

Microarray Data Analysis



Dhammika Amaratunga

*Senior Research Fellow in Nonclinical Biostatistics
Johnson & Johnson*

Javier Cabrera

*Director Institute of Biostatistics, Professor,
Dept. of Statistics, and Biostatistics, Rutgers University*



Webinar in ASA Biopharmaceutical Section Webinar Series, March 2009

Outline

1. Data Reduction: Selection, Weighting, PCA, PLS
2. Clustering: ABC clustering.
3. Classification: Enriched Classifiers: Random Forest
4. Classification: Ensemble Classifiers
5. LASSO and other (if there is time)

Microarray data as Multivariate Data

Gene Expressions: $G \times p$ matrix Genes are variables $\Rightarrow G \gg p$

Correlation Matrix:
$$R = \begin{pmatrix} 1, r_{12}, \dots, r_{1G} \\ r_{21}, 1, \dots, r_{2G} \\ \dots \dots \dots \\ r_{G1}, r_{G2}, \dots, 1 \end{pmatrix}$$

Dim(R) = $G \times G$ and G is generally up to 50000 or more, this is too big \Rightarrow Curse of Dimensionality .

Dimension Reduction: PCA, Response or grouping variable \Rightarrow PLS

Gene Selection: Response: Calculate the F–statistic for each individual gene and select those genes with the lowest p–value.

It may not work: High dimensionality introduces spurious signal.

Weights: Assign high weight to most important genes.

No response case: Important genes have high variance.

Response case: Calculate p–values and transform into q–values.

Make the weights equal to the reciprocal or log of the q–values.

Principal Components Analysis and Partial Least Squares

Calculate Principal components without calculating any GxG matrix

Singular Value Decomposition: $X = U D V'$
Gxp Gxp pxp pxp

The Correlation Matrix takes the form: $R = U D^2 U'$
GxG Gxp pxp pxG

S is GxG but we do not need to write it down to calculate U and D²

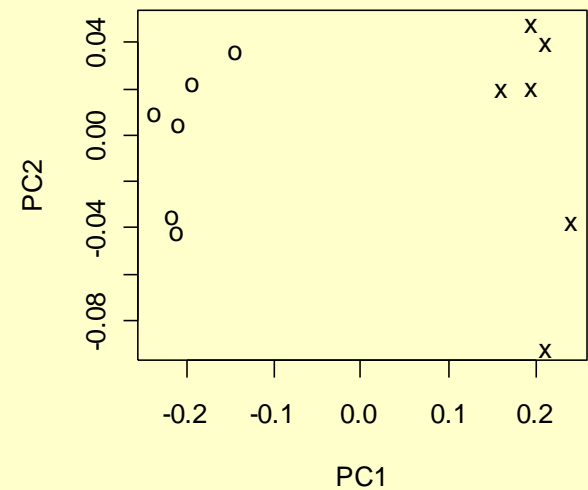
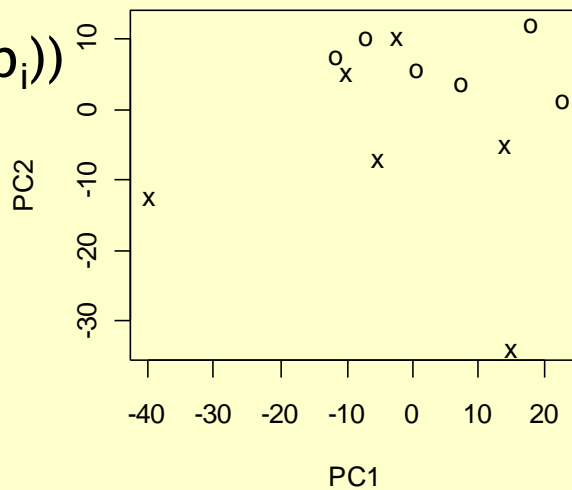
Weighted PCA : Day 0 data
Unweighted PCA's Weighted PCA's

$$W_i = 1/q_i \text{ (or } W_i = -\log(p_i) \text{)}$$

$$W = \text{Diag}(W_i)$$

$$X^* = W X = U^* D^* V^{*'}$$

$$R^* = W R W = U^* D^{*2} U^{*'}$$



Note: FDR & q-values

p_1, p_2, \dots, p_k are p-values. t = Threshold for significance, $0 < t \leq 1$

π_0 = proportion of truly null genes. $S(t) = \#\{p_i \leq t\}$ $F(t) = \#\{\text{null } p_i \leq t\}$

$$FDR(t) = E[F(t)/S(t)] \approx E[F(t)]/E[S(t)]$$

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{p(1-\lambda)}$$

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 p t}{\#\{p_i \leq t\}}$$

$$q_i = \min_{t \geq p_i} \widehat{FDR}(t)$$

Partial Least Squares Regression

\mathbf{y} has $\text{mean}(\mathbf{y})=0$, $\text{Var}(\mathbf{y})=1$, \mathbf{x}_j has $\text{mean}(\mathbf{x}_j)=0$, $\text{Var}(\mathbf{x}_j)=1$ for all j .

1. $\hat{b}_j = \langle \mathbf{x}_j, \mathbf{y} \rangle$: slope of regressing \mathbf{y} on each \mathbf{x}_j , $\mathbf{z}_1 = \sum_{j=1}^p \hat{b}_j \mathbf{x}_j$
2. $\hat{\beta}_1 = \langle \mathbf{z}_1, \mathbf{y} \rangle / \langle \mathbf{z}_1, \mathbf{z}_1 \rangle$ coefficient of regressing \mathbf{y} on \mathbf{z}_1 ,
3. Update the \mathbf{x}_j 's by orthogonalizing them w/r \mathbf{z}_1 . Update \mathbf{y} by the residuals of the previous linear fit.

Iterate 1-3 produces a set of orthogonal vectors $\{\mathbf{z}_i\}$ and estimators $\{\hat{\beta}_i\}$

Relation to Weights: $\hat{b}_j \approx -\log(p_j)$ hence $\mathbf{z}_1 \approx \mathbf{z}_1^* = \sum -\log(p_j) \mathbf{x}_j$

Use our weights $\mathbf{z}_1^W = \sum W_j \mathbf{x}_j = \sum -\log(q_j) \mathbf{x}_j$

If feature set is small and correlated to \mathbf{y} then \mathbf{Z}_1^W performs like \mathbf{z}_1

If feature set is big and only a few features are correlated to \mathbf{y} then \mathbf{Z}_1^W outperforms \mathbf{z}_1

\mathbf{Z}_1^W provides a way adapt PLS to high dimensional data with low signal



"Hey! I've just had a great idea!
How about a light bulb....?"

ABC clustering

- 1. A Bootstrap approach called ABC Refers to the Bagging of genes and samples from Microarray data. Genes are bagged using weights proportional to their variances.*
- 2. By creating new datasets out of subsets of columns and genes we are able to create estimates of the class response several hundred times.*
- 3. These estimates are then used to obtain a dissimilarity (distance) measure between the samples of the original data.*
- 4. This dissimilarity matrix is then adopted to cluster the data.*

Data

Gene	S1	S2	S3	S4	S5	S6
G8521	1003	1306	713	1628	1268	1629
G8522	890	705	566	975	883	1005
G8523	680	749	811	669	724	643
G8524	262	311	336	1677	1286	1486
G8525	254	383	258	1652	1799	1645
G8526	81	140	288	298	241	342
G8527	4077	2557	2600	3394	2926	2755
G8528	2571	1929	1406	2439	1613	5074
G8529	55	73	121	22	141	44
G8530	1640	1693	1517	1731	1861	1550
G8531	168	229	284	220	310	315
G8532	323	258	359	345	308	315
G8533	12131	11199	14859	11544	11352	11506
G8534	11544	11352	12131	11199	14859	12529
G8535	1929	1406	2439	254	383	258
G8536	191	140	288	298	241	342
G8537	4077	2557	2600	3394	2926	2755
G8538	2571	1613	5074	1652	1799	1645
G8539	55	73	121	22	91	24
G8540	1640	1693	1517	1731	1861	1750
G8541	168	229	284	220	312	335
G8542	323	258	359	345	298	325
G8543	2007	1878	1502	1758	2480	1731
G8544	2480	1731	2007	1878	1502	1758
G8545	1652	1799	1645	254	383	258
G8546	298	241	342	81	150	298
G8547	2607	3394	2926	2755	3077	2227
G8548	2571	1929	1406	2439	1613	5074
G8549	121	22	55	730	201	35
G8550	1640	1693	1517	1731	1861	1550

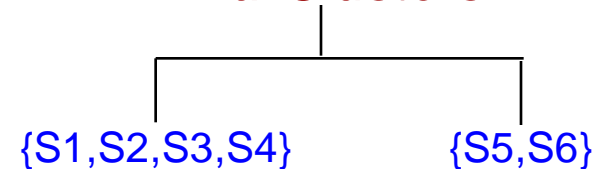
Select n samples and g genes

Gene	S1	S2	S4	S5	S6
G8523	680	749	669	724	643
G8524	262	311	1677	1286	1486
G8528	2571	1929	2439	1613	5074
G8530	1640	1693	1731	1861	1550
G8537	4077	2557	3394	2926	2755
G8545	1652	1799	254	383	258
G8547	2607	3394	2755	3077	2227

Compute similarity

Similarity	S1	S2	S3	S4	S5	S6
S1	0	6	7	7	0	0
S2	6	0	5	5	1	1
S3	7	5	0	8	0	0
S4	7	5	8	0	2	2
S5	0	2	0	2	0	10
S6	0	2	0	2	10	0

Final Clusters



Examples

For each data set:

Genes Selected = \sqrt{G} ,

Simulations = 500

Genes Bagged By Variance

	Armstrong	Colon	Tao	Golub	Iris
BagWeight	0.01	0.1	0.2	0.17	0.05
BagEquiWeight	0.07	0.48	0.2	0.36	0.11
BagWholeData	0.08	0.48	0.3	0.4	0.05
NoBagWeight	0.01	0.1	0.2	0.17	0.08
NoBagEquiWeight	0.03	0.37	0.2	0.4	0.13
Ward	0.1	0.48	0.4	0.29	0.09
Kmeans	0.06	0.48	0.4	0.21	0.11

Random Forest with weights (Enriched Random Forest)

1. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
2. Grow a "random" tree, where at each node, the best split is chosen among m randomly selected variables. The tree is grown to maximum size and not pruned back.
3. Use the tree to predict out-of-bag data.
4. use the predictions on out-of-bag data to form majority votes.
5. Repeat 1-4 N times and collect an ensemble of N trees. Prediction of test data is done by majority votes from predictions from the ensemble of trees.

Incorporate Weights

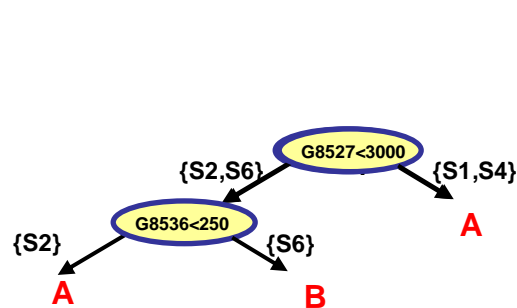
- 2.* Grow a "random" tree, where at each node, the best split is chosen among m randomly selected variables according to the weights $\{W_i\}$. The tree is grown to maximum size and not pruned back.

OUT-OF-BAG SET

Data

Y	A	A	B	A	A	B
Gene	S1	S2		S4		S6
G8521	1003	1306		1628		1629
G8522	890	705		975		1005
G8523	680	749		669		643
G8524	262	311		1677		1486
G8525	254	383		1652		1645
G8526	81	140		298		342
G8527	4077	2557		3394		2755
G8528	2571	1929		2439		5074
G8529	55	73		22		44
G8530	1640	1693		1731		1550
G8531	168	229		220		315
G8532	323	258		345		315
G8533	12131	11199		11544		11506
G8534	11544	11352		11199		12529
G8535	1929	1406		254		258
G8536	191	140		298		342
G8537	4077	2557		3394		2755
G8538	2571	1613		1652		1645
G8539	55	73		22		24
G8540	1640	1693		1731		1750
G8541	168	229		220		335
G8542	323	258		345		325
G8543	2007	1878		1758		1731
G8544	2480	1731		1878		1758
G8545	1652	1799		254		258
G8546	298	241		81		298
G8547	2607	3394		2755		2227
G8548	2571	1929		2439		5074
G8549	121	22		730		35
G8550	1640	1693		1731		1550

Enriched Random Forest(ERF)



OUT-OF-BAG SET

Y	A	A	B	A	A	B
Gene	S1	S2	S3	S4	S5	S6
G8523	680	749	871	669	724	643
G8524	262	311	386	1677	1286	1486
G8528	2571	1929	1406	2439	1613	5074
G8530	1640	1693	1517	1731	1861	1550
G8537	4077	2557	600	3394	2926	2755
G8545	1652	1799	1645	254	383	258
G8547	2607	3394	2926	2755	3077	2227
G8549	121	22	55	730	201	35
G8550	1640	1693	1517	1731	1861	1550

OUT-OF-BAG prediction

SAMPLE	S1	S2	S3	S4	#A	#B	PRED
S1		A			113	45	A
S2			B	A	187	11	A
S3	B		A		98	110	B
S4				A	145	110	A
S5	A		A		199	2	A
S6		A		B	108	102	A

Out of Bag Error Rates

- Random forest (RF),
- Random forest with p -value based filtering (RF(p)),
- Random forest with g -based filtering (RF(g)),
- Enriched random forest with t -based weights (ERF(t)) and
- Enriched random forest with Ct-based weights (ERF(Ct)).
- $R=1000$ iterations were used in each procedure.

	RF	RF(p)	RF(g)	ERF- OUT (t)	ERF- OUT (Ct)	ERF- IN (t)	ERF- IN (Ct)	ERF- CV (t)	ERF- CV (Ct)
Breast cancer	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.000
Diabetes	0.543	0.514	0.286	0.143	0.343	0.457	0.543	0.486	0.571
Epilepsy	0.154	0.077	0.000	0.077	0.154	0.154	0.154	0.077	0.154
HIV Encephalitis	0.357	0.429	0.357	0.179	0.179	0.250	0.250	0.250	0.286
Slc17A5 Day 0	0.583	0.583	0.250	0.083	0.000	0.167	0.000	0.083	0.000
Slc17A5 Day 18	0.083	0.083	0.083	0.000	0.000	0.000	0.000	0.000	0.000
<i>Slc17A5 Day 0 (scrambled)</i>	<i>0.750</i>	<i>0.750</i>	<i>0.333</i>	<i>0.833</i>	<i>0.750</i>	<i>0.833</i>	<i>0.667</i>	<i>0.750</i>	<i>0.667</i>
<i>Slc17A5 Day 18 (scrambled)</i>	<i>0.583</i>	<i>0.667</i>	<i>0.583</i>	<i>0.667</i>	<i>0.750</i>	<i>0.667</i>	<i>0.667</i>	<i>0.417</i>	<i>0.583</i>

Ensemble Classifiers

1. Draw a bootstrap sample from the data. Call those not in the bootstrap sample the "out-of-bag" data.
2. Generate m randomly selected features according to the weights $\{W_i\}$ and use them together with the bootstrap sample to construct a classifier using method "A".
3. Use the Classifier to predict out-of-bag data to form majority votes.
4. Repeat 1-3 N times and collect an ensemble of N trees. Prediction of test data is done by majority votes from predictions from the ensemble of trees.
5. Classifier "A" can be any standard classifier: LDA, ANN, PLS, SVM, KNN, LASSO,....

OUT-OF-BAG SET

Data

Y	A	A	B	A	A	B
Gene	S1	S2		S4		S6
G8521	1003	1306		1628		1629
G8522	890	705		975		1005
G8523	680	749		669		643
G8524	262	311		1677		1486
G8525	254	383		1652		1645
G8526	81	140		298		342
G8527	4077	2557		3394		2755
G8528	2571	1929		2439		5074
G8529	55	73		22		44
G8530	1640	1693		1731		1550
G8531	168	229		220		315
G8532	323	258		345		315
G8533	12131	11199		11544		11506
G8534	11544	11352		11199		12529
G8535	1929	1406		254		258
G8536	191	140		298		342
G8537	4077	2557		3394		2755
G8538	2571	1613		1652		1645
G8539	55	73		22		24
G8540	1640	1693		1731		1750
G8541	168	229		220		335
G8542	323	258		345		325
G8543	2007	1878		1758		1731
G8544	2480	1731		1878		1758
G8545	1652	1799		254		258
G8546	298	241		81		298
G8547	2607	3394		2755		2227
G8548	2571	1929		2439		5074
G8549	121	22		730		35
G8550	1640	1693		1731		1550

Ensemble Classifiers

OUT-OF-BAG SET

Y	A	A	B	A	A	B
Gene	S1	S2	S3	S4	S5	S6
G8523	680	749	811	669	724	643
G8524	262	311	336	1677	1286	1486
G8528	2571	1929	1406	2439	1613	5074
G8530	1640	1693	1517	1731	1861	1550
G8537	4077	2557	2600	3394	2926	2755
G8545	1652	1799	1645	254	383	258
G8547	2607	3394	2926	2755	3077	2227
G8549	121	22	55	730	201	35
G8550	1640	1693	1517	1731	1861	1550

CLASSIFIER

OUT-OF-BAG prediction

SAMPLE	S1	S2	S3	S4	#A	#B	PRED
S1		A			113	45	A
S2			B	A	187	11	A
S3	B		A		98	110	B
S4				A	145	110	A
S5	A		A		199	2	A
S6		A		B	108	102	A

Out of Bag Error Rates

	RF	ERF	ERF (Ct)	LDA-PCA (BagE)	LDA PCA (SimE)	LDA PCA (EnrE)	DLDA (BagE)	DLDA (SimE)	DLDA (EnrE)	LASSO (BagE)	LASSO (SimE)	LASSO (EnrE)
Slc17A5 Day 0	0.583	0.167	0.000	0.417	0.583	0.000	0.500	0.583	0.250	0.000	0.500	0.000
Slc17A5 Day 18	0.083	0.000	0.000	0.000	0.000	0.000	0.083	0.083	0.000	0.083	0.000	0.083
<i>Slc17A5 Day 0 (scrambled)</i>	<i>0.750</i>	<i>0.833</i>	<i>0.667</i>	<i>0.833</i>	<i>0.750</i>	<i>0.833</i>	<i>0.833</i>	<i>0.667</i>	<i>0.667</i>	<i>0.583</i>	<i>0.833</i>	<i>0.833</i>
<i>Slc17A5 Day 18 (scrambled)</i>	<i>0.583</i>	<i>0.667</i>	<i>0.667</i>	<i>0.667</i>	<i>0.583</i>	<i>0.583</i>	<i>0.583</i>	<i>0.583</i>	<i>0.583</i>	<i>0.750</i>	<i>0.750</i>	<i>0.750</i>
Astrocytoma	0.214	0.000	0.071	0.143	0.143	0.071	0.429	0.429	0.214	0.214	0.143	0.214
Breast Cancer	0.029	0.029	0.029	0.000	0.000	0.000	0.314	0.029	0.029	0.000	0.029	0.029
Epilepsy	0.154	0.154	0.154	0.077	0.154	0.077	0.538	0.077	0.077	0.077	0.077	0.077
HIV _Encephalitis	0.357	0.250	0.250	0.143	0.179	0.179	0.286	0.286	0.286	0.179	0.214	0.179
Human Lymph _Node Sinus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Macular - _degeneration	0.111	0.083	0.083	0.083	0.111	0.056	0.139	0.139	0.111	0.028	0.083	0.028

Out of Bag Error Rates

	KNN (BagE)	KNN (SimE)	KNN (EnrE)	PLS (BagE)	PLS (SimE)	PLS (EnrE)	SVM (BagE)	SVM (SimE)	SVM (EnrE)
Slc17A5 Day 0	0.333	0.833	0.167	0.333	0.417	0.250	0.500	0.583	0.500
Slc17A5 Day 18	0.083	0.000	0.000	0.000	0.000	0.000	0.167	0.250	0.000
<i>Slc17A5 Day 0 (scrambled)</i>	<i>0.833</i>	<i>0.667</i>	<i>0.667</i>	<i>0.583</i>	<i>0.667</i>	<i>0.667</i>	<i>0.833</i>	<i>0.750</i>	<i>0.833</i>
<i>Slc17A5 Day 18 (scrambled)</i>	<i>0.583</i>	<i>0.583</i>	<i>0.667</i>	<i>0.583</i>	<i>0.500</i>	<i>0.667</i>	<i>0.583</i>	<i>0.583</i>	<i>0.583</i>
Astrocytoma	0.214	0.286	0.071	0.143	0.214	0.071	0.214	0.214	0.071
Breast Cancer	0.029	0.029	0.029	0.000	0.000	0.000	0.000	0.029	0.029
Epilepsy	0.077	0.154	0.077	0.077	0.077	0.077	0.077	0.077	0.077
HIV Encephalitis	0.321	0.250	0.286	0.143	0.179	0.107	0.107	0.107	0.107
Human Lymph Node Sinus	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Macular degeneration	0.194	0.194	0.111	0.083	0.111	0.056	0.056	0.083	0.056

Wrap up



◆ References:

- D. Amaratunga and J. Cabrera (2004), *Exploration and Analysis of DNA Microarray and Protein Array Data*, New York: John Wiley.
- D. Amaratunga, J. Cabrera and Y. S. Lee (2009). Growing better random forests with megavariable data, *in review*.
- D. Amaratunga and J. Cabrera (2007). A conditional t suite of tests for identifying differentially expressed genes in a DNA microarray experiment with little replication, in advance access publication in *Statistics in Biopharmaceutical Research*.
- D. Amaratunga, J. Cabrera and V. Kovtun (2007). Microarray learning with ABC, in advance access publication in *Biostatistics*.
- J. Cabrera and C. Yu.(2007) Estimating the proportion of differentially expressed genes in comparative DNA Microarray Experiments, *IMS Lect. Notes-Monograph Series 54*.
- N. Raghavan, D. Amaratunga, J. Cabrera, A. Nie, Q. Jie and M. McMillian (2006), On methods for gene function scoring as a means of facilitating the interpretation of microarray results, *Journal of Computational Biology*, 13: 798-809
- D. Amaratunga and J. Cabrera (2001), Statistical analysis of viral microchip data, *Journal of the American Statistical Association*, 96(454): 1101-1110

◆ Website:

www.rci.rutgser.edu/~cabrera/DNAMR

Thank you