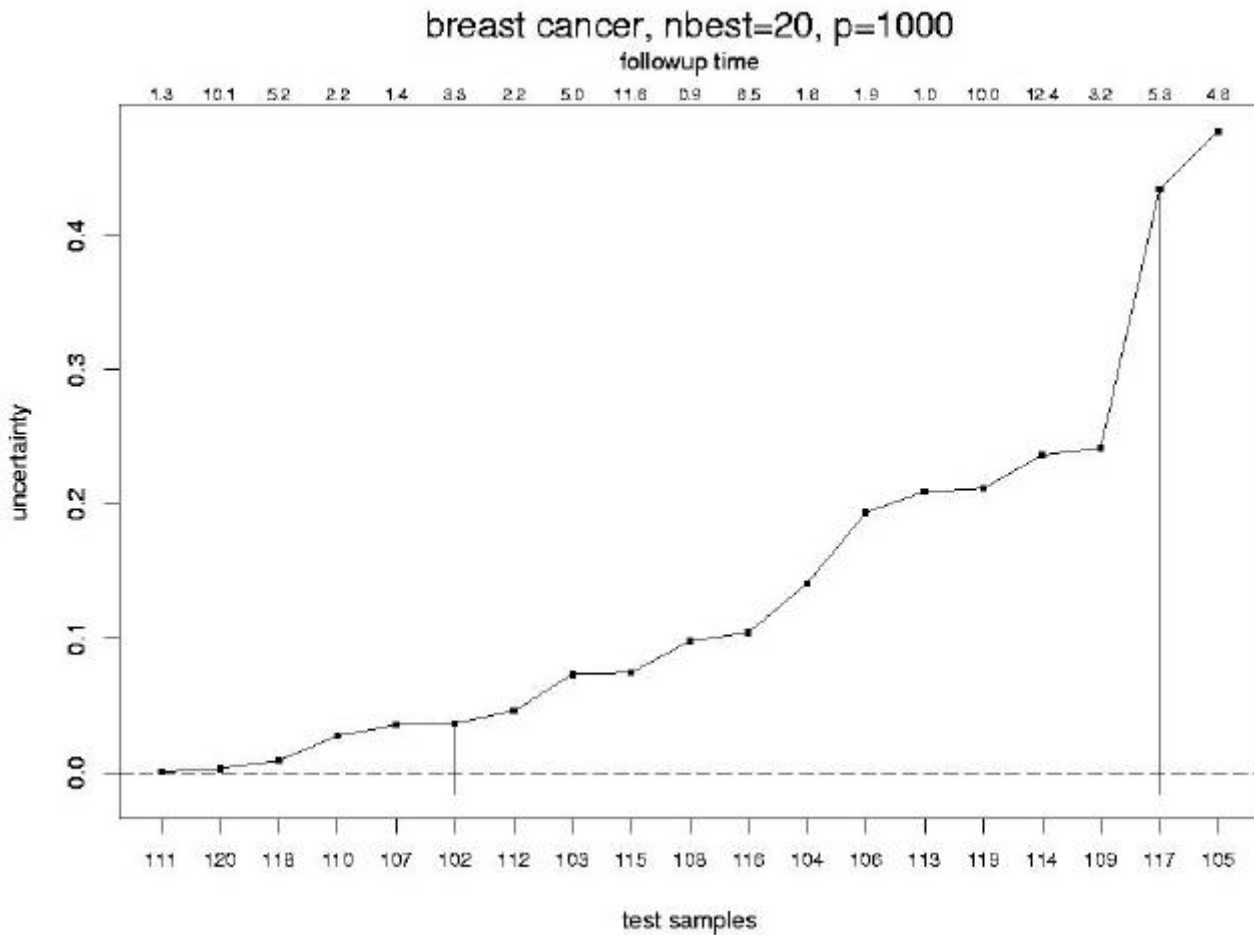
Table A.3.a: Selected genes and their corresponding BSS/WSS ranks, posterior probabilities and membership in the 70-gene signature chosen by van't Veer et al.

selected genes	probne0 (%)	BSS/WSS rank	in 70-gene signature?	gene description
AL080059	100	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
NM_003315	57.3	423	no	tetratricopeptide repeat domain 2
Contig55979_RC	41.4	772	no	ESTs

Table A.3.b: Models selected (total 11 models using 6 genes) and their corresponding posterior probabilities sorted in descending order of model posterior probabilities.

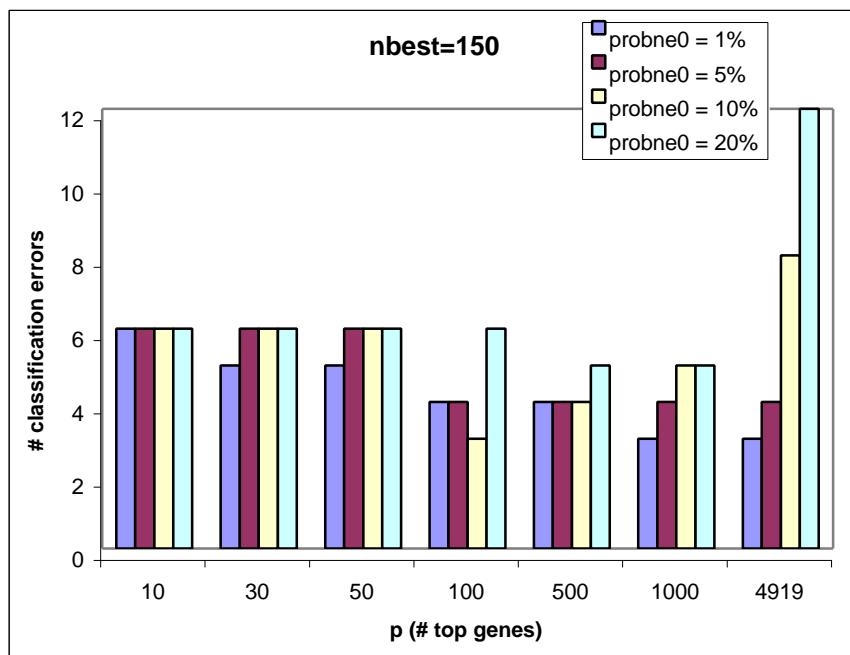
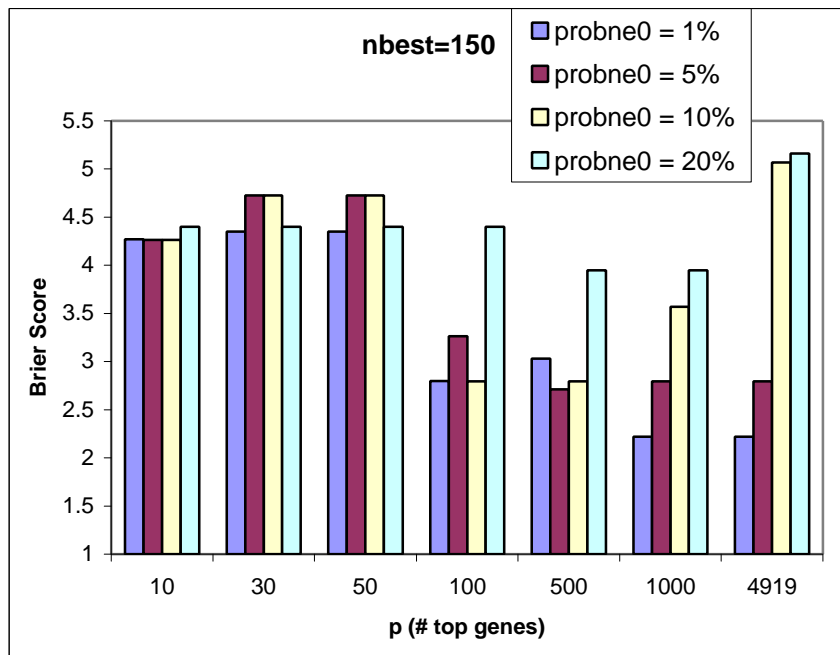
Model #	Genes in the model	Model posterior probability
1	AL080059,NM_012214,Contig49670_RC	19.1%
2	AL080059,Contig59951,NM_012214,Contig49670_RC	13.5%
3	AL080059,Contig59951,Contig55979_RC,NM_003315,Contig49670_RC	13%
4	AL080059,Contig59951,Contig55979_RC,NM_003315	12.3%
5	AL080059,Contig59951,NM_003315,NM_012214,Contig49670_RC	10.5%
6	AL080059,Contig55979_RC,NM_012214,Contig49670_RC	10.1%
7	AL080059,NM_003315,NM_012214,Contig49670_RC	7.3%
8	AL080059,Contig55979_RC,NM_003315,NM_012214,Contig49670_RC	6.1%
9	AL080059,Contig59951,NM_003315,NM_012214	4.1%
10	AL080059,Contig59951,NM_003315	2.8%
11	AL080059,Contig59951,NM_003315,Contig49670_RC	1.1%

Figure A.3.c: Uncertainty plot for the predicted probabilities on the test set (19 samples). The y-axis represents the uncertainty (1 – predicted probability of Y=1), and the x-axis represents the 19 test samples sorted in increasing uncertainties. The followup time of patients is used to label the upper x-axis. The vertical bars represent classification errors. In this case, test samples #102 and #117 with follow-up times 3.3 and 5.3 respectively were mis-classified.



4. Effects of parameters on predictive performance

Figure A.4: Effect of varying parameters p (# top univariate genes) and $probne0$ threshold when $nbest$ is fixed at 150 on the breast cancer prognosis data in terms of the Brier Score and classification errors on the test set of the breast cancer prognosis data. Both the Brier Score and the number of classification errors are reduced (improved) when the number of top univariate genes (p) is high (more than 1000) and when $probne0$ threshold = 1%. We observed similar results at different values of $nbest$.



B. Results on the 2-class Leukemia data

1. Predictive performance over different p and different nbest

Table B.1: Brier Score and # classification errors on the test set (out of 34 samples) of the 2-class leukemia data. Columns 4-7 correspond to results from the iterative BMA algorithm without the adaptive thresholding step, hence not all p genes are guaranteed to be considered in the algorithm. Columns 8-11 correspond to results from the iterative BMA algorithm with the adaptive thresholding step, hence all p genes are considered.

nbest	probne0	p					<u>Adaptive Threshold</u>				
			# errors	Brier Score	# genes	# models	# errors	Brier Score	# genes	# models	
10	1	10	2	1.9289	8	19					
10	1	50	3	1.9247	14	35					
10	1	100	3	1.8489	11	29					
10	1	500	2	1.6026	11	19					
10	1	1000	2	1.5637	11	20					
10	1	2000	2	1.5317	10	18					
10	1	3051	NA unstable								
20	1	10	2	1.8475	10	26					
20	1	50	1	2.2232	19	74					
20	1	100	1	1.8892	18	73					
20	1	500	2	1.6213	20	45					
20	1	1000	2	1.4967	20	38					
20	1	2000	NA unstable								
20	1	3051	NA unstable								
50	1	10	2	1.9635	27	111					
50	1	50	2	1.9635	27	111					
50	1	100	3	1.8129	28	110					
50	1	500	2	1.8202	30	108	2	1.701	26	90	
50	1	1000	2	1.8202	30	108	2	1.635	25	83	
50	1	2000	2	1.8202	30	108	NA unstable				
50	1	3051	2	1.8202	30	108	NA unstable				
100	1	10	2	1.8475	10	26					
100	1	50	2	1.8528	27	117	2	1.827	30	111	
100	1	100	2	1.8528	27	117	1	1.759	29	119	
100	1	500	2	1.8528	27	117	1	1.69	28	121	
100	1	1000	2	1.8528	27	117	1	1.758	27	120	
100	1	2000	2	1.8528	27	117	NA unstable				
100	1	3051	2	1.8528	27	117	NA unstable				
150	1	10	2	1.8475	10	26					
150	1	50	2	1.9779	30	118	1	1.691	29	112	
150	1	100	2	1.9779	30	118	1	1.603	29	121	
150	1	500	2	1.9779	30	118	2	1.578	29	117	
150	1	1000	2	1.9779	30	118	1	1.598	23	105	
150	1	2000	2	1.9779	30	118	2	1.56	26	108	
150	1	3051	2	1.9779	30	118	NA unstable				

Observations:

- If the adaptive threshold procedure is not used, iterative BMA cannot process all the given genes when nbest is high (50, 100, 150).
- Iterative BMA with adaptive threshold guarantees that all genes are processed and generally produces better results (lower Brier scores and reduced number of classification errors on the test set).
- No results from the adaptive threshold procedure are shown for nbest=10 or 20 because all genes are considered in iterative BMA without increasing the probne0 threshold.
- When nbest = 20, the predictive performance is similar to higher nbest values while the number of selected genes and models are lower.

2. Detailed results for nbest=20 and p=1000

When nbest=20 and p=1000, Brier Score = 1.5 and the number of classification errors = 2 on the test set (34 samples).

Table B.2.a: Selected genes and their corresponding BSS/WSS ranks, and posterior probabilities.

selected genes	BSS/WSS rank	post prob (%)
M27891_at	1	44.5
X62535_at	83	4.7
M13792_at	70	4.7
U94319_at	988	4.6
J04132_at	875	4.6
X95735_at	7	57.8
M16336_s_at	681	4.5
U27460_at	185	4.7
U23852_s_at	552	4.7
L10373_at	379	4.6
S54005_s_at	618	4.7
X66401_cds1_at	252	4.7
Z14982_rna1_at	393	4.7
X59871_at	478	4.6
X03934_at	748	4.7
D83920_at	596	4.6
L05148_at	180	4.7
L01087_at	964	4.6
L07633_at	589	4.7
J05243_at	29	4.7

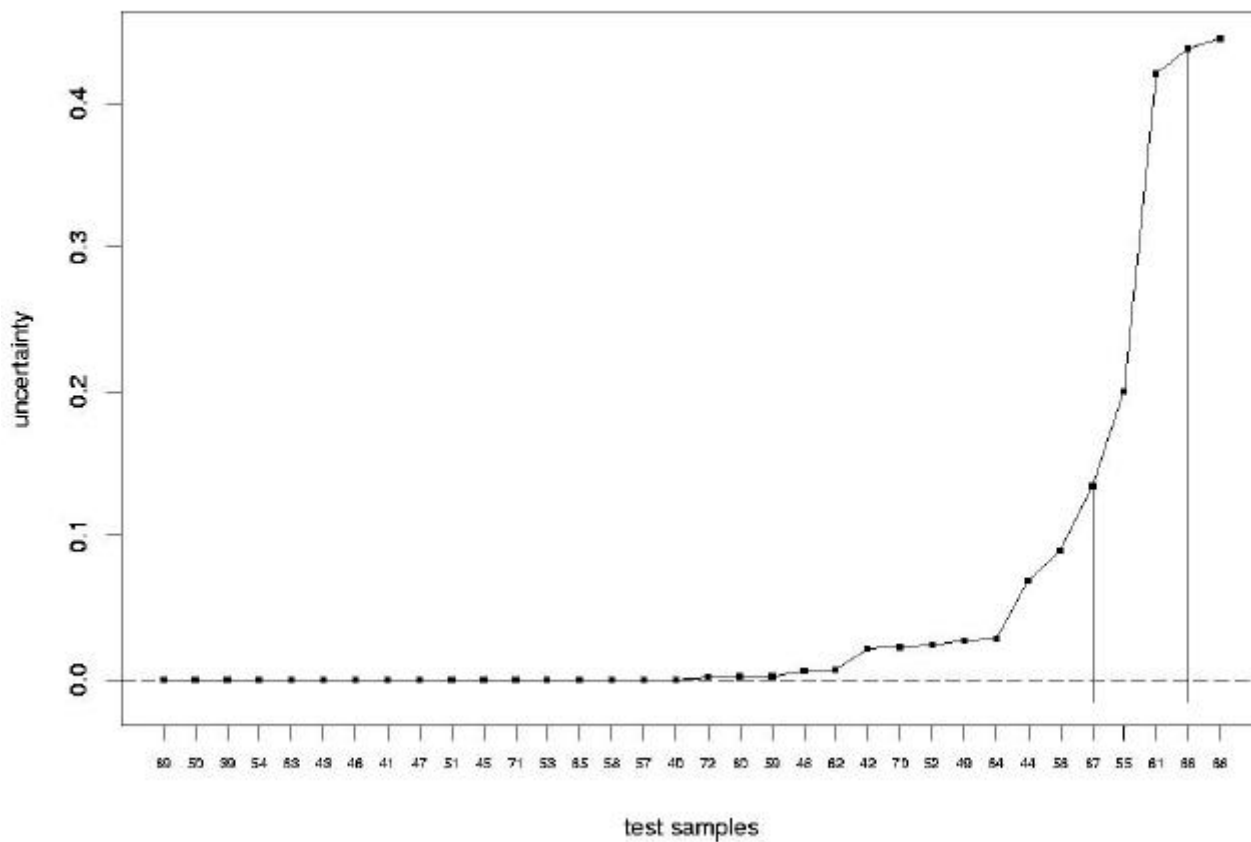
Table B.2.b: Models selected (total 38 models using 20 genes) and their corresponding posterior probabilities sorted in descending order of model posterior probabilities.

Model #	Genes in the model	Model post prob
1	X95735_at	13.8%
2	X95735_at,Z14982_rna1_at	2.4%
3	M27891_at,U27460_at	2.4%
4	M27891_at,J05243_at	2.4%
5	M27891_at,U23852_s_at	2.4%

6	M27891_at,X03934_at	2.4%
7	M27891_at,L05148_at	2.4%
8	M27891_at,X62535_at	2.4%
9	X95735_at,L07633_at	2.4%
10	X95735_at,J05243_at	2.4%
11	M27891_at,J04132_at	2.4%
12	X62535_at,X95735_at	2.4%
13	M27891_at,X59871_at	2.4%
14	M27891_at,L10373_at	2.4%
15	X95735_at,X66401_cds1_at	2.4%
16	M27891_at,X95735_at	2.4%
17	M27891_at,D83920_at	2.4%
18	M27891_at,Z14982_rna1_at	2.4%
19	X95735_at,U27460_at	2.4%
20	M27891_at,S54005_s_at	2.4%
21	X95735_at,L05148_at	2.3%
22	M27891_at,M13792_at	2.3%
23	M13792_at,X95735_at	2.3%
24	U94319_at,X95735_at	2.3%
25	M27891_at,L01087_at	2.3%
26	X95735_at,U23852_s_at	2.3%
27	X95735_at,X03934_at	2.3%
28	X95735_at,S54005_s_at	2.3%
29	M27891_at,L07633_at	2.3%
30	M27891_at,X66401_cds1_at	2.3%
31	M27891_at,M16336_s_at	2.3%
32	X95735_at,L10373_at	2.3%
33	M27891_at,U94319_at	2.3%
34	X95735_at,D83920_at	2.3%
35	X95735_at,X59871_at	2.2%
36	X95735_at,L01087_at	2.2%
37	X95735_at,M16336_s_at	2.2%
38	J04132_at,X95735_at	2.2%

Table B.2.c: Uncertainty plot for the predicted probabilities on the test set (34 samples). The y-axis represents the uncertainty ($1 - \text{predicted probability of } Y=1$), and the x-axis represents the 34 test samples sorted in increasing uncertainties. The vertical bars represent classification errors. In this case, test samples #57 and 66 are mis-classified.

ALL AML 2 class, nbest=20, p=1000



C. Results on the 3-class Leukemia data

1. Brier Score and # classification errors on the test set over different p and different nbest

nbest	probne0	p	# errors	Brier Score	# genes	# models
10	1	10	5	4.946	11	49
10	1	50	6	4.502	11	50
10	1	100	4	3.135	11	66
10	1	500	4	3.063	10	62
10	1	1000	3	3.059	12	75
10	1	2000	NA in both binary logistic regression			
10	1	3051	NA in both binary logistic regression			
20	1	10	6	4.970	15	68
20	1	50	3	3.249	16	107
20	1	100	5	4.204	18	142
20	1	500	1	2.137	17	57
20	1	1000	1	1.496	15	46
20	1	2000	NA in both binary logistic regression			
20	1	3051	NA in both binary logistic regression			
50	1	10	5	3.974	20	102
50	1	50	4	3.477	29	137
50	1	100	3	2.829	21	123
50	1	500	2	1.791	20	111
50	1	1000	3	2.954	29	110
50	1	2000	NA in both binary logistic regression			
50	1	3051	NA in both binary logistic regression			
100	1	10	5	3.974	20	102
100	1	50	5	3.569	28	149
100	1	100	4	4.037	26	144
100	1	500	3	3.135	29	126
100	1	1000	3	2.842	26	125
100	1	2000	NA in both binary logistic regression			
100	1	3051	NA in both binary logistic regression			
150	1	10	5	3.974	20	102
150	1	50	4	3.371	28	147
150	1	100	1	1.826	28	117
150	1	500	2	2.195	28	111
150	1	1000	5	4.920	29	123
150	1	2000	NA in both binary logistic regression			
150	1	3051	NA in both binary logistic regression			

Observations:

- Results are similar to those from the 2-class case.
- Nbest=20 gives low Brier scores and # classification errors without choosing many genes.
- It is amazing that the number of classification errors and Brier scores are similar to the 2-class case.

2. Detailed results for nbest=20 and p=1000

When $n_{best}=20$ and $p=1000$, Brier Score = 1.5 and the number of classification errors = 1 on the test set (34 samples).

Table C.1.a: Selected genes and their corresponding BSS/WSS ranks, and posterior probabilities.

selected genes	probne0 (%)	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 C.1.b: Uncertainty plot for the predicted probabilities on the test set (34 samples). The y-axis represents the uncertainty ($1 - \text{max predicted probability}$), and the x-axis represents the 34 test samples sorted in increasing uncertainties. The vertical bars represent classification errors. In this case, test sample #61 is mis-classified.

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

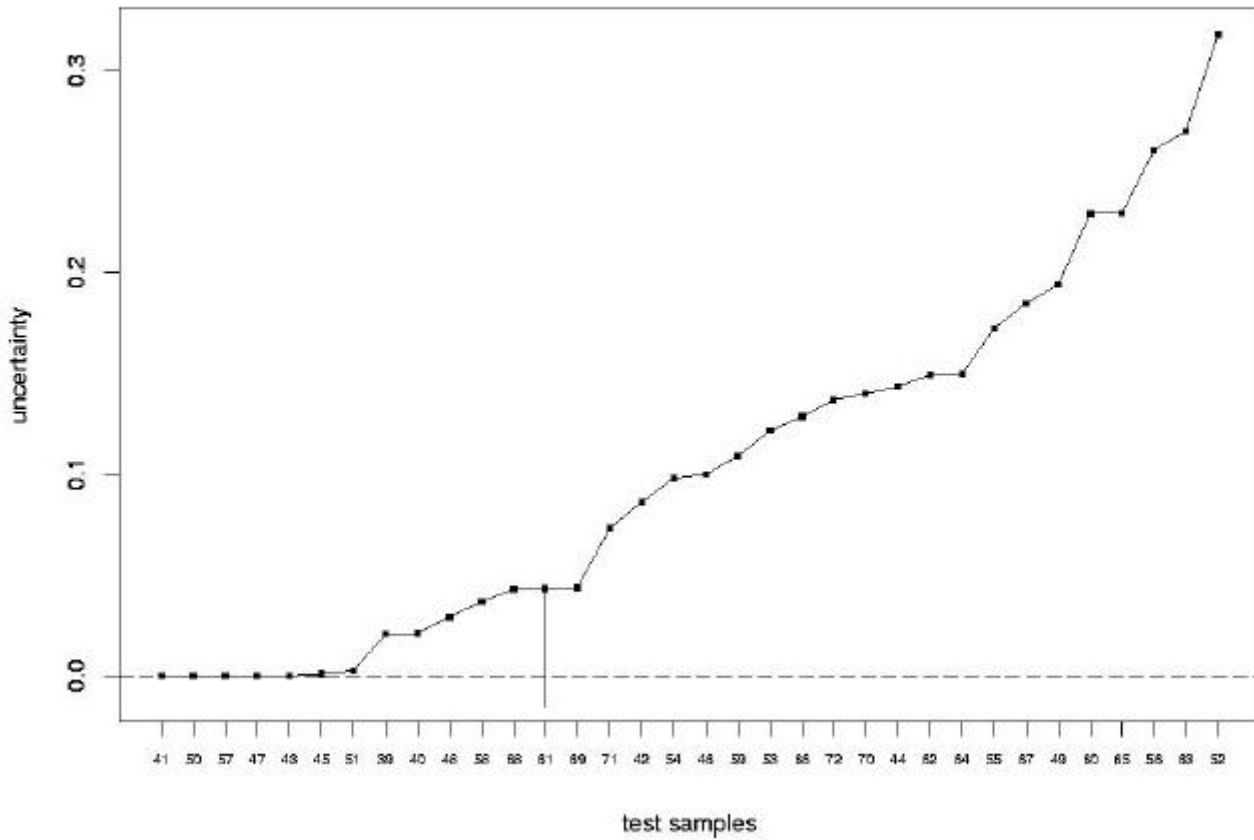


Table C.1.c: Detailed predicted probabilities for each class. Each sample is assigned to the class with the highest predicted posterior probability.

test sample #	predicted probabilities			predicted class	truth
	AML Y=0	ALL-B cell Y = 1	ALL-T cell Y = 2		
39	0.02	0.98	0.00	1	1
40	0.02	0.98	0.00	1	1
42	0.04	0.91	0.04	1	1
47	0.00	1.00	0.00	1	1
48	0.00	0.97	0.03	1	1
49	0.13	0.81	0.06	1	1
41	0.00	1.00	0.00	1	1
43	0.00	1.00	0.00	1	1
44	0.06	0.86	0.09	1	1
45	0.00	1.00	0.00	1	1
46	0.02	0.90	0.08	1	1
70	0.05	0.86	0.09	1	1

71	0.00	0.93	0.07	1	1
72	0.06	0.86	0.08	1	1
68	0.00	0.96	0.04	1	1
69	0.00	0.96	0.04	1	1
67	0.14	0.04	0.82	2	2
55	0.17	0.83	0.00	1	1
56	0.26	0.74	0.00	1	1
59	0.02	0.89	0.08	1	1
52	0.68	0.02	0.30	0	0
53	0.88	0.00	0.12	0	0
51	1.00	0.00	0.00	0	0
50	1.00	0.00	0.00	0	0
54	0.90	0.04	0.06	0	0
57	1.00	0.00	0.00	0	0
58	0.96	0.00	0.04	0	0
60	0.77	0.00	0.23	0	0
61	0.04	0.96	0.00	1	0
65	0.77	0.00	0.23	0	0
66	0.87	0.02	0.11	0	0
63	0.73	0.02	0.25	0	0
64	0.85	0.13	0.02	0	0
62	0.85	0.02	0.13	0	0

D. Results on the BRCA Hereditary Breast Cancer data

1. Brier Score and # classification errors on the test set over different p and different nbest

Table D.1: The number of classification errors (out of 22 samples) and Brier Score are shown for different nbest and different p on the hereditary breast cancer data. Since there is no test set for this data, the classifiers were evaluated using leave-one-out cross validation (LOOCV). Since a different classifier is built for each test sample, the number of relevant genes and selected models may vary across classifiers. Therefore, we show the minimum, average and maximum number of relevant genes and selected models for each nbest and p.

nbest	probne0	p	# errors	Brier Score	# genes			# models		
					min	avg	max	min	avg	max
20	1	10	17	10.008	7	11.545	17	27	86.545	172
20	1	50	8	7.193	13	16.227	20	110	145.318	182
20	1	100	9	7.046	14	17.318	20	123	154.864	190
20	1	500	13	7.392	13	16.409	21	135	155.545	182
20	1	1000	11	7.155	14	16.318	19	132	157.182	181
20	1	2000	9	6.987	13	15.273	18	127	154.091	185
20	1	3226	13	7.311	11	14.818	17	121	152.773	185
50	1	10	16	8.956	6	14.818	20	16	105.136	184
50	1	50	9	6.802	16	18.591	21	118	149.545	190
50	1	100	8	6.270	16	17.955	20	109	154.500	207
50	1	500	5	5.571	15	17.045	20	111	151.364	186
50	1	1000	8	5.907	14	17.000	20	113	148.682	190
50	1	2000	5	5.343	14	16.273	19	118	149.955	180
50	1	3226	6	5.477	13	16.000	18	112	144.727	181
150	1	10	16	9.103	16	18.591	20	43	96.182	184
150	1	50	11	7.076	16	21.455	22	121	158.818	195
150	1	100	10	6.344	19	21.318	22	103	172.318	210
150	1	500	10	5.949	18	21.227	22	144	179.273	201
150	1	1000	8	5.695	18	20.955	22	132	171.500	201

Observations:

- Nbest=50 generally yields the best results in terms of Brier Scores and # classification errors.

E. Algorithmic and implementational details

1. BMA

The idea of BMA is to take model uncertainty into consideration by averaging over the product of the posterior probability given a model M_k , and the posterior probability of the model M_k , summing over a set of models M .

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

The posterior probability for model M_k is given by:

$$\Pr(M_k | D) = \frac{\Pr(D | M_k) \Pr(M_k)}{\sum_{l \in M} \Pr(D | M_l) \Pr(M_l)} \quad \text{where } \Pr(D | M_k) = \int \Pr(D | \mathbf{q}_k, M_k) \Pr(\mathbf{q}_k | M_k) d\mathbf{q}_k$$

$\Pr(D|M_k)$ is the integrated likelihood of model M_k , and θ_k represents the vector of regression parameters (b_0, b_1, \dots, b_p) of model M_k .

We adopted the BMA implementation for generalized linear models (“bic.glm”) from (Volinsky; Raftery 1995) in which a prior of 0.5 is used on all variables by default.

We used logistic regression to model $\Pr(Y=1|D, M_k)$ in which $\ln[\Pr(Y=1|D, M_k)/ \Pr(Y=0|D, M_k)] = b_0 + b_1x_1 + \dots + b_px_p$, where x_i 's represent the expression levels of selected genes and b_i 's are the regression parameters

Outline of bic.glm (written by Adrian Raftery and Chris Volinsky in S+)

Input: Data, parameter “nbest”

1. Fit all p given variables (genes)
2. If # variables > 30, then use stepwise backward elimination to reduce # variables to 30
3. Use the leaps and bounds algorithm (Furnival and Wilson 1974) to enumerate the best “nbest” models up to 30 variables.
4. Use the Occam’s Window method and BIC to eliminate models that are not supported by the data.

Output:

- posterior probability for each selected model
- Probne0: posterior probability that each variable is non-zero
- Maximum likelihood estimates of regression parameter b_i 's

2. Binary iterative BMA algorithm

An outline of the algorithm is described in the manuscript. In our implementation, we have a user-specified parameter “maxIter” which represents the maximum number of iterations applying the traditional BMA algorithm (bic.glm) to the dataset. This is necessary because of S+’s memory management problem: S+ is unhappy if we go through the “while” loop too many times. So, we usually set “maxIter” to 100 or 200 depending on the size of the dataset. If we are unable to finish processing all p genes and have reached “maxIter” iterations, we spit out the temporary results, which will be read in next time when the procedure is re-started.

Note that the number of iterations of the “while” loop depends on the number of genes removed (i.e. number of genes with probne0 < 1% or under the adaptive threshold).

Prediction

Suppose our iterative BMA algorithm has selected genes g_1, g_2, \dots, g_q as the relevant genes. Assume that these genes have expression levels/ratios x_1, x_2, \dots, x_q . We extracted the maximum likelihood

estimates of the regression parameters for these selected genes ($b_1^k, b_2^k, \dots, b_q^k$) of each model k from the output of the iterative BMA algorithm. The posterior predicted probability of $Y=1$ is:

$$\Pr(Y = 1 | D) = \sum_{k \in M} \left[\frac{\exp(\sum b_i^k x_i)}{1 + \exp(\sum b_i^k x_i)} \right] * \Pr(M_k | D)$$

It follows that $\Pr(Y=0|D) = 1 - \Pr(Y=1|D)$.

3. Additional modifications to the iterative BMA algorithm

We have attempted various ways to further improve the performance of the binary iterative BMA algorithm. Here is a brief description of our attempts:

- Interaction terms
- Wrap around

Interaction terms

We use logistic regression without any interaction or higher order terms to model $\Pr(Y=1|D)$. We experimented with adding interaction and second order terms. However, our preliminary results show that adding interaction and second order terms does not improve predictive performance. In particular, after a set of relevant genes S are selected using our iterative BMA algorithm, we create new variables.

- For each gene g_i in S , add residuals ($\text{lm}(g_i^2 \sim g_i)$)
- For each pair of genes g_i and g_j , add residuals ($\text{lm}(g_i * g_j \sim g_i + g_j)$)

Then, apply the iterative BMA algorithm to the input data with the above additional variables. We also tried different methods of ordering these new variables before applying the iterative BMA algorithms, eg. the deviance of the single variable models. But nothing seems to improve the prediction accuracy.

Wrap around

In the case of the breast cancer prognosis data, we did not increase the probe threshold by the adaptive thresholding step in all of our empirical results. However, we did employ the adaptive thresholding step in both the leukemia and the hereditary breast cancer datasets. In these cases, we experimented with the following idea:

- Keep track of the genes removed due to adaptive threshold, call it S_{adapt} .
- After we exhaust all p genes, we give the genes in S_{adapt} a second chance to get into the 30-gene window in the iterative BMA algorithm.

Our experiments show that this “wrap around” idea does not improve prediction accuracy or reduce the number of relevant genes in general.

4. multi-class iterative BMA

For simplicity, let us start with the 3-class case ($K=3$) although our algorithm and implementation are applicable for any number of classes.

Polychotomous logistic regression

Suppose our multi-class iterative BMA algorithm has selected genes g_1, g_2, \dots, g_q as the relevant genes. Assume that these genes have expression levels/ratios x_1, x_2, \dots, x_q . As usual, let Y be the response variable (class) and Y takes values 0, 1, or 2.

$$g_1(x) = \ln \left[\frac{\Pr(Y=1|D)}{\Pr(Y=0|D)} \right] = b_{10} + b_{11}x_1 + \dots + b_{1q}x_q$$

$$g_2(x) = \ln \left[\frac{\Pr(Y=1|D)}{\Pr(Y=0|D)} \right] = b_{20} + b_{21}x_1 + \dots + b_{2q}x_q$$

The goal of logistic regression is to estimate the regression parameters (b_{ij}), and it follows that the conditional probability of each response variable is:

$$\Pr(Y=0|D) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\Pr(Y=1|D) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\Pr(Y=2|D) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

Generalization to the K-class data, where $K \geq 3$.

$$g_k(x) = \ln \left[\frac{\Pr(Y=k|D)}{\Pr(Y=0|D)} \right] = b_{k0} + b_{k1}x_1 + \dots + b_{kq}x_q \quad \text{where } k = 1, 2, \dots, (K-1)$$

Assuming $g_0(x)=0$, the conditional probability of each response variable is:

$$\Pr(Y=k|D) = \frac{e^{g_k(x)}}{\sum_{j=0}^{K-1} e^{g_j(x)}} \quad \text{where } k = 0, 1, \dots, (K-1)$$

Please refer to Chapter 8 of (Hosmer and Lemeshow 2000) for further details.

Multi-class iterative BMA algorithm

An outline of the algorithm is described in the manuscript.

Prediction

The maximum likelihood estimates of each regression parameters are extracted. We also keep track of the binary logistic regression from which each variable (gene) is selected. For example, suppose gene g is selected from the binary logistic regression ($Y=0$ vs. $Y=k$). The regression parameter (b) for gene g estimated using the iterative BMA algorithm will be used in the function $g_k(x)$.

Note that some genes are selected in only a subset of the binary logistic regression. If a variable is not selected in a binary logistic regression, we set the regression parameter to 0.

$$\Pr(Y=k|D) = \sum_{k \in M} \left[\frac{e^{g_k(x)}}{\sum_{j=0}^{K-1} e^{g_j(x)}} \right] * \Pr(M_k | D) \quad \text{where } k = 0, 1, \dots, (K-1)$$