Advanced Statistical Methods for the Analysis of Gene Expression and Proteomics Gene/Feature Selection and Classification

Veera Baladandayuthapani
(pronounced as Veera B)

University of Texas M.D. Anderson Cancer Center
Houston, Texas, USA
veera@mdanderson.org

Course Website:
http://odin.mdacc.tmc.edu/~kim/TeachBioinf/AdvStatGE-Prot.htm

Story till now....

- What is Bioinformatics?
  - As the generation, organization, and analysis of biological data (initially genomic data)
  - Attracted lot of interest in different fields: Computer Science, Physics, Engineering and of course Statistics
- Microarrays
  - What are they?; What they measure?
  - Pre-processing issues: normalization, technical vs biological variation
  - Downstream analysis

Microarray Technology

- High-throughput assays for understanding molecular biology
- Simultaneously measure expression levels for thousands of genes
- By understanding how "gene expression" changes across multiple conditions
  - Researches gain clues about gene functions
  - How genes work together to carry out biological functions
- Many applications in a variety of studies; attracted considerable statistical literature
- Other techniques to measure gene expression
  - Serial analysis of gene expression (SAGE); cDNA library sequencing; differential display; cDNA subtraction; multiplex quantitative RT-PCR

Final Data for Analysis

- What statisticians work with: Gene Expression Matrix

Samples	Gene 1	Gene 2	$\cdots$	Gene p
1	X	X	$\cdots$	X
2	X	X	$\cdots$	X
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X	X	$\cdots$	X

  - X = Gene expression intensities (some form)
  - p = Number of genes (usually in thousands)
  - n = Number of samples (micorarrays) ($n \ll p$)
  - Y (tissue type/phenotype) = 0 if Normal; 1 if Cancer (binary)
  - Z = Design variables for controlled experiments (e.g. Drug A/B) OR Covariates

Statistical Issues with Microarray Data

- Preprocessing of the data
  - Assess spot quality, reliability of signal, normalize data
- Differential expression (Last two classes and next class)
  - Identify which genes are up-/down-regulated in different sets of experimental conditions
- Classification/Discrimination (supervised learning)
  - Use gene expression profile to predict type of tumor (class prediction)
- Clustering (unsupervised learning)
  - Determine genes that are coexpressed or new subtypes of disease (class discovery)
- Feature (gene) selection

Dimension reduction

- Often in microarrays: $n << p$
  - Order of n: tens or hundreds
  - Order of p: thousands or more
- Therefore it is advisable/essential from a practical and methodological point of view to reduce the dimension i.e. p; not all genes affect the process
- Termed Variable/Gene/Feature selection
- Statistical theory: Model selection i.e. different set(s) of variables(genes) different models
- Rich literature in non-microarray context also: stepwise, backward, forward regression; AIC; BIC.

Feature Selection in a context

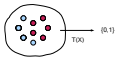
- Variable important by itself
  - Gene independently ranked by some criteria
- Gene important in a context
  - Combine variables
  - Model for combining variables is needed
- Important genes not in a context
  - Model averaging; ensemble learning
- Today's lecture: Gene selection in a context: Classification

Classification

- Objective: assign objects to classes (groups) on the basis of measurements on the objetcs
- Unsupervised: classes are unknown and want to discover them from data *apriori*
- Supervised: classes are known *apriori* and want to use a traning/learning set of labeled objects to form a classifier for classification of future observations
- In microarray context
  - Objects are microarrays here, and are to be classified as belonging to one of a number of predefined classes $\{1, 2, \ldots, K\}$
  - Each array has a class label: $Y \in \{1, 2, \ldots, K\}$ and associated feature vector of G genes: $X = \{X_1, X_2, \ldots, X_G\}$ and the aim is to predict Y from X.

Classification

Suppose there are two populations, healthy and disease individuals. Let the class labels (arbitrary) $Y_j = 0$ if individual j is healthy and $Y_j = 1$ is individual j has disease. The classifier function, $T(X_j)$, predicts Y_j, given variables X_j. The function is a mapping from X to the class labels, $T : X \to \{0, 1\}$.



Different nomenclature in different fields:

- Discriminant analysis (multivariate statistics)
- Supervised learning (machine learning/artificial intelligence in computer science)
- Pattern recognition (engineering)
- Prediction, predictive classification (Bayesian)

Why is it important

- In Tumor classification: reliable and precise classification essential for successful cancer treatment
- Characterizing molecular variations among tumor by monitoring gene expression
- Hope is that microarrays will lead to more reliable tumor classification

Statistical Decision Theory

- Useful to view classification as a statistical decision theory problem
- Suppose observation Y's are iid from an unknown multivariate distribution. Denote population proportion of objects of class k as $\pi_k = p(Y = K)$. Objects in class k have feature vectors with class conditional density $p_k(x) = p(x|Y = k)$.
- A loss function $L(i, j)$ quantifies the loss incurred by erroneously classifying a member of class i as class j.
- The risk function for a classifier $T(X)$ is just the expected (average) loss

$$R(T) = E[L(Y, T(X))] = \sum_k E[L(k, T(X))|Y = k]\pi_k$$

- For symmetric loss i.e. $L(i, j) = 1$ for $i \neq j$ then the risk turns out to be simple missclassification rate: $p(T(X) \neq Y)$

Classification and Bayes Rule - I

- In the (unlikely) situation that we know both $p_k(x)$ and π_k, we can use Bayes rule to express posterior probability $p(k|x)$ of class k given a feature gene vector x

$$p(k|X) = \frac{\pi_k p_k(x)}{\sum_l \pi_l p_l(x)}$$

- Bayes' rule predicts class with highest posterior probability

$$T_B(X) = \mathbf{argmax_k p(k|X)}$$

- Bayes rule minimizes the risk function/misclassifiation rate under a symmetric loss function – Bayes risk.

$$T_B(X) = \mathbf{argmax_k L_k p(k|X)}$$

Classification and Bayes Rule - II

- Many classfiers can viewed as versions of this general rule, with either parametric or nonparametric estimators of $p(k|X)$. There are two general paradigms to estimate $p(k|X)$.
- Density estimation approaches e.g. Gaussian maximum likelihood discriminant rules (discrimnant analysis); mostly linear
- Direct function estimation approach: Regression methods e.g. logistic/probit regression, neural networks, classification trees; can be adapted to be more flexible

Maximum Likelihood Discriminant Rules

- Frequentist analogue of Bayes Rule
- MLE chooses the parameter value that makes the chance of the observations the highest
- For known class conditional densities $p_k(x) = p(x|Y = k)$, the ML rule predicts the class of x that gives the largest likelihood to x: $C_M(x) = \mathbf{argmax}_k p_k(x)$.
- In case of equal class priors: π_k, this is same as Bayes Rule
- Otherwise, ML rule is not optimal => does not minimize the risk function

Discriminant Analysis

- Fisher Linear Discrimnant Analysis (FLDA)
- Finds linear combinations $(a'X)$ of the gene expression profiles $X = X_1, ..., X_p$ with large ratios of between-groups to within-groups sums of squares ($\frac{a'Ba}{a'Wa}$) - discriminant variables
- Predicts the class of an observation X by the class whose mean vector is closest to X in terms of the discriminant variables
- Classifier: $T(X) = \mathbf{argmin}_k d_k(x)$ where $d_k^2(x) = \sum_{i=1}^{S}[x - (\bar{x})]_k)v_i$; v_l are discriminating variables.
- Standard method in most multivariate statistics books
- Two main steps: (1) Dimension reduction via eigen values (2) Classification using the discriminant variables.
- Note: No distribution over X's - Nonparametric method

Gaussian Discriminant Rules

- If we assume multivariate Gaussian (normal) class densities for $X|Y = k \sim N(\mu_k, \Sigma_k)$, the ML classifier is

$$T(X) = \mathbf{argmin}_k\{(X - \mu_k)\Sigma_k^{-1}(X - \mu_k)' + log|\Sigma_k| - 2\log\pi_k\}$$

- In general, this is a quadratic rule (Quadratic discriminant analysis, or QDA) in standard multivariate analysis; function of the Mahalanobis distance: $(X - \mu_k)\Sigma_k^{-1}(X - \mu_k)'$
- In practice, population mean vectors μ_k and covariance matrices Σ_k are estimated by corresponding sample quantities
- Most common classifiers are variations of the Gaussian discriminant rule

Common Classifiers

- QDA: $T(X) = \text{argmin}_k\{(X-\mu_k)\Sigma_k^{-1}(X-\mu_k) + log|\Sigma_k| - 2\log\pi_k\}$
- Linear discriminant analysis (LDA): If $\Sigma_k = \Sigma$ and π_k is constant for all k then

$$\begin{aligned} T(X) &= \text{argmin}_k\{(X-\mu_k)\Sigma^1(X-\mu_k)\} \\ &= \text{argmin}_k\{\mu_k\Sigma^1\mu_k - 2x\Sigma^{-1}\mu_k'\} \end{aligned}$$

- Diagonal quadratic discriminant analysis (DQDA): If $\Sigma_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kG}^2)$,

$$T(X) = \text{argmin}_k \sum_{g=1}^{G}\{\frac{x_g - \mu_{kg})^2}{\sigma_{kg}^2} + \log\sigma_{kg}^2\}$$

- Diagonal linear discriminant analysis (DLDA): If $\Sigma_k = \text{diag}(\sigma_1^2, \ldots, \sigma_G^2)$,

$$T(X) = \text{argmin}_k \sum_{g=1}^{G}\{\frac{x_g - \mu_{kg})^2}{\sigma_g^2}\}$$

Various Modifications

- Nearest Centroid ($\Sigma_k = I_G$); G is the number of genes
- Flexible discriminant analysis; Penalized Discriminant Analysis; Mixture Discriminant Analysis
- These are widely used especially for microarray data for a variety of reasons
  - Simple and intuitive: predict class closest to sample mean
  - Estimated Bayes Rule: LDA is Bayes rule with Gaussian distributions
  - Easy to implement
  - Reasonable performance: low classification error

Possible Drawbacks

- Microarray data are very rich and complex; linear or even quadratic classification boundaries may not be flexible enough
- Features (genes) may have mixture distributions within classes
- Curse of dimensionality: for large number of genes the performance may degrade rapidly due to over-parameterization and high variance of parameter estimates
- There are methods and algorithms to overcome some of these problems (later in the course)
- Very nice article comparing common classification methods: Dudoit, Fridlyand, Speed (JASA, 2002)

How do we evaluate Classifiers?

- Error rates
- Resubstitution estimation: fit a single classifier to the data, and applies this classifier in turn to each data observation
  Problem: downward bias; underestimates classification error (sometimes severely)
- Test and training data: divide cases in learning set into two sets, S1 and S2; classifier built using S1, error rate computed for S2. S1 and S2 must be iid (crucial).
  Problem: reduced effective sample size
- V-fold Crossvalidation: learning set randomly divided into V subsets of (nearly) equal size. Build classifiers leaving one set out; test set error rates computed on left out set and averaged.
  Problem: Bias-variance tradeoff: smaller V can give larger bias but smaller variance
- Other methods: Aggregating;Bagging;Boosting

Feature Selection in Classification

- Two ways to do this
  - Do feature selection first and then build a classifier (Filter methods)
  - Implicitly as an inherent part of the classifier building procedure (Wrapper methods)
- Filter methods
  - Simplest: one-gene-at-a-time approaches using univariate test statistics e.g. t or F test, signal to noise ratio, Wilcoxon statistics, p-values
  - More advanced methods: consider joint distribution of genes; ordering methods such as random forests
- Wrapper methods: depends on classifier
  - Some Bayesian classifiers inherently take care of this (more later)
- Bottomline: Feature selection important and is an aspect of classifier training

Revisit Classification Rules

Suppose independent random variables (possibly vectors) $X_{i1}, ..., X_{iN_i}$ are observed from popuations $i = 1, ..., K$, each with probability distribution $f_i(\theta_i)$.
The likelihood of the data is

$$\prod_{i=1}^{K} \prod_{j=1}^{N_i} f_i(X_{ij}|\theta_i),$$

where the θ_i's are unobserved population parameters.

Classification Rules

In classical frequentist parametric classification (as discussed before), a new observation Z is classified by estimating θ_i from the training observations, $\hat{\theta}_i$, and plugging $\hat{\theta}_i$ back into the likelihood to form prediction rules. Z is assigned to the class i for which

$$f_i(Z|\hat{\theta}_i) > f_{i'}(Z|\hat{\theta}_{i'})$$

for all i', and assigned randomly in the event of ties.
There are some disadvantages to this approach. To a Bayesian, θ is unknown, and therefore the uncertainty in θ should be taken into account when making predictions. See Lehmann (1990) for discussion of bias/variance tradeoff in classification.

Classification Rules

In Bayesian parametric classification, a new observation Z is classified by assigning a prior distribution to the θ_i's, $\pi(\theta_1, ..., \theta_K)$, and updating the prior distribution to obtain a posterior distribution

$$\pi(\theta_1, ..., \theta_K|X) \propto \prod_{i=1}^{K} \prod_{j=1}^{N_i} f_i(X_{ij}|\theta_i)\pi(\theta_1, ..., \theta_K).$$

The predictive distribution for the i-th population of a new observation Z is

$$f_i(Z|X) = \int_{\Theta_i} f_i(Z|\theta_i)\pi(\theta_i|X)d\theta_i,$$

for all i, integrating over $\theta|X$.

Classification Rules

The Bayesian prediction rule assigns Z to the population i for which

$$\pi_i f_i(Z|X) > \pi_{i'} f_{i'}(Z|X)$$

for all i', again at random in the event of ties. The posterior distribution of $f_i(Z|\theta_i)$ is known given X, at least up to a normalizing constant.

$$P(Z = i) = \frac{\pi_i f_i(Z|X)}{\sum_{i'} \pi_{i'} f_{i'}(Z|X)}$$

Classification Rules

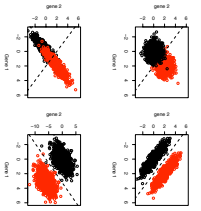
Frequentist methods sometimes resort to large sample or resampling theory in order to determine the uncertainty in prediction.
Measuring the uncertainty in the Bayesian classification rule is straightforward, once $\pi(\theta_i|X)$ is obtained.

Bayesian Linear Classifiers

Suppose that independent random (p-dim) variables $X_{i1}, ..., X_{iN_i}$ are observed from popuations $i = 1, ..., K$, with $j = 1, ..., N_i$ observations each, with probability distributions $N(\mu_i, \Sigma_i)$, where $\theta_i = (\mu_i, \Sigma_i)$ are the unobserved population mean and covariance of X_{ij}.
The likelihood for the data is

$$p(X|\mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_K) = \prod_{i=1}^{K} \prod_{j=1}^{N_i} N\left(X_{ij}|\mu_i, \Sigma_i\right)$$

Classification Rules



Bayesian Linear Classifiers

In the context of microarray data, X_{ij} denotes the vector of gene expression intensity values for individual j in population i. In typical studies, k=2 or 3, for example comparing cancer to normal gene expression, or different types, or stages, of cancer. These studies tend to be large, $N > 100$. Although for microarray classification $N_i < 200$ is considered small.
A convenient, non-informative prior for $\mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_K$ is

$$\pi(\mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_K) \propto \prod_{i=1}^{K} \mid \Sigma_i \mid^{(p+1)/2} .$$

Bayesian Linear Classifiers

The predictive distribution of a new observation Z is

$f_i(z|\bar{x}_i, S_i, \pi_i) \propto$

$$\frac{N_i}{N_i+1}^{p/2} \left[1 + \frac{N_i(z-\bar{X}_i)'S_i^{-1}(z-\bar{X}_i)}{(N_i+1)(N-k)}\right]^{-(N-k+1)/2}$$

where $N = \sum_{i=1}^{k} N_i$, $\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} X_{ij}'$ and

$(N_i - 1)S_i = \sum_j (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'$

Proof: see Press (2003) Bayesian Statistics

Example

Example Consider the case of two populations (classes), where $\Sigma_1 = \Sigma_2$, the frequentist rule is to assign $\mathbf{x}$ to class 1 if

$$P = \frac{q_1 f_1(\mathbf{x})}{q_1 f_1(\mathbf{x}) + q_2 f_2(\mathbf{x})} = [1 + (q_1/q_2)\exp(-L)]^{-1}$$

is greater than 1, class 2 if $P < 1$ and at random if $P = 1$, where L is the log-density ratio,

$$L = \log(f_1(\mathbf{x})/f_2(\mathbf{x})) = (\hat{\delta}_2 - \hat{\delta}_1)/2$$

for $\hat{\delta}_i = (x - \bar{x}_i)'\hat{\Sigma}^{-1}(x - \bar{x}_i)$, and q_i is the known probability that a randomly selected observation is from population i for $i = 1, 2$.

Example

Example Con't

$$\hat{\Sigma} = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2}{(N_1 + N_2 - 2)}.$$

The Bayesian rule, in the case of vague prior knowledge, is to compare

$$P_B = [1 + (q_1/q_2)\exp(-L_B)]^{-1}$$

with 1 analogously,
where

$$L_B = \frac{1}{2}(\nu + 1)\log[(\nu + r_2\hat{\delta}_2)/(\nu + r_1\hat{\delta}_1)] + \frac{1}{2}p\log[r_1/r_2],$$

$r_i = N_i/(N_i + 1)$ and $\nu = N_1 + N_2 - 2$. $P_B = p(Z \in \pi_i|X) = E[P|X]$
(Rigby 1997, *JASA*).

Feature Selection revisited

- Suppose some subset of genes from the microarray are truely differentially expressed in different populations, while the rest of the genes have no information for discrimination.
- Based on non-informative priors, how do you account for the uncertainty in the feature selection? How would a frequentist? Typically the heuristic approach is to select the featuers first, based on some criterion, univariate or multivariate, and then fit the classifier.
- Either way, in applications with array data, there is uncertainty in choosing the features.

Feature Selection revisited

Suppose $\Sigma_i \equiv \Sigma$ for all i. We introduce the indicator variable γ_g, such that $\gamma_g = 1$ if gene g is included in the model and $\gamma_g = 0$ if gene g is excluded. A robust noninformative prior for $(\Sigma, p, \gamma_1, ..., \gamma_G)$ is

$$\begin{aligned} \pi(p, \gamma_1, ..., \gamma_M) &= \pi(\gamma_1, ..., \gamma_M | p)\pi(p) \\ &\propto I\left(\sum_{g=1}^{M} \gamma_g = p\right) \times \frac{\lambda^p e^{-\lambda}}{p!} \mid \Sigma \mid^{(p+1)/2} \end{aligned}$$

Feature Selection

For any given feature set, of size p,
$f_i(z|X_i, \gamma, p) \propto$

$$\frac{N_i}{N_i + 1}^{p/2} \left[1 + \frac{N_i(z - \bar{x}_{i\gamma})' S_\gamma^{-1}(z - \bar{x}_{i\gamma})}{(N_i + 1)(N - k)}\right]^{-(N-k+1)/2}$$

where $\bar{x}_{i\gamma}$ and S_γ are derived from the selected subset of genes. Accounting for uncertainty in feature selection involves integrating of the posterior distribution of $p, \gamma_1, ..., \gamma_M$.

$$f_i(z|X) \propto \int_P \int_\Gamma f_i(z_\gamma | X_\gamma) \pi(\gamma, p|X) d\gamma dp$$

for $i = 1, .., .K$.

Feature Selection revisited

In practice investigating the posterior density for all possible subsets of X_γ of size p is infeasible. Fortunately the unnormalized posterior of (γ, p) may be evaluated as

$$\tilde{\pi}(p, \gamma | X) \propto \prod_{i=1}^{K} \prod_{j=1}^{N_i} f_i(X_{ij} | \gamma, p) \pi(\gamma, p)$$

and $f_i(z|X)$ may be obtained by

$$f_i(z|X) = \frac{\tilde{f}_i}{\sum_{i=1}^{K} \tilde{f}_i}$$

where where

$$\tilde{f}_i = \int_P \int_\Gamma f_i(z_\gamma | X_\gamma) \tilde{\pi}(\gamma, p|X) d\gamma dp.$$

Summary: Frequentist vs Bayesian Classification

- Bayesian and Frequentist classification rules depend on the likelihood function
- Bayesian rules allow prior information
- Bayesian rules flexibly account for all uncertainty in θ (features).
- Bayesian classifiers yield exact measures of prediction uncertainty.
- Intuitively Bayesian Classifiers can reduce variance, by averaging over the uncertainty in θ, see Lehmann (1990) for discussion of bias/variance tradeoff in classification.

Detour: Bayesian Analysis of a Linear Model

The linear model, is frequenty used in many biostatistical applications, including

1. dose response modeling
2. polynomial regression
3. exposure assessment
4. analysis of variance (ANOVA) problems comparing treatment groups

(See *Case studies in Biometry* by Lange et al., John Wiley & Sons.)

Bayesian Analysis of a Linear Model

The linear model can be written as

$$Y = X\beta + \epsilon$$

where Y is a $n \times 1$ response, X is a $n \times p$ matrix of covariates, β is a $p \times 1$ vector of coefficients (unobserved) and

$$\epsilon \sim N_n(0, \sigma^2 I).$$

Let $M = X(X'X)^{-}X'$, and $\tau = \sigma^{-2}$, where $-$ denotes generalized inverse. Recall that the UMVUE of $\mu = E(Y) = X\beta$ is MY. We would like to derive the posterior distribution of β and τ under noninformative priors.

Bayesian Analysis of a Linear Model

Theorem 1

Suppose τ is known, X is of full rank p, and

$$\pi(\beta) \propto 1.$$

Then

$$\beta | y, \tau \sim N_p\left(\hat{\beta}, \tau^{-1}(X'X)^{-1}\right),$$

where

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Proof:

$$\begin{aligned} p(\beta|y,\tau) &\propto \exp\left\{-\frac{\tau}{2}(Y-X\beta)'(Y-X\beta)\right\} \\ &= \exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right\} \\ &= \exp\left\{-\frac{\tau}{2}\left[(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right\}. \end{aligned}$$

Note that

$$\begin{aligned} & Y'(I-M)Y+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \\ = \;& Y'(I-M)Y+\beta'X'X\beta-2\hat{\beta}'X'X\beta+\hat{\beta}'X'X\hat{\beta} \\ = \;& Y'(I-M)Y+\beta'X'X\beta-2Y'X(X'X)^{-1}(X'X)\beta+Y'MY \end{aligned}$$

$$= Y'Y+\beta'X'X\beta-2Y'X\beta=(Y-X\beta)'(Y-X\beta).$$

Thus

$$p(\beta|y,\tau)\propto \exp\left\{-\frac{\tau}{2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right\}.$$

We can recognize this as a normal kernel with mean $\hat{\beta}$ and covariance matrix $\tau^{-1}(X'X)^{-1}$. Thus,

$$\beta|y,\tau\sim N_p\left(\hat{\beta},\tau^{-1}(X'X)^{-1}\right).$$

Theorem 2

When τ is known, Jeffreys prior for β is a uniform prior, i.e.,

$$\pi(\beta)\propto 1.$$

Proof:

$$\begin{aligned} \log[p(y|\beta,\tau)] &= -\frac{n}{2}\log(2\pi)+\frac{n}{2}\log(\tau)-\frac{\tau}{2}(Y-X\beta)'(Y-X\beta) \\ \frac{\partial}{\partial\beta}\log[p(y|\beta,\tau)] &= \frac{\partial}{\partial\beta}\left[-\frac{\tau}{2}(Y-X\beta)'(Y-X\beta)\right] \\ &= \frac{\partial}{\partial\beta}\left[-\frac{\tau}{2}\left[Y'Y-2\beta'X'Y+\beta'X'X\beta\right]\right] \\ &= \tau X'Y-\tau(X'X)\beta. \end{aligned}$$

Also,

$$\frac{\partial^2}{\partial\beta\partial\beta'}\log[p(y|\beta,\tau)]=-\tau(X'X),$$

and therefore,

$$I(\beta)=\tau(X'X).$$

Thus Jeffreys prior for β is given by

$$\pi(\beta|\tau)\propto|\tau(X'X)|^{1/2}\propto 1.$$

Bayesian Analysis of a Linear Model

Theorem 3
Consider the linear model where **both** β and τ are unknown. Then **Jeffreys joint prior** for (β, τ) is given by

$$\pi(\beta|\tau) \propto \tau(X'X)^{p/2-1}.$$

Proof: Exercise

Theorem 4
Consider the linear model with both β and τ unknown, and suppose

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Bayesian Analysis of a Linear Model

Then

$$\beta|y \sim S_p\left(n-p, \hat{\beta}, s^2(X'X)^{-1}\right),$$

where $s^2 = Y'(I-M)Y/(n-p)$ and
$t|y \sim$ gamma$((n-p)/2, s^2(n-p)/2)$.

Proof:
We have

$$\begin{aligned} p(\beta, \tau|y) &\propto \tau^{n/2-1}\exp\left\{-\frac{\tau}{2}(Y-X\beta)'(Y-X\beta)\right\} \\ &= \tau^{n/2-1}\exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right\} \end{aligned}$$

Bayesian Analysis of a Linear Model

Thus,

$$\begin{aligned} p(\beta|y) &\propto \int_0^\infty \tau^{n/2-1}\exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y + (Y-X\beta)'(Y-X\beta)\right]\right\} d\tau \\ &= \left[Y'(I-M)Y + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]^{-n/2}. \end{aligned}$$

Let $s^2 = Y'(I-M)Y/(n-p)$. Then the above integral is

$$\begin{aligned} &= \left[(n-p)s^2 + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]^{-(n-p+p)/2} \\ &= \left[1 + \frac{1}{s^2(n-p)}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]^{-(n-p+p)/2} \end{aligned}$$

Thus, $\beta|y \sim S_p(n-p, \hat{\beta}, s^2(X'X)^{-1})$.

Bayesian Analysis of a Linear Model

Now,

$$\begin{aligned} p(\tau|y) &\propto \int_{-\infty}^\infty \tau^{n/2-1}\exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y + (Y-X\beta)'(Y-X\beta)\right]\right\} d\beta \\ &= \tau^{n/2-1}\exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y\right]\right\} \\ &\quad \int_{-\infty}^\infty \exp\left\{-\frac{\tau}{2}(\beta-\hat{\beta})'(\beta-\hat{\beta})\right\} d\beta \\ &\propto \tau^{(n-p)/2-1}\exp\left\{-\frac{\tau}{2}\left[Y'(I-M)Y\right]\right\} \\ &= \tau^{(n-p)/2-1}\exp\left\{-\frac{\tau}{2}\left[(n-p)s^2\right]\right\} \end{aligned}$$

Thus $\tau|y \sim$ gamma $((n-p)/2, s^2(n-p)/2)$.

Theorem 5

Consider the linear model with β and τ unknown, and suppose

$$\beta|\tau \sim N_p(\mu_o, \tau^{-1}\Sigma_o)$$
$$\tau \sim \text{gamma}\left(\frac{\delta_o}{2}, \frac{\gamma_o}{2}\right)$$

Then $\beta|y \sim S_p\left(n+\delta_o, \beta^*, \tilde{s}^2(X'X+\Sigma_o^{-1})^{-1}\right)$, where

$$\begin{aligned}\beta^* &= \Lambda\mu_o + (I-\Lambda)\hat{\beta},\\ \Lambda &= (X'X+\Sigma_o^{-1})^{-1}\Sigma_o^{-1},\\ \hat{\beta} &= (X'X)^{-1}X'Y,\end{aligned}$$

and

$$\tilde{s}^2 = (n+\delta_o)^{-1}[Y'(I-M)Y + (\hat{\beta}-\mu_o)'(\Lambda'X'X)(\hat{\beta}-\mu_o) + \gamma_o],$$

and

$$\tau|y \sim \text{gamma}\left(\frac{(n+\delta_o)}{2}, \frac{\tilde{s}^2(n+\delta_o)}{2}\right).$$

Proof: Exercise

Hint:

$$\pi(\beta,\tau|Y) \propto \tau^{\frac{n+p+\delta_o}{2}-1}e^{-\frac{\tau}{2}Q}$$

where

$$Q = (Y-X\beta)'(Y-X\beta) + (\beta-\mu_o)'\Sigma_o^{-1}(\beta-\mu_o) + \delta_o$$

Notice that

$$\begin{aligned}(Y-X\beta)'(Y-X\beta) &= \beta'X'X\beta - \beta'X'Y - Y'X\beta + Y'Y\\ &= (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) + Y'(I-M)Y.\end{aligned}$$

Setting $Q = (\beta-\beta^*)'\Sigma^{*-1}(\beta-\beta^*)$ we have

$$\begin{aligned}Q &= \beta'X'X\beta - \hat{\beta}'X'X\beta - \beta'X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} + [Y'Y - Y'MY]\\ &\quad +[\beta'\Sigma_o^{-1}\beta - \mu_o'\Sigma_o^{-1}\beta - \beta'\Sigma_o^{-1}\mu_o + \mu_o'\Sigma_o^{-1}\mu_o] + \delta_o\\ &= \beta'\Sigma^{*-1}\beta - \beta^{*\prime}\Sigma^{*-1}\beta - \beta'\Sigma^{*-1}\beta^{*\prime} + \beta^{*\prime}\Sigma^{*-1}\beta^*\end{aligned}$$

Rearranging terms, and equating quadratic and linear terms we find that

$$\begin{aligned}\Sigma^* &= \left(X'X+\Sigma_o^{-1}\right)^{-1}\\ \Sigma^{*-1}\beta^* &= X'X\hat{\beta} + \Sigma_o^{-1}\mu_o\\ \beta^* &= \left(X'X+\Sigma_o^{-1}\right)^{-1}\left(X'X\hat{\beta} + \Sigma_o^{-1}\mu_o\right)\end{aligned}$$

Bayesian ANOVA Models for Gene Expression Data

The One-Way ANOVA model, for gene g is defined for a single response vector Y_g as

$$Y_g = x^T \beta_g + \epsilon_g \quad (1)$$

where x^T is a matrix of indicator variables for $j = 1, \ldots, k$ treatments ($k = 2$ often, in marker studies) and β_g

$$\beta_g = (\beta_{g1}, \ldots, \beta_{gk}) \quad (2)$$

is the k-dimensional (unknown) vector of treatment effects for gene g, and σ_g^2 is the unknown variance of ϵ_g .

Bayesian ANOVA Models for Gene Expression Data

For $\beta_g = (\beta_{g1}, \ldots, \beta_{gk})$ one could assume either non-informative and informative prior specifications (depending on the case). See

Lindley and Smith (1972) article for an extensive treatment of the Bayesian linear model

The ANOVA model is very powerful, and popular, for microarray analysis. The model has a strong basis in normal theory, and may be applied in many settings.

Note: In this model setup, the genes are assumed independent, largely out of convenience and admittedly naivete.

Bayesian Feature Selection

Note that in biomarker discovery we are interested in variable selection, i.e. determining the set of genes responsible for significant variation between the j = 1, ..., k treatment groups. Variable selection algorithms for high-dimension are discussed in work by:

- George and McCulloch (1997): Bayesian variable selection via Gibbs Sampline
- Brown, Vannucci and Fearn (1998): Multivariate extension
- Storey (2003): FDR based
- Lee (2003), Sha (2006): Probit binary/multinomial regression with variable selection
- Ishwaran and Rao (2003): ANOVA models for gene expression
- Ibrahim, Chen and Gray (2002): Threshold models

Extending Basic ANOVA Model

- One of the first Bayesian models for differential expression was that of Ibrahim, Chen and Gray (2002)
- Propose a general parametric Bayesian model that accomplishes two goals.
- Determines which genes best discriminate between different types of cancer
- Characterize the expression patters in the tumor tissues
- Model the expression under each tissue condition (normal/tumor) as coming from a mixture of a point mass and a log-normal distribution

Extending Basic ANOVA Model

Model gene expression x as,

$$x = \begin{cases} c_o & \text{with probability } p \\ c_o + y & \text{with probability } 1 - p \end{cases}$$

where $c_o > 0$ is the threshold level at which x is considered not expressed. This is a truncated distribution, where c_o is the lower bound, and y, is the continuous part.

Extending Basic ANOVA Model

Let x_{jig} denote the gene expression where j indexes the tissue type (e.g. 1=normal, 2=tumor), i indexes the individual, $i = 1, \ldots, n_j$ and g indexes the gene, $g = 1, \ldots, G$. Similarly, y_{jig} denote the continuous component if the gene expression level for the jth tissue type for the ith individual and the gth gene.

Assume y_{jig} are independently log-normal distributed as,

$$\begin{aligned} p(y_{ijg}|\mu_{jg}, \sigma^2_{jg}) &= (2\pi)^{-1/2} y_{jig}^{-1} \sigma_{jg}^{-1} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma^2_{jg}} (\log(y_{jig}) - \mu_{jg})^2 \right\} \end{aligned}$$

Let $\delta_{jig} = 1$ is $x_{jig} = c_o$ and 0 otherwise. Further, the prior probability $P(\delta_{jig} = 1) \equiv P(x_{jig} = c_0) = p_{jg}$

Extending Basic ANOVA Model

Let $\theta = (\mu, \sigma^2, p)$ be the collection of all parameters for $j = 1, 2$ and $g = 1, \ldots, G$. Then conditional on the observed data $D = (x, \delta)$, the likelihood for θ is given by,

$$L(\theta|D) = \prod_{jig} p_{jg}^{\delta_{jig}} (1 - p_{jg})^{1-\delta_{jig}} p(y_{jig}|\mu_{jg}, \sigma^2_{jg})^{1-\delta_{jig}}$$

With this formulation, all the fundamental questions can be answered by the summary characteristics of the posterior distribution of θ. For example, a quantity of interest is the expection,

$$\begin{aligned} \psi_{jg} &= E_{\delta,y}[c_o \delta_{jig} + (1 - \delta_{jig})(c_o + y_{jig})|p_{jg}, \mu_{jg}, \sigma^2_{jg}] \\ &= c_o p_{jg} + (1 - p_{jg}) \left(c_o + \exp\left\{ \mu_{jg} + \frac{\sigma^2_{jg}}{2} \right\} \right). \end{aligned}$$

Extending Basic ANOVA Model

For gene-wise treatment comparisons, e.g. normal versus tumor expression in gene g, the summarize the posterior distribution of,

$$\xi_g = \psi_{2g}/\psi_{1g} \tag{3}$$

for each gene $g = 1, \ldots, G$.

Priors:

$$\begin{aligned} \mu_{jg}|\mu_{jo}, \sigma^2_{jg} &\sim N(\mu_{jo}, \tau_o \sigma^2_{jg}/\bar{n}_j) \\ \sigma^2_{jg} &\sim \text{IG}(a_{jo}, b_{jo}) \\ \mu_{jo} &\sim N(m_{jo}, \nu^2_{jo}) \\ \text{logit}(p_{jg}) &\sim N(u_{jo}, k_{jo}\omega^2_{jo}) \\ u_{jo} &\sim N(\hat{u}_{jo}, h_{jo}\omega^2_{jo}) \end{aligned}$$

where $\bar{n}_j = \frac{1}{G} \sum_{g=1}^{G} \left(n_j - \sum_{g=1}^{n_j} \delta_{jig} \right)$

Extending Basic ANOVA Model

Note this that in this model formulation, the priors induce *a priori* correlation between the genes. It can be shown that $(\mu_{jg}, \mu_{jg'}) \sim N_2(\mu^*, \Sigma^*)$, with $\mu^* = (m_{jo}, m_{jo})'$ and

$$\Sigma^* = \begin{pmatrix} \frac{\tau_o \sigma^2_{jg}}{\bar{n}_j} + \nu^2_{jo} & \nu^2_{jo} \\ \nu^2_{jo} & \frac{\tau_o \sigma^2_{jg'}}{\bar{n}_j} + \nu^2_{jo} \end{pmatrix}.$$

This implies that $\text{Corr}(\mu_{jg}, \mu_{jg'} | \sigma^2_{jg}, \sigma^2_{jg'}, \nu_{jo}) \to 1$ as $\bar{n}_j \to \infty$ or $\nu^2_{jo} \to \infty$, thus borrowing strength across genes.

Extending Basic ANOVA Model

The general gene selection algorithm under the specified model proceeds as,

- Compute posterior distributions of ξ_g's for $g = 1, ..., G$ and find $\gamma_g = P(\xi_g > 1|D)$
- Select a threshold γ_o for γ_g
- If gene g is declared differentially expressed, require $\mu_{1g} \neq \mu_{2g}$, else $\mu_{1g} = \mu_{2g}$, and create a submodel.
- Create several submodels using different γ_o = .7, .8, .9, ...
- Compare models by the L-measure (see Ibrahim and Laud, 1994;Laud and Ibrahim, 1995)
- L-measure defined as:

$$L = E[(z - x)'(z - x)]$$

where the expectation is with respect to the posterior predictive distribution

$$p(Z|D) = \int p(z|\theta) p(\theta|D) d\theta$$

Bayesian Analysis of Variance for Microarrays (BAM)

- Ishwaran and Rao (2003, 2005a, 2005b)
- An extension of the ANOVA model to detect differential expression in genes within a model selection framework
- BAM approach uses a special inferential regularization known as spike-and-slab shrinkage that provides an optimal balance between total false detections and total false non-detections
- Use a parameteric stochastic variable selection procedure first proposed by Mitchell and Beauchamp (1988)
- Recast the problem of finding differentially expressing genes as determining which factors are significant in a Bayesian ANOVA model

Bayesian Variable Selection in Linear Models

- Mitchell and Beauchamp (JASA, 1988)

$$\begin{aligned} Y_i &= x_i^T \beta + \epsilon_i \\ (Y_i | X_i, \beta, \sigma^2) &\sim N(X_i^T \beta, \sigma^2), \quad i = 1, \ldots, n \\ (\beta_g | \gamma_g, \tau_g^2) &\sim N(0, \gamma_g \tau_g^2), \quad g = 1, \ldots, G \\ (\gamma_g | \lambda_g) &\sim (1 - \lambda_g) \delta_{\gamma^*}(\cdot) + \lambda_g \delta_1(\cdot) \\ \lambda_g &\sim \text{U}(0, 1) \\ (\tau_g^{-2} | a_1, a_2) &\sim \text{Gamma}(a_1, a_2) \\ (\sigma^{-2} | b_1, b_2) &\sim \text{Gamma}(b_1, b_2). \end{aligned}$$

where Y_i is the response/gene expression, X_i is the G-dimensional covariate with β as the associated regression coefficients and σ^2 the measurement error

Bayesian Variable Selection in Linear Models

$$\begin{aligned}(Y_i|X_i,\beta,\sigma^2) &\sim N(X_i^T\beta,\sigma^2), \quad i=1,\ldots,n\\ (\beta_g|\gamma_g,\tau_g^2) &\sim N(0,\gamma_g\tau_g^2), \quad g=1,\ldots,G\\ (\gamma_g|\lambda_g) &\sim (1-\lambda_g)\delta_{\gamma^*}(\cdot)+\lambda_g\delta_1(\cdot)\end{aligned} \tag{4}$$

The key feature in this model is that the prior variance $\nu_g^2=\gamma_g\tau_g^2$ on a given coefficient β_g has a bimodal distribution, which is calibrated via the choice of priors on τ_g^2 and γ_g. For example, a large value of ν_g^2 occurs when $\gamma_g=1$ and τ_g^2 is large, thus inducing a large values for β_g, indicating the covariate could be potentially informative. Similarly, small values of ν_g^2 occur when $\gamma_g=\gamma^*$ (fixed to a pre-specified small value), which leads to shrinkage of β_g.

Bayesian Variable Selection in Linear Models

Under the above model formulation, the conditional posterior mean of β is,

$$E(\beta|\nu^2,\sigma^2,Y)=(\sigma^2\Gamma^{-1}+X^TX)^{-1}X^TY, \tag{5}$$

where $\Gamma=\mathbf{diag}(\nu_1^2,\ldots,\nu_G^2)$, $\tau^2=(\tau_1^2,\ldots,\tau_G^2)$ and $Y=(Y_1,\ldots,Y_n)$. This is the (generalized) ridge regression estimate of Y on X with weights $\sigma^2\Gamma^{-1}$. Shrinkage is induced via the small diagonal elements of Γ, which are determined by the posteriors of γ, τ^2 and λ.

BAM

- IR extend this variable selection framework to microarray data, via an ANOVA model and its corresponding representation as a linear regression model
- Note: ANOVA can be written as a regression and vice-versa
- The two-group setting is discussed in Ishwaran and Rao (2003)

BAM

For a group $l=1,2$, let Y_{gil} denote the gene expression from array/individual $i=1,\ldots,n_{g,l}$ of gene $g=1,\ldots,G$. The interest then is to identify differnetially expressed genes between two groups say, control($l=1$) versus treatment group($l=2$). To this end, the ANOVA model can then be written as,

$$Y_{gil}=\theta_{g,0}+\mu_{g,0}I\{l=2\}+\epsilon_{gil}$$

where the errors ϵ_{gil} are asssumed iid $N(0,\sigma^2)$. $\theta_{g,0}$ model the mean of the gth gene in the control group. In this model those genes that are differentially expressed correspond to $\mu_{g,0}\neq 0$ i.e. turned on or off depending on the sign on $\mu_{g,0}$.

BAM

The authors then go through a series of tranformations of the data, before they fit the above model. There are two primary transformation: centering and rescaling the data. They transformed data used for down-stream analysis is,

$$\tilde{Y}_{gil} = (Y_{gil} - \bar{Y}_{g1})\sqrt{n/\hat{\sigma}_n^2}$$

where

$$\hat{\sigma}_n^2 = (n-p)^{-1}\sum_{gil}(Y_{gil} - \bar{Y}_{g2}I\{l=2\} - \bar{Y}_{g1}I\{l=1\})^2$$

is the usual unbaised (pooled) estimator of σ_0^2, $n = \sum_{g=1}^{p} n_j$ is the total number of observations, $\bar{X}_{gl}$ is mean of group l.

1. Centering: reduces the number of parameters and correlation between the model parameters θ_g and μ_g.
2. Rescaling to force the variance σ^2 to be approximately equal to n

BAM

Finally the transformed model that is fit to the data is,

$$\tilde{Y} = \tilde{X}^T\tilde{\beta}_0 + \tilde{\epsilon} \quad (6)$$

where $\tilde{Y}$ is a vector of expression values obtained by concatenating the values $\tilde{Y}_{gil}$ in a vector, $\tilde{\beta}_0$ are the new vector regression coefficients under scaling and $\tilde{\epsilon}$ is the vector of measurement errors. $\tilde{X}$ is the rescaled design matrix such that the second moments are equal to 1 is of dimension $n \times 2p$.

BAM

The effect of these transformations is, for genes that are differentially expressed, to induce a conditional mean and variance for μ_g

$$\mu_g \approx \frac{\sqrt{n_{g,2}}}{\sigma^2_n}(\bar{Y}_{g,l=1} - \bar{Y}_{g,l=2})$$

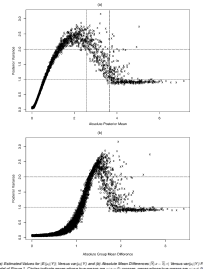
$$\frac{\nu_{2g}^2}{\nu_{2}^2 g + 1} \quad a \approx 1$$

Their "Bayes Test Statistic" is

$$\mu_g^* = \mathbf{E}(\mu_g|\mathbf{Y})\sqrt{\mathbf{n_{g,1}/n_g}}$$

This $\mathbf{E}(\mu_g|\mathbf{Y})$ is the compared to a $N(0, n_{g,1}/n_g)$ distribution to test whether $\mu_{g,0}$ is non-zero. This forms the basis of the *Zcut* procedure for differential gene expression. IR further discuss an extension called *FDRMix* to control the FDR via a hybrid version of the Benjamini and Hochberg (1995) procedure.

BAM Illustration: Shrinkage



BAM Illustration

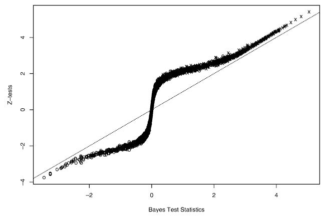
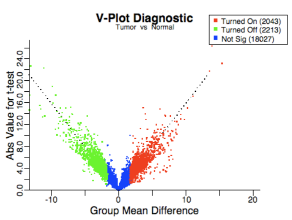


Figure 3. BAM Test Statistics $\mu^*_{j,n}$ Versus $Z_{j,n}$ From Simulations in Figure 1. Expressed genes are represented by crosses; nonexpress by circles. (Nonexpressed genes are the values mostly near 0 on the x-axis that have been shrunken by BAM.)

BAM Example

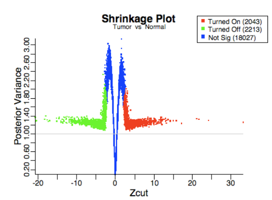
- Lung cancer Affymetrix microarray dataset of Wachi, Yoneda and Wu (2005). Expression values of 22283 genes collected from 10 patients, 5 of whom had squamous cell carcinoma (SCC) of the lung and 5 were normal patients. The dataset is available for download at: `http://www.ncbi.nlm.nih.gov/∼geo/query/acc.cgi?acc=GSE3`
- BAM software available at `http://www.bamarray.com/`

BAM Example



BAM assumes equal variance for each group, and uses a CART variance stabilization algorithm.

BAM Example



Genes that are truly differentially expressed will have posterior variances converge to 1 in the far right and left side of the plot. The cut-off values are determined in a data adaptive manner by balancing the total false detections against total false non-detections.

Summary

- Microarray data: large n small p
- Classification and feature selection
- Frequentist and Bayesian perspectives
- Both have their advantages and disadvantages