**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://veera.mda.tmc.edu/MICROARRAY/MICROARRAY.html

---

**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:  
  - Researchers 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 statisticsian work with:** Gene Expression Matrix

<table>
<thead>
<tr>
<th>Sample</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Sample 2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

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

---

**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
**Statistical Issues with Microarray Data**

- Preprocessing of the data
  - Assures spot quality, reliability of signal, normalization 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 expressed or new subtypes of disease (class discovery)
- Feature (gene) selection

**Dimension Reduction**

- Often in microarrays: \( n \ll p \)
  - 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/Genes 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 objects
- Unsupervised: classes are unknown and want to discover them from data alone
- Supervised: classes are known apriori and want to use a training/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 \( X \).
Suppose there are two populations, healthy and disease individuals. Let the class labels (arbitrary) \( Y_i = 0 \) if individual \( i \) is healthy and \( Y_i = 1 \) if individual \( i \) has disease. The classifier function, \( f_k(x) \), predicts \( Y_i \); given variable \( X_i \).

The function is a mapping from \( X_i \) to the class labels, \( f: X_i \rightarrow \{0, 1\} \).

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 that microarrays will lead to more reliable tumor classification

**Classification and Bayes Rule – I**
- In the (unlikely) situation that we know both \( p_d(x) \) and \( p_v \), we can use Bayes rule to express posterior probability \( p_k(x) \) of \( k \) given a feature gene vector \( x \):
  \[ p_k(x) = \frac{p_d(x)p_k}{\sum_k p_d(x)p_k} \]
- Bayes' rule predicts class with highest posterior probability
  \[ f_k(x) = \text{argmax}_k p_k(x) \]
- Bayes rule minimizes the risk function: misclassification rate under a symmetric loss function – Bayes risk:
  \[ \tau_d = \text{argmax}_k \tau_k(x) \]
Many classifiers can be viewed as versions of this general rule, with either
parametric or nonparametric estimates of \( p(Y|X) \). There are two
general paradigms to estimate \( p(Y|X) \):

- Density estimation approaches, e.g., Gaussian maximum likelihood
discriminant rules (discriminant 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
- ML chooses the parameter value that maximizes the chance of the
  observations, the highest
- For known class conditional densities \( p(Y|X=x) \), the ML rule
  predicts the class \( y \) that gives the largest likelihood to \( x \):
  \[ C_{ML}(x) = \arg\max_y \{ p(Y=y|X=x) \} \]
  - In case of equal class priors: this is same as Bayes Rule
  - Otherwise, ML rule is not optimal \( \alpha \) does not minimize the risk function

**Gaussian Discriminant Rules**

- If we assume multivariate Gaussian (normal) class densities for
  \( X \sim N(\mu_k, \Sigma_k) \), the ML classifier is
  \[ C_{ML}(x) = \arg\max_k \{ N(x|\mu_k, \Sigma_k) \} \]
  - In general, this is a quadratic rule (Quadratic discriminant analysis, or
    QDA) in standard multivariate analysis; function of the Mahalanobis
    distance: \( d(x) = (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \)
  - In practice, population mean vectors \( \mu_k \) and covariance matrices \( \Sigma_k \) are
    estimated by corresponding sample quantities
  - Most common classifiers are variations of the Gaussian discriminant rule

**Discriminant Analysis**

- Fisher Linear Discriminant Analysis (FLDA)
- Finds linear combinations \( d^T(X) \) of the gene expression profiles
  \( X = X_1, ..., X_p \), with large ratios of between-groups to within-groups
  sums of squares: \( \sum \frac{p(Y|X)}{p(Y)} \) 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: \( C(X) = \arg\max_k d^T(X) \) where \( d^T(X) = \sum \frac{p(Y|X)}{p(Y)} \)
  are discriminating variables.
- Standard method in most multivariate statistics books
- Two main steps: (1) Dimension reduction via eigen values (2)
  Classification using the discriminating variables.
- Note: no distribution over \( X \)'s; Nonparametric method

**Classification and Bayes Rule - II**

- There are two general paradigms to estimate \( p(Y|X) \):
  - Density estimation approaches, e.g., Gaussian maximum likelihood
discriminant rules (discriminant 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.
**Common Classifiers**

- QDA: $T(X) = \arg \min_k \left( (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k) + \log |I| - 2 \log \pi_k \right)$
  - Linear discriminant analysis (LDA): if $\Sigma_k = \Sigma$ and $\mu_k$ is constant for all $k$ then $T(X) = \arg \min_k ((X - \mu_k)^T \Sigma^{-1} (X - \mu_k))$
  - Diagonal quadratic discriminant analysis (QDDA): if $\Sigma_k = diag(\sigma^2_1, \ldots, \sigma^2_d)$.
  - Diagonal linear discriminant analysis (LDDA): if $\Sigma_k = diag(\sigma^2_1, \ldots, \sigma^2_d)$.

**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).

**Various Modifications**

- Nearest Centroid ($T(X) = \hat{X}_k$): $G$ 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 closest to sample mean.
  - Estimated Bayes Rule: LDA is Bayes rule with Gaussian distributions.
  - Easy to implement.
  - Reasonable performance: low classification error.

**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 (crude).
  - 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 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

---

**CLASSIFICATION RULES**

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

\[
L(Z|\theta_i) > L(Z|\theta_j)
\]

for all \( j \), and assigned randomly in the event of ties.

There are some disadvantages to this approach. To a Bayesian, \( \theta \) is unknown, and therefore the uncertainty in \( \theta \) should be taken into account when making predictions. See Lehmann (1996) for discussion of bias/variance tradeoff in classification.

---

**REVISE CLASSIFICATION RULES**

Suppose independent random variables (possibly vectors) \( X_1, \ldots, X_N \) are observed from populations \( i = 1, \ldots, K \), each with probability distribution \( \mathcal{F}(\theta_i) \).

The likelihood of the data is

\[
L(X) = \prod_{i=1}^{K} \mathcal{F}(X|\theta_i)
\]

where the \( \theta_i \)'s are unobserved population parameters.

---

**CLASSIFICATION RULES**

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

\[
\pi(\theta_1, \ldots, \theta_K|X) = \prod_{i=1}^{K} \mathcal{F}(X|\theta_i)\pi(\theta_i)
\]

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

\[
L(Z|\theta_i) = \int_\theta L(Z|\theta)\pi(\theta|X)\,d\theta,
\]

for all \( i \), integrating over \( \theta \).
**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 \( f_i(Z|X) \) 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 \( \pi_i f_i(Z|X) \) is obtained.

**BAYESIAN LINEAR CLASSIFIERS**

Suppose that independent random (p-dm) variables \( X_1, \ldots, X_n \) are observed from populations \( i = 1, \ldots, K \), with \( j = 1, \ldots, N \) observations each, with probability distributions \( N(\mu_i, \Sigma_i) \), where \( \mu_i = (\mu_i, \Sigma_i) \) are the unknown population mean and covariance of \( X \).

The likelihood for the data is

\[
p(X | \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) = \prod_{j=1}^{N} \prod_{i=1}^{K} N(X | \mu_i, \Sigma_i)
\]
Bayesian Linear Classifiers

In the context of microarray data, $X_i$ denotes the vector of gene expression intensity values for individual $i$ in population $j$. In typical studies, $k=2$ or $k=3$, 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 $\beta_1, \ldots, \beta_k, \Sigma_1, \ldots, \Sigma_k$ is

$$p(\beta_1, \ldots, \beta_k, \Sigma_1, \ldots, \Sigma_k) \propto \prod_{j=1}^k |\Sigma_j|^{1/2}.$$ 

Example

**Example** Consider the case of two populations (classes), where $\alpha_j = \Pr(\theta = j)$. The frequentist rule is to assign $x$ to class $j$ if

$$P(x | \theta = j) > P(x | \theta = k),$$

where $P(x | \theta = j) = \alpha_j f_j(x)$ is the density of the $j$th population, and $f_j(x)$ is the known probability density function of the $j$th population.

**Bayesian Linear Classifiers**

The predictive distribution of a new observation $Z$ is

$$p(Z | \Sigma_1, \Sigma_2) = \frac{N_1}{(N_1 + 1)(N_2 + 1)} \left[ 1 + \frac{N_1(x - X_j)^T \Sigma_1^{-1}(z - X_j)}{(N_1 + 1)(N_2 + 1)} \right]^{-(N_1 + N_2 + 2)/2},$$

where $N_j = \sum_{i=1}^N X_{ij}$ and $\Sigma_j = \sum_{i=1}^N (X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)^T$. 

**Proof**: see Press (2003) *Bayesian Statistics*

**Example Con’t**

$$L_2 = \frac{(N_i + 1)\beta_i + (N_j + 1)\beta_j}{(N_i - N_j - 2)},$$

where

$$L_2 = \frac{(N_i - 1)\beta_i + (N_j - 1)\beta_j}{(N_i - N_j - 2)}.$$ 

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

$$P_2 = \frac{1 + \langle \delta \rangle}{2}$$

with 1 analogously.

Pug (1997, JASA)
Suppose some subset of genes from the microarray are truly 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 features 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**

For any given feature set, of size \( p \),

\[
\ell_{(i)}(X \mid \gamma, \beta) = \frac{N!}{\gamma_1! \cdots \gamma_k! (N-k)!} \left( \frac{1}{(N-k)} \right)^{N-k} \left( \frac{1}{N-k} \right)^{\gamma_1 + \cdots + \gamma_k} \]

where \( X_i \) and \( S_i \) are derived from the selected subset of genes.

Accounting for uncertainty in feature selection involves integrating over the posterior distribution of \( \beta \).

\[
\ell_{(i)}(X) = \int \ell_{(i)}(X \mid \gamma, \beta) \pi(\beta) d\beta
\]

for \( i = 1, \ldots, K \).

**Feature Selection revisited**

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

\[
\tilde{\ell}(\gamma, \beta) = \prod_{i=1}^N \ell_{(i)}(X \mid \gamma, \beta)
\]

and \( \ell(X) \) may be obtained by

\[
\ell(X) = \frac{\tilde{\ell}}{\sum_{\gamma, \beta} \tilde{\ell}(\gamma, \beta)}
\]

where

\[
\tilde{\ell} = \int \ell_{(i)}(X \mid \gamma, \beta) \pi(\beta) d\beta.
\]
**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 (1998) for discussion of bias/variance tradeoff in classification.

**Detour: Bayesian Analysis of a Linear Model**

The linear model is frequently 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, \( \beta \) is a \( p \times 1 \) vector of coefficients (unobserved) and

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

Let \( M = X(X'X)^{-1}X' \), and \( \tau = \sigma^{-2} \), where \( - \) denotes generalized inverse. Recall that the UMVUE of \( \mu = E(Y) = X\beta \) is \( M\hat{Y} \). We would like to derive the posterior distribution of \( \beta \) and \( \tau \) under noninformative priors.

**Theorem 1**

Suppose \( \tau \) is known, \( X \) is of full rank \( p \), and \( \pi(\beta) \propto 1 \).

Then

\[ \beta|\tau \sim N_p \left( \tau^{-1} \mu, \tau^{-1} M \right) \]

where

\[ \hat{\beta} = (X'X)^{-1}X'Y \].
Bayesian Analysis of a Linear Model

Proof:

\[
\begin{align*}
p(\beta|y, \tau) &\propto \exp\left\{-\frac{\tau^2}{2} (Y - X\beta)'(Y - X\beta)\right\} \\
&= \exp\left\{-\frac{\tau^2}{2} (Y - Mu)'(Y - Mu) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta})\right\} \\
&= \exp\left\{-\frac{1}{2} (\beta - \hat{\beta})'X'(\beta - \hat{\beta})\right\}.
\end{align*}
\]

Note that

\[
\begin{align*}
(Y - Mu)'(Y - Mu) &= (Y - Mu)'(Y - Mu) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) \\
&= (Y - Mu)'(Y - Mu) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) \\
&= (Y - Mu)'(Y - Mu) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) \\
&= (Y - Mu)'(Y - Mu) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) + (\beta - \hat{\beta})'X'(\beta - \hat{\beta}).
\end{align*}
\]

Bayesian Analysis of a Linear Model

Theorem 2

When \( \tau \) is known, Jeffreys prior for \( \beta \) is a uniform prior, i.e.,

\[ p(\beta) \propto 1. \]

Proof:

\[
\begin{align*}
\frac{d}{d\beta} \log p(\beta|\tau) &= -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\tau) - \frac{1}{2} (Y - X\beta)'(Y - X\beta) \\
&= -\frac{1}{2} \frac{d}{d\beta} \left[ \frac{1}{2} (Y - X\beta)'(Y - X\beta) \right] \\
&= -\frac{1}{2} \frac{d}{d\beta} \left[ \frac{1}{2} (Y - X\beta)'(Y - X\beta) \right] \\
&= -\frac{1}{2} \frac{d}{d\beta} \left[ \frac{1}{2} (Y - X\beta)'(Y - X\beta) \right].
\end{align*}
\]
Bayesian Analysis of a Linear Model

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

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

**Proof:** Exercise

**Theorem 4**
Consider the linear model with both $\beta$ and $\tau$ unknown, and suppose

$$\pi(\beta, \tau) \propto \tau^{-3}.$$ 

Then

$$\beta|\tau \sim \mathcal{N}(n - p, \hat{\beta}, \tau(X'X)^{-1})$$

where $\hat{\beta} = Y'(I - M)Y/(n - p)$ and $t|\nu \sim \text{gamma}(n - p)/2, \tau^2(n - p)/2$.

**Proof:** We have

$$p(\beta|\tau) = \tau^{n-1} \exp\left\{-\frac{1}{2} \tau \left[ (Y - X\beta)'(Y - X\beta) \right] \right\}$$

and

$$p(\tau) = \tau^{-3} \exp\left\{-\frac{1}{2} \tau \left[ (Y - X\hat{\beta})'X(X'X)^{-1}X(Y - X\hat{\beta}) \right] \right\}.$$

Thus $t|\nu \sim \text{gamma}(n - p)/2, \tau^2(n - p)/2$.

Bayesian Analysis of a Linear Model

Thus, for $a(0)$,

$$a(0) = \int e^{0^3(\beta)} \left\{ \frac{1}{2} \left[ (Y - X\beta)'(Y - X\beta) \right] \right\} \, d\beta = \left[ 0^3 + (0 - \beta)'X(X'X)^{-1}X(0 - \beta) \right]^{-1/2}.$$ 

Let $\hat{\beta} = Y'(I - M)Y/(n - p)$. Then the above integral is

$$\left[ (n - p)^2 + (n - p - 1) \right]^{-1/2}.$$ 

Thus, $\beta|\tau \sim \mathcal{N}(n - p, \hat{\beta}, \tau(X'X)^{-1})$.
Bayesian Analysis of a Linear Model

**Theorem 5**
Consider the linear model with $\beta$ and $\tau$ unknown, and suppose

$$
\beta | \tau \sim N(\mu_\beta, \tau^{-1}\Sigma_\beta) \\
\tau \sim \text{gamma} \left(\frac{1}{2}, \frac{1}{2}\right)
$$

Then $\beta | y \sim N(\mu_\beta + (1 - n\lambda)\bar{y}, \tau^{-1}(X'X + \Sigma_\beta)^{-1})$, where

$$
\beta' = A\mu_\beta + (1 - n\lambda)\bar{y}, \\
A = (X'X + \Sigma_\beta)^{-1}X'y, \\
\beta = (X'X)^{-1}X'y.
$$

**Proof:** Exercise

**Hint:**
$$
\tau(\beta, \tau | Y) \propto \left(\frac{n + \delta_\tau}{2}\right)^{\frac{1}{2}}
$$

where

$$
\delta_\tau = \max(0, n - 3) + \frac{1}{2}.
$$

---

Bayesian Analysis of a Linear Model

$$
Q = (Y - X\beta)'(Y - X\beta) + (\beta - \mu_\beta)'\Sigma_\beta^{-1}(\beta - \mu_\beta) + \delta_\lambda
$$

Notice that

$$
(Q - X\beta)'Q^{-1}(Q - X\beta) = (x'X'X - x'X'y + y'y - y'y - y'y) + y'y
$$

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

$$
Q = 3x'X'X - 2x'X'y - x'y'y + y'y
$$

$$
= \frac{3}{2}(I_X'X)^{-1} - y'y + \frac{3}{2}(I_X'X)^{-1} - y'y + \frac{3}{2}(I_X'X)^{-1} - y'y
$$

Bayesian Analysis of a Linear Model

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

$$
\Sigma_\beta = (X'X + \Sigma_\beta)^{-1}
$$

$$
\beta' = (X'X + \Sigma_\beta)^{-1}(X'y + \Sigma_\beta^{-1}\mu_\beta)
$$

$$
\beta'' = (X'X + \Sigma_\beta)^{-1}(X'y + \Sigma_\beta^{-1}\mu_\beta)
$$
The One-Way ANOVA model, for gene $g$ is defined for a single response vector $Y_g$ as

$$Y_g = X_g^T \beta_g + \epsilon_g$$

(1)

where $X^T$ is a matrix of indicator variables for $f = 1, \ldots, k$ treatments ($k = 2$ often, in marker studies) and $\beta_g$ is the $k$-dimensional (unknown) vector of treatment effects for gene $g$, and $\epsilon_g$ is the unknown variance of $\epsilon_g$.

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 naively.

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 $j = 1, \ldots, k$ treatment groups. Variable selection algorithms for high-dimensions are discussed in work by:

- George and McCulloch (1997): Bayesian variable selection via Gibbs Sampling
- Storey (2003): FDR based
- Lee (2003); Shi (2006): Probit binary/multinomial regression with variable selection
- Ibrahim, Chen and Gray (2002): Threshold models

Bayesian ANOVA Models for Gene Expression Data

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 patterns 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
Lognormal Model for Gene Expression

Let \( \mathbf{x} \) denote the gene expression where \( i \) indexes the tissue type, e.g., 1 = normal, 2 = tumor, \( j \) indexes the individual, \( j = 1, \ldots, n_j \), and \( g \) indexes the gene, \( g = 1, \ldots, G \). Similarly, \( y_{ijg} \) denote the continuous component of the gene expression level for the \( i \)th tissue type for the \( j \)th individual and the \( g \)th gene.

Assume \( y_{ijg} \) are independently log-normal distributed as,

\[
P(y_{ijg} | \mu_{ijg}, \sigma^2_{ijg}) = \frac{1}{\sqrt{2\pi} \sigma_{ijg}} \exp\left(-\frac{(\ln(y_{ijg}) - \mu_{ijg})^2}{2 \sigma^2_{ijg}}\right)
\]

Let \( \lambda_{ijg} = 1 \) if \( y_{ijg} = \epsilon_{ijg} \) and 0 otherwise. Further, the prior probability

\[
P(\lambda_{ijg} = 1) = P(y_{ijg} = \epsilon_{ijg}) = \rho_{ijg}^\lambda
\]


ewpage

EXTENDING BASIC ANOVA MODEL

Model gene expression \( x \) as,

\[
x = \begin{cases} 
  c_x & \text{with probability } p \\
  c_x + y & \text{with probability } 1 - p
\end{cases}
\]

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

EXTENDING BASIC ANOVA MODEL

Let \( \mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_G) \) be the collection of all parameters for \( j = 1, 2 \) and \( g = 1, \ldots, G \). Then conditional on the observed data \( \mathbf{D} = (x, \hat{d}) \), the likelihood for \( \mathbf{y} \) is given by,

\[
L(d | \mathbf{y}) = \prod_{j} p(y_j | x_j, \mu_j, \sigma_j^2)^{1/2} y_j^{1/2} \exp\left(-\frac{y_j}{2 \sigma_j^2}\right)
\]

With this formulation, all the fundamental questions can be answered by the summary characteristics of the posterior distribution of \( \mathbf{y} \). For example, a quantity of interest is the expectation,

\[
E(y_{ijg}) = E(x_{ijg}) + (1 - \delta_{ijg})(c_x + y_{ijg}) \left( \hat{\sigma}_{ijg}^2 + \frac{\sigma^2_{ijg}}{2} \right)
\]
Extending Basic ANOVA Model

Note that in this model formulation, the priors induce a prior correlation between the genes. It can be shown that
\[
(p, r^2) \sim N(\mu^*, \Sigma^*), \quad \mu^* = (m_0, m_0^2)^T \text{ and }
\]
\[
\Sigma^* = \begin{pmatrix}
\frac{\sigma^2}{\nu_0} & \frac{\nu_0}{\nu_0 + \nu_1} \nu_1^* \\
\frac{\nu_0}{\nu_0 + \nu_1} \nu_1^* & \frac{\nu_1}{\nu_0 + \nu_1}
\end{pmatrix}
\]
This implies that \(\text{Corr}(\mu_0, r^2) \rightarrow 1 \text{ as } \nu_0 \rightarrow \infty \text{ or } \nu_1^* \rightarrow \infty\), thus borrowing strength across genes.

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 parametric 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)
\[
Y_i = x_i^T \beta + \epsilon_i
\]
\[
(Y_i; x_i, \beta, \sigma^2) \sim N(x_i^T \beta, \sigma^2)
\]
\[
(\beta_p; \sigma^2) \sim N(0, \sigma^2 \gamma_p^2)
\]
\[
(\gamma_p; \lambda_p) \sim \Gamma(1 - \lambda_p, \lambda_p) = \text{Beta}(1, \lambda_p)
\]
\[
(\tau_i; \alpha_i, \beta_i) \sim \Gamma(\alpha_i, \beta_i)
\]
\[
(\sigma^2; \alpha, \beta) \sim \Gamma(\alpha, \beta)
\]
where \(Y_i\) is the response/gene expression, \(X_i\) is the \(G\) dimensional covariate with \(j\) as the associated regression coefficients and \(\sigma^2\) the measurement error

Extending Basic ANOVA Model

The general gene selection algorithm under the specified model proceeds as,
- Compute posterior distributions of \(\gamma_g\) s for \(g = 1, \ldots, G\) and find \(\gamma_{g*} = P(\gamma_g > 1|D)\)
- Select a threshold \(\gamma_{g*}\) for \(\gamma_g\)
- If gene \(g\) is declared differentially expressed, require \(\mu_g \neq 0\) and \(\mu_g = 0\) and create a submodel
- Create several submodels using different \(\gamma_g\) s = 7, 8, 9, ...
- Compare models by the L-measure (see Ibrahim and Laud, 1994, Laud and Ibrahim, 1995)
- L-measure defined as:
\[
L = E[(x - \hat{x})^2]/(x - x)^2
\]
where the expectation is with respect to the posterior predictive distribution
\[
p(Z|D) = \int p(z|\theta)p(\theta|D)d\theta
\]
BAYESIAN VARIABLE SELECTION IN LINEAR MODELS

\[(Y_i|X_i, \beta, \sigma^2) \sim N(X_i^T \beta, \sigma^2), \quad i = 1, \ldots, n \]

\[(\beta_g|\gamma_g, \tau_g^2) \sim N(0, \gamma_g \tau_g^2), \quad \beta \sim (1 - \lambda_g) \gamma_g \tau_g^2 + \lambda_g \sigma^2 \]  

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

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)

BAYESIAN VARIABLE SELECTION IN LINEAR MODELS

Under the above model formulation, the conditional posterior mean of \(\beta\) is

\[E(\beta|\gamma, \tau^2, Y) = -\delta^{-1} + X^T \delta^{-1} Y, \]  

where \(\delta = \text{diag}(\gamma_1^2, \ldots, \gamma_p^2), \tau^2 = (\tau_1^2, \ldots, \tau_p^2)\) and \(Y = (Y_1, \ldots, Y_n)\). This is the (generalized) ridge regression estimate of \(Y\) on \(X\) with weights \(\tau^2\). Shrinkage is induced via the small diagonal elements of \(\delta\), which are determined by the posteriors of \(\gamma\), \(\tau^2\) and \(\lambda\).

BAM

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

\[Y_g = \beta_g + \epsilon_g, \quad I = 1, 2 \]  

where the errors \(\epsilon_g\) are assumed iid \(N(0, \sigma^2)\). \(Y_g\), model the mean of the \(g\)th gene in the control group. In this model those genes that are differentially expressed correspond to \(\beta_g \neq 0\) i.e. turned on or off depending on the sign of \(\beta_g\)
The authors then go through a series of transformations of the data, before they fit the above model. There are two primary transformation: centering and rescaling the data. They transformed data used for downstream analysis is:

$$\tilde{Y}_{it} = (Y_{it} - \bar{Y}_i)/\sqrt{\sigma^2}$$

where

$$\sigma^2 = (n-1)^{-1} \sum_{it} (Y_{it} - \bar{Y}_i)^2$$

is the usual unbiased (pooled) estimator of $$\sigma^2$$; $$n$$ is the total number of observations, $$\bar{Y}_i$$ is mean of group $$i$$.

Centering: reduces the number of parameters and correlation between the model parameters $$\tilde{y}_i$$ and $$\mu_y$$.

Rescaling to force the variance $$\sigma^2$$ to be approximately equal to $$n$$.

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

$$\tilde{Y} = X\tilde{\beta} + \tilde{\varepsilon}$$

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

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

$$\mu_y = \sum_{i=1}^{n} \rho_{i} (\tilde{y}_{it} - \bar{Y}_{i})$$

$$\sigma^2_{\mu_y} = \frac{\sigma^2}{2p+1} \quad a = 1$$

Then “Bayes Test Statistic” is

$$\tilde{\nu}^2 = X(\mu_y|Y)\|\tilde{\nu}\|$$

This $$\nu^2(\mu_{Y})$$ is the compared to a $$\nu^2_{n-1}/\nu_{p}$$ distribution to test whether $$\mu_y$$ is non-zero. This forms the basis of the Z-stat procedure for differential gene expression, (if further discuss an extension called FORMA to control the FDR via a hybrid version of the Benjamin and Hochberg (1995) procedure.)

BAM illustration: Shrinkage
BAM ILLUSTRATION

BAM EXAMPLE

Lung cancer Affymetrix microarray dataset of Weiss, Yereda 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:

BAM software available at http://www.bamarray.com/

BAM EXAMPLE

BAM EXAMPLE

BAM EXAMPLE

BAM EXAMPLE

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

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