Advanced Statistical Methods for the Analysis of Gene Expression and Proteomics
Nonlinear Methods for 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

Till now...

- Microarray Classification
- Various approaches: last lecture
  - Linear/Quadratic Discriminant Analysis.
  - Maximum Likelihood Discriminant Rules
  - Bayesian linear classifiers; Linear models for Differential expression
- Today: Nonlinear Methods
  - Regression Methods: Generalized (Non)-linear Models (GLMs)
  - Splines; SVM; Kernel methods
  - Theory motivated in a Bayesian framework but estimation can be any method.

Generalized Linear Models

The class of generalized linear models is a natural generalization of the classical linear model. Generalized linear models include as special cases, linear regression and analysis of variance models, logit and probit models for quantal response data, log-linear models and multinomial response models for counts, some commonly used models for survival data.

To simplify the transition from the classical normal linear model, i.e. $Y = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I)$ to generalized linear models, it will be important to characterize specific aspects of the **linear** model

Generalized Linear Models

1. **Random component:** $Y \sim N_n(\mu, \sigma^2 I)$, where $\mu = X\beta$. Note that the linear model has constant variance.
2. **Systematic component:** The **covariate** comprises the systematic component of the model. For the i^{th} observation, we let

$$\eta_i = x_i'\beta, \quad i = 1, ..., n.$$

We call η_i the **linear predictor**.

Generalized Linear Models

Thus $y_i \sim N(x_i'\beta, \sigma^2) = N(\eta_i, \sigma^2)$, $i = 1, ..., n$ and the y_i's are independent, given the x_i's and β. Note here that for the usual normal linear model, the relationship between the **mean** of y_i and η_i is given by

$$\mu_i \equiv E(y_i|x_i, \beta) = x_i'\beta = \eta_i, \quad i = 1, ..., n.$$

Thus

$$\mu_i = \eta_i, i = 1, ..., n.$$

Generalized linear models involve **2 extensions** of the normal linear model.

Generalized Linear Models

1. The distribution of y is from the **exponential family**
2. The relationship between $\mu_i = E(y_i|x_i, \beta)$ can be made more general, so that

$$g(\mu_i) = \eta_i \equiv x_i'\beta$$

$g(\mu_i)$ is called the μ**-link** function and relates the **mean** of y_i (i.e., μ_i) to the linear predictor η_i. y has a distribution in the exponential family with **canonical parameter** θ and dispersion ϕ

$$p(y|\theta, \phi) = \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\right\}$$

Generalized Linear Models

Without loss of generality, we assume $a(\phi) = \phi$, so that

$$p(y|\theta, \phi) = \exp\left\{[y\theta - b(\theta)]/\phi + c(y, \phi)\right\}.$$

Here

$$\int_y \exp\left\{[y\theta - b(\theta)]/\phi + c(y, \phi)\right\} dy = 1,$$

so that

$$\exp\left\{\frac{b(\theta)}{\phi}\right\} = \int_y \exp\left\{\frac{y\theta}{\phi} + c(y, \phi)\right\} dy.$$

Generalized Linear Models

Here $b(\cdot)$ and $c(\cdot)$ are **known** functions. If ϕ is unknown, then the above may or may **not** be an exponential family. θ is called the **canonical parameter**. An excellent book on generalized linear models is McCullagh & Nelder (Chapman Hall).
The class of generalize linear models has many uses in biostatistics. Binomial models are often used to model dose response. Gamma models are often used to model survival or time-to-event data. Poisson models are used to model count data, such as yearly pollen counts, number of cancerous nodes, etc.
Distributions included in the exponential family are the normal, binomial, gamma, poisson, beta, multinomial, and inverse gaussian distributions.

To see how the normal distribution, for example, fits into the framework above, suppose,

$$y \sim N(\mu, \sigma^2).$$

Then

$$\begin{aligned} p(y|\mu,\sigma^2) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\left(y\mu - \mu^2/2\right)/\sigma^2 - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \end{aligned}$$

so that in this case,

$$\begin{aligned} \theta &= \mu \\ a(\phi) &\equiv \phi = \sigma^2 \\ b(\theta) &= \frac{\theta^2}{2} \\ c(y,\phi) &= -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]. \end{aligned}$$

- Similar representations exist for Binomial, Poisson, Gamma etc.
- For Binomial it turns out that $\boldsymbol{b(\theta) = log(1 + e^{\theta})}$ and hence the transformation $\boldsymbol{log(\frac{p}{1-p})}$ is called the logit transformation.
- One can prove that in general

$$\begin{aligned} \boldsymbol{E(y|\theta,\phi)} &= \boldsymbol{b'(\theta)} \\ \boldsymbol{V(y|\theta,\phi)} &= \boldsymbol{\phi b''(\theta)} \end{aligned}$$

- Thus once we know the $\boldsymbol{b(.)}$ function, we can get the mean and variance of the exponential family model.

Now suppose we have n independent observations $y_1, ..., y_n$ from an exponential family. Then the density for the i^{th} observation can be written as

$$p(y_i|\theta_i,\phi) = \exp\left\{\phi^{-1}(y_i\theta_i - b(\theta_i)) + c(y_i,\phi)\right\}.$$

The density based on n observations is

$$p(y|\theta,\phi) = \prod_{i=1}^{n} p(y_i|\theta_i,\phi),$$

where $y = (y_1, ..., y_n)$, $\theta = (\theta_1, ..., \theta_n)$.

GENERALIZED LINEAR MODELS

To construct the regression model, (i.e., the generalized linear model), we let the θ_i's depend on the linear predictor $\eta_i = x_i'\beta$ through the equation

$$\theta_i = \theta(\eta_i), \text{for} \quad i = 1, ..., n,$$

i.e., the link function $\theta(\cdot)$, where $x_i' = (x_{i1}, ..., x_{ip})$, and $\beta = (\beta_1, ..., \beta_p)'$. The link function is called the θ-**link** and is often more convenient to use than the μ-link. The θ-link is a one-to-one function of the μ-link. Once $\theta_i = \theta(\eta_i)$ is given, one can write the likelihood function as a function in (β, ϕ). When $\theta_i = \eta_i$, we say that we have a **canonical link**. The function $\theta_i = \theta(\eta_i)$ can be any **monotonic** function.

GENERALIZED LINEAR MODELS

Example

Suppose $y_i \sim$ Binomial(1,p_i), the y_i's are independent, $i = 1, ..., n$. We have

$$\begin{aligned} p(y_i|p_i) &= \exp\left\{y_i \log\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{1}{1-p_i}\right)\right\} \\ &= \exp\left\{y_i\theta_i - \log\left(1+e^{\theta_i}\right)\right\}. \end{aligned}$$

If a canonical link is used, the we set $\theta_i = \eta_i = x_i'\beta$. Substituting $\theta_i = x_i'\beta$ into $p(y_i|p_i)$ above, we get

$$p(y_i|\beta) = \exp\left\{y_i x_i'\beta - \log\left(1+e^{x_i'\beta}\right)\right\}.$$

GENERALIZED LINEAR MODELS

Thus, the likelihood function of β based on all n observations is given by

$$\begin{aligned} p(y|\beta) &= \prod_{i=1}^{n} p(y_i|\beta) \\ &= \prod_{i=1}^{n} \exp\left\{y_i x_i'\beta - \log\left(1+e^{x_i'\beta}\right)\right\}. \\ &= \exp\left[\sum\left\{y_i x_i'\beta - \log\left(1+e^{x_i'\beta}\right)\right\}\right] \end{aligned}$$

GENERALIZED LINEAR MODELS

For this model, the relation between θ_i and μ_i is

$\theta_i = \log(\frac{\mu_i}{1-\mu_i})$, where $\mu_i = E(y_i|p_i) \equiv p_i$.

Thus $\mu_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$. Suppose, we consider a **probit model**. The μ-**link** for the probit model is given by

$$\begin{aligned} \Phi^{-1}(\mu_i) &= \eta_i \\ \mu_i &= \Phi(\eta_i) \\ \eta_i &= x_i'\beta, \\ \Phi(\eta_i) &= \frac{e^{\theta_i}}{1+e^{\theta_i}}. \end{aligned}$$

Generalized Linear Models

Any model that satisfies

$$p(y_i|\theta_i, \phi) = \exp\left\{\phi^{-1}(y_i\theta_i - b(\theta_i)) + c(y_i, \phi)\right\}$$

and $\theta_i = \theta(\eta_i)$, $\eta_i = x_i'\beta$, is called a generalized linear model (GLM). Below we give some distributions with their canonical links.

Distribution	**Canonical μ-link**
Normal	$\eta = \mu$
Poisson	η =log(μ)
Binomial	$\eta = \log(\frac{\mu}{1-\mu})$
Gamma	$\eta = \mu^{-1}$

Estimation in GLM's

- Frequentist inference
  - MLE of β does not have closed form; Newton-Raphson or Fisher Scoring used
  - The resulting equations are non-linear functions of β
  - The likelihood equations are of β are independent of ϕ
  - Often use Large Sample theory for Hypothesis testing
- Bayesian inference
  - Put prior on β
  - No conjugate priors exist; posteriors not of closed form
  - However in most cases they are log-concave: attractive methods exist to sample from them: Adaptive Rejection sampling (Gilks and Wild (1992, Applied Statistics)

Bayesian Model Selection in GLM's

The computation of Bayes factor, HPD intervals, or posteriormodelprobabilities will require MCMC techniques since the posterior distributions are not available in closed form. It turns out that some novel MCMC algorithms can be developed for computing posterior model probabilities, in cases in which noninformative priors or informative priors are used. We now discuss some of these methods.

A popular method for computing posterior model probabilities using non-informative (but proper) priors was developed by George and McCulloch (1993, *JASA*), and George, McCulloch and Tsay (1996).

Bayesian Model Selection in GLM's

Consider the model

$$Y = X\beta + \epsilon, \;\; \epsilon \sim N_n(0, \sigma^2 I).$$

George, McCulloch and Tsay consider a prior for each β_i, $\beta = (\beta_1, ..., \beta_p)'$ to be a mixture of two normal densities, and thus

$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2\tau_i^2),$$

where γ_i is a binary random variable with

$$p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = p_i.$$

Bayesian Model Selection in GLM's

Note that when $\gamma_i = 0$, $\beta_i \sim N(0, \tau_i^2)$ and when $\gamma_i = 1$, $\beta_i \sim N(0, c_i^2\tau_i^2)$. The interpretation of this is as follows. Set $\tau_i(\tau_i > 0)$ small so that is $\gamma_i = 0$, then β_i would probably be so small that it could "safely" be estimated by 0. Second, if $c_i(c_i > 1$ always) is set large so that if $\gamma_i = 1$, then a non-zero estimate of β_i would probably be included in the model. Thus, the user must specify (τ_i, c_i), for $i = 1, ..., p$. Note here, that a priori, the β_i's are **not** necessarily independent.

Bayesian Model Selection in GLM's

Based on this interpretation, p_i may not be thought of as the prior probability tha β_i is **not** zero, or equivilantly that X_i should be included in the mode, where X_i denotes the i^{th} covariate. The mixture prior for $\beta_i|\gamma_i$ can be written in vector form as

$$\beta \sim \gamma \sim N_p(0, D_\gamma R D_\gamma),$$

where $\gamma = (\gamma_1, ..., \gamma_p)$, R is the prior correlation matrix and

$$D_\gamma = \text{diag}(a_1\tau_1, ..., a_p\tau_p),$$

where $a_i = 1$ if $\gamma_i = 0$ and $a_i = c_i$ if $\gamma_i = 1$. Thus D_γ determines the scaling of the prior covariance matrix.

Back to Microarrays

Now back to Microarrays....

Bayesian Probit Classification

Consider C-classes with class labels $y_i \in \{1, 2, \ldots, C\}$, for $i = 1, \ldots, n$ individuals with associated p covariate measurements $x_i = (x_{i1}, \ldots, x_{ip})$. The idea is to fit classifier model that can predict the class (label) well given the p measurements.

Binary or multinomial regression using GLMS is popular, although inference using Bayesian GLMs is not trivial in practice, as conjugate priors do not exist.

Bayesian Probit Regression

For binary classification, $y \in \{0, 1\}$ we write,

$$\begin{aligned} f(y|\beta) &= [\pi(\beta)]^y [1 - \pi(\beta)]^{1-y} \\ \pi(\beta) &= \Phi(\eta), \eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j. \text{ (probit)} \end{aligned}$$

Other choices are logit and log-log link functions. There are no conjugate priors & computation can be difficult.

Albert and Chib (1993) demonstrated an auxillary variable approach to simplify binary probit regression.

Bayesian Probit Regression

Auxillary variables method
Define $z = \eta + \epsilon$, where $\epsilon \sim N(0, 1)$. Then

$$y = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \end{cases}$$

and the marginal distributions of y is

$$\begin{aligned} & p(y = 1|\beta) \\ & = p(y = 1|z > 0, \beta)p(z > 0|\beta) + p(y = 1|z < 0, \beta)p(z < 0|\beta \\ & = 1 \cdot p(z > 0|\beta) + 0 \cdot p(z < 0|\beta) \\ & = p(z - \eta > -\eta|\beta) = \Phi(\eta). \end{aligned}$$

Conditional upon the auxillary variable z, β is updated.

Bayesian Probit Regression

We have $\mathcal{D} = \{y_i, \mathbf{x}_i\}_1^n$ and $\mathbf{z} = (z_1, \ldots, z_n)$. The hierarchical model is

$$\begin{aligned} y_i|z_i, \beta &\sim I(z > 0)\delta_1 \\ z_i &\sim N(\mathbf{x}_i'\beta, \sigma^2) \\ \ldots &\quad \ldots \\ \beta &\sim N(\mu, \sigma^2 V) \\ \sigma^2 &\sim IG(a, d) \end{aligned}$$

Bayesian Probit Regression

Sample proceeds by sampling all of the parameters conditional on $\mathbf{z}$ and then sampling $\mathbf{z}$ conditional upon $\mathbf{y}$ from a truncated normal distribution:

$$\begin{aligned} p(z_i|y_i = 1, \beta) &\propto p(y_i = 1|z_i, \beta)p(z_i|\beta) \\ &= I(z > 0) \cdot N(\mathbf{x}_i'\beta, \sigma^2), \end{aligned}$$

and

$$\begin{aligned} p(z_i|y_i = 0, \beta) &\propto p(y_i = 0|z_i, \beta)p(z_i|\beta) \\ &= I(z < 0) \cdot N(\mathbf{x}_i'\beta, \sigma^2). \end{aligned}$$

Bayesian Probit Regression

Suppose you want to sample $z \sim N(\mu, \sigma^2) \cdot I(a < z < b)$. This can be accomplished by

1. Setting $u_1 = \Phi(a; \mu, \sigma^2)$ and $u_2 = \Phi(b; \mu, \sigma^2)$
2. Sampling $u \sim U(u_1, u_2)$
3. Setting $z = \Phi^{-1}(u; \mu, \sigma^2)$

Bayesian Probit Regression

How do we classify?
Suppose we have $\mathbf{x}_i$ from $i = 1, ..., m$ individuals (think of the binary responses z_i as missing). Given $\mathbf{x}_i, i = m+1, \ldots, n$, we want to assign class labels to the remaining individuals.

Given the sampled parameters from the posterior distributions based on the first m individuals, we sample $z_i, i = m+1, \ldots, n$. If the estimated $\hat{z}_i > 0$, then $\hat{y}_i = 1$ and 0 otherwise.

Bayesian Probit Regression

Example: Feature Selection (Lee et al. 1993 *Bioinformatics*)
The probit model was used

$$P(Y_i = 1) = \Phi(X_i'\beta)$$

where X_i are measured gene expression values for the ith individual. The variable γ is introduced, such that $\gamma_j = 0$ if $\beta_j = 0$ and $\gamma_j = 1$ if $\beta_j \neq 0$. Conditional upon γ, the prior for β is

$$\beta_\gamma \sim N(0, c(X_\gamma^T X_\gamma)^{-1})$$

for some positive scaler constant c.

Bayesian Probit Regression

The γ_j's are tacken to be *a priori* independent with

$$p(\gamma_j) = \pi_j$$

for π_j small.

Sampling

1. Initialize $[\gamma^{(0)}, Z^{(0)}, \beta^{(0)}]$
2. Draw $\gamma^{(1)}$ from $p(\gamma|Z^{(0)})$
3. Draw $Z^{(1)}$ from $p(Z|\gamma^{(1)}, \beta^{(0)})$
4. Draw $\beta^{(1)}$ from $p(\beta|\gamma^{(1)}, Z^{(1)})$
5. Repeate 2-4 for $b = 2, ..., B$ iterations.

The MC estimateif the $P(Y_{new} = 1|X)$ is

$$\hat{p}(Y_{new}|X) = \frac{1}{m} \sum_{b=1}^{B} p(Y_{new} = 1|X, Z^{(b)}, \beta^{(b)}, \gamma^{(b)})$$

Model Comparison by Cross Validation

1. Model 1 : Use all strongly significant genes
2. Model 2 : Use genes with selected more than 5%
3. Model 3 : Use genes with selected more than 6%
4. Model 4 : Use genes with selected more than 7%

Table 1. Breast cancer data: strongly significant genes for the classification of BRCA1 versus BRCA2 or sporadic

Frequency* (%)	Image clone ID	Gene description
8.6	897781	keratin 8
8.4	823940	TOB1
7.8	26184	'phosphofructokinase, platelet'
7.5	840702	SELENOPHOSPHATE SYNTHETASE; Human selenium donor protein
7.1	376516	cell division cycle 4-like
6.9	47542	small nuclear ribonucleoprotein D1 polypeptide (16 kD)
6.6	366647	butyrate response factor 1 (EGF-response factor 1)
6.6	293104	phytanoyl-CoA hydroxylase (Refsum disease)
6.2	28012	O-linked *N*-acetylglucosamine (GlcNAc) transferase
6.1	212198	'tumor protein p53-binding protein, 2'
5.9	247818	ESTs
5.5	26082	very low density lipoprotein receptor
5.4	667598	PC4 and SFRS1 interacting protein 1
5.2	30093	RAN binding protein 1
5.1	73531	nitrogen fixation cluster-like
5	950682	'phosphofructokinase, platelet'
5	47681	'splicing factor, arginine/serine-rich (transformer 2 Drosophila homolog)'
4.9	46019	minichromosome maintenance deficient (*S. cerevisiae*) 7
4.9	307843	ESTs
4.8	548957	'general transcription factor II, i, pseudogene 1'
4.7	788721	KIAA0090 protein
4.7	843076	signal transducing adaptor molecule (SH3 domain and ITAM motif)
4.7	204897	'phospholipase C, gamma 2 (phosphatidylinositol-specific)'
4.7	812227	'solute carrier family 9 (sodium/hydrogen exchanger), isoform 1'
4.6	566887	heterochromatin-like protein 1
4.6	563598	'gamma-aminobutyric acid (GABA) A receptor, pi'
4.5	324210	sigma receptor (SR31747 binding protein 1)

* Percentage of times the genes appeared in the posterior samples.

Breast Cancer: Hedenfalk *et al.* (2001)

Table 2. Crossvalidated classification probabilities and deviances of the 4 models for the breast cancer data set

Y	Model 1 $Pr(Y = 1\|X)$	Model 2 $Pr(Y = 1\|X)$	Model 3 $Pr(Y = 1\|X)$	Model 4 $Pr(Y = 1\|X)$
1	1	1	0.9993	0.9998
1	1	1	1	0.9969
1	1	1	0.9999	1
1	1	1	0.9999	0.8605
1	1	1	0.9999	0.7766
1	1	1	0.9998	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0.0002
0	0	0	0.0018	0.0867
0	0	0	0.0005	0.007
0	0	0	0	0
0	0	0	0	0.2864
1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
Deviance	$1.2683e-12$	$3.1464e-7$	0.0071	1.6843
Number of misclassifications	0	0	0	1

Breast Cancer: Hedenfalk *et al.* (2001)

Table 3. Cross-validation errors of different models for the breast cancer data set

	Model	Cross-validation error*
1	Feed-forward neural networks (3 hidden neurons, 1 hidden layer)	1.5 (Average error)
2	Gaussian kernel	1
3	Epanechnikov kernel	1
4	Moving window kernel	2
5	Probabilistic neural network ($r = 0.01$)	3
6	kNN ($k = 1$)	4
7	SVM linear	4
8	Perceptron	5
9	SVM Nonlinear	6

* Number of misclassified samples.
Feature Selection: 51 Features used in the paper 'Gene-expression profiles in hereditary breast cancer' (Hedenfalk *et al.*, 2001).

Bayesian Probit Regression

Leukemia: Golub *et al.* (1999)

Table 4. Leukemia data: predictions on the test set using genes with frequencies higher than 2.5%.

Y	$Pr(Y\|X_{new})$	Y	$Pr(Y\|X_{new})$
1	1.0000	1	0.2504
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	0.9972	1	0.9999
1	1.0000	1	1.0000
1	1.0000	1	1.0000
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
1	1.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
1	0.9962		
0	0.0000		
0	0.0000		
0	0.0000		
1	1.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0000		
0	0.0012		

Multiclass Classification

In the auxillary variable approach, all the regression tools (MARS, NNs, etc) fit easily in the classification paradigm. Multiclass classification is just an extension of the Albert & Chib (1993) approach.

Define $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iC})$ such that $y_{ij} = 1$ if the ith data point falls in class j. Assume a set of coefficients, $\beta_1, \ldots, \beta_C$, one for each class and

$$p(\mathbf{y}_i|\beta) = \prod_{j=1}^{C} \pi(\beta_i)^{y_{ij}}$$

Multiclass Classification

Also define auxillary variables, $z_{ij} = \mathbf{x}_i'\beta_j + \varepsilon_{ij}$ for each y_{ij} with $\varepsilon_{ij} \sim N(0, \sigma^2)$. Then define the response as

$$p(y_{ij} = 1|\mathbf{z}, \beta) = \begin{cases} 1, & \text{if } z_{ij} > z_{il}, l \neq j \\ 0, & \text{otherwise.} \end{cases}$$

Conditional on the current model, $z_{ij} \sim N(\mathbf{x}_i'\beta_j, \sigma^2)$ subject to $z_{ij} > z_{il}$ for all $l \neq j$, if the ith data point is from the jth category. Y_{new} is predicted to be in class j if

$$P(Y_{new,j} = 1|\mathbf{X}) > P(Y_{new,l} = 1|\mathbf{X})$$

for all $l \neq j$; based on the predictive distribution of Y_{new}, integrating out the parameters.

Multiclass Classification

Example:Finney Data (Alber & Chib, 1993)

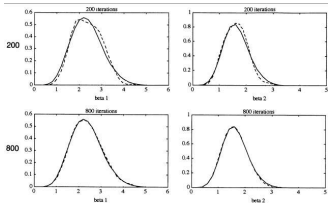
The probit model in Finney (1947) is

$$\pi_i = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}), i = 1, \ldots, 39$$

where x_{i1} - volume of air inspired, x_{i2} - rate of air inspired & the binary outcome is the occurrence or non-occurrence on a transient vasorestriction on the skin of the digits. A uniform prior is placed on β.

The posterior distn of β_1, β_2 are plotted for simulated samples of size 200 and 800, against the exact posterior distn in solid line.

Multiclass Classification



Nonlinear Classification

Probit model:

$$\mathbf{Pr(Y_i = 1|\beta) = \phi(X'\beta)}$$

Nonlinear Probit model:

$$\mathbf{Pr(Y_i = 1|\beta) = \phi\{f(X)\}}$$

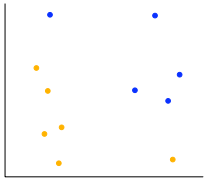
How to model f as X is very high dimension.

- Kernel Methods
- Spline based methods
- Both closely related

Support Vector Machines

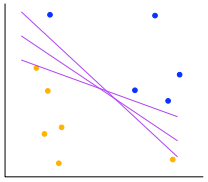
- Excellent performance without lot of tweaking (on par with neural networks)
- Based on simple and elegant principles with nice theoretical properties; used a lot in computer science, machine learning literature
- Construction based on two principles
  - Maximum margin hyperplanes
  - Kernelization

Kernel Methods



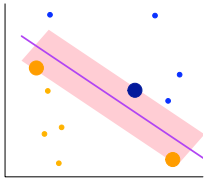
Courtesy: Matt Wand

Kernel Methods



Courtesy: Matt Wand

Kernel Methods



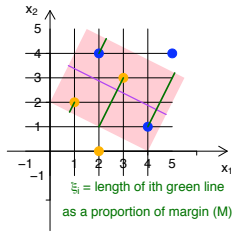
Courtesy: Matt Wand

Support Vector Machines

- Minimize distance of points from this margin subject to penalty constraints

$$\sum_{i=1}^{N} \xi_i \leqslant C$$

Kernel Methods



Courtesy: Matt Wand

Support Vector Machines

- Minimize distance of points from this margin subject to penalty constraints

$$\sum_{i=1}^{N} \xi_i \leqslant C$$

- C is some version of smoothing parameter
- If the points cant be separated by a straight line: transform axis
- Kernelization: the transformation can be written generally as a Kernel matrix: K
- Works very well in high dimensional data problems: microarrays

Kernel Methods

$K_{ij} = K(x_i, x_j|\theta)$: Kernel Matrix.

- Gaussian Kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \text{Exp}\{-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/\theta\}$ (Corrsponding to Radial basis function)
- Polynomial Kernel:$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^\theta$ (Corresponding to Polynomial Basis function)

Kernel Method: Fundamental Theorem (Mallick et al., JRSSB,2005)

Theorem: If K is a reproducing kernel for the function space (Hilbert Space) then the family of functions $K(\cdot.t), t \in \mathbf{x}$ span the space.
So with a choice of a kernel function K, f can be presented as

$$f(\mathbf{x}) = \sum_{k=1}^{n} \beta_k K(\mathbf{x}, \mathbf{x}_k|\boldsymbol{\theta})$$

This is now a n dimensional problem rather than p.

Support Vector Machine

$$p(y_i|p_i) \sim Binary(p_i);$$

$$p_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \stackrel{\text{ind}}{=} \Phi[\mathbf{K}_i'\boldsymbol{\beta}]; \quad (1)$$

$$\boldsymbol{\beta}, \Sigma \sim \text{N}_{n+1}(\boldsymbol{\beta}|\mathbf{0}, \Sigma)\text{IG}(\Sigma|\gamma_1, \gamma_2), \quad (2)$$

$$\boldsymbol{\theta} \sim \Pi_{q=1}^{p} U(a_{q1}, a_{q2}) \quad (3)$$

This is also known as Relevance Vector Machine (RVM).

Nonlinear Probit Model

- Also a Kernel based method
- Difference is the likelihood function
- Based on optimizing the loss function L
- Convert Loss to Likelihood

$$\textbf{Likelihood} \propto \exp[-\mathbf{L}]$$

Likelihood

- We code the class as $Y_i = 1$ or $Y_i = -1$. Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002) and Herbrich (2002). The idea behind support vector machines is to find a linear hyperplane that separates the observations with $y = 1$ from those with $y = -1$ that has the largest minimal distance from any of the training examples. This largest minimal distance is known as the *margin*.
- Shown by Wahba (1999) or Pontil *et al.* (2000), the optimization problem of SVM amounts to finding β which minimizes $\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\{1 - y_i f(\boldsymbol{x}_i)\}_+$, where $[a]_+ = a$ if $a > 0$ and is 0 otherwise, $C \geq 0$ is a penalty term.

Bayesian Hierarchical SVM

- In a Bayesian formulation, the optimization problem is equivalent to finding the posterior mode of $\boldsymbol{\beta}$, where the likelihood is given by $\exp[-\sum_{i=1}^{n}\{1 - y_i f(\boldsymbol{x}_i)\}_+]$, while $\boldsymbol{\beta}$ has the $N(\mathbf{0}, C_{n+1})$ prior.

$$\begin{aligned} p(\boldsymbol{y}|f) &\sim \exp[-\sum_{i=1}^{n}\{1 - y_i f(\boldsymbol{x}_i)\}_+]; \\ f_i|\boldsymbol{\beta}, \boldsymbol{\theta} &= \mathbf{K}_i'\boldsymbol{\beta}; \\ \boldsymbol{\beta}, \Sigma &\sim \mathrm{N}_{n+1}(\boldsymbol{\beta}|\mathbf{0}, \Sigma)\mathrm{IG}(\Sigma|\gamma_1, \gamma_2), \\ \boldsymbol{\theta} &\sim \Pi_{q=1}^{p} U(a_{q1}, a_{q2}) \end{aligned} \tag{4}$$

Bayesian Normalized SVM

- The SVM likelihood does not contain the normalizing constant which may contain f.
- If you do complete normalization then the density comes out to be

$$p(y_i|f_i) = \begin{cases} \{1 + \exp(-2y_i f_i)\}^{-1} & \text{for } |f_i| \leq 1, \\ [1 + \exp\{-y_i(f_i + \mathrm{sgn}(f_i))\}]^{-1} & \text{otherwise,} \end{cases}$$

where $\mathrm{sgn}(u) = 1, 0$ or -1 according as u is greater than, equal or less than 0.

Using this distribution to develop the likelihood we obtain Bayesian Normalized SVM (BNSVM).

Bayesian SVM

We can extend this model using multiple smoothing parameters so that the prior for (β, σ^2) is

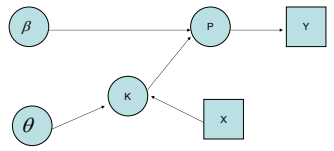
$$\beta, \Sigma \sim \mathrm{N}_{n+1}(\beta|0, \Sigma\mathbf{D}^{-1})\mathrm{IG}(\Sigma|\gamma_1, \gamma_2),$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_{n+1}$. Once again λ_1 is fixed at a small value, but all other λ's are unknown. We assign independent Gamma(m, c) priors to them. Let $\lambda = (\lambda_1, \ldots, \lambda_{n+1})'$.

Bayesian SVM

To avoid the problem of specifying the hyperparameters m and c of λ, we can use Jeffreys' independence prior $p(\lambda) \propto \Pi_{i=1}^{n+1}\lambda_i^{-1}$. This is a limiting form of the gamma prior when both m and c go to 0. Figueirdo (2002) observed that this type of prior promoted sparseness, thus reducing the effective number of parameters in the posterior. Sparse models are preferable as they predict accurately using fewer parameters.

Hierarchical Model



Latent Variable Scheme

The hierarchical model will be

$$p(y_i|z_i) \propto \exp\{-l(y_i, z_i)\},\ i = 1, \ldots, n,$$

where the $y_1, y_2, \cdots, y_n$ are conditionally independent given $z_1, z_2, \cdots, z_n$ and l is any specific choice of the loss function as explained in the previous section.
We relate z_i to $f(\boldsymbol{x}_i)$ by $z_i = f(\boldsymbol{x}_i) + \epsilon_i$, where the ϵ_i are residual random effects.
The random latent variable z_i is thus modeled as

$$z_i = \beta_0 + \sum_{j=1}^{n} \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{\theta}) + \epsilon_i = \mathbf{K}_i'\boldsymbol{\beta} + \epsilon_i, \qquad (1)$$

where the ϵ_i are independent and identically distributed $N(0, \Sigma)$ variables

Bayesian Analysis

Introduction of the latent variables z_i simplify computation (Holmes and Held, 2003), as we now show.
From the Bayes Theorem,

$$p(\beta, \theta, \mathbf{z}, \Sigma, \lambda | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{z}, \beta, \theta, \Sigma, \lambda) p(\beta, \mathbf{z}, \theta, \lambda, \Sigma).$$

This distribution is complex, and implementation of the Bayesian procedure requires MCMC sampling techniques, and in particular, Gibbs sampling (Gelfand and Smith, 1990) and Metropolis−Hastings algorithms (Metropolis *et al.*, 1953). The Gibbs sampler generates posterior samples using conditional densities of the model parameters which we describe

Bayesian Analysis

β and Σ, whose posterior density conditional on $\mathbf{z}, \theta, \lambda$ is Normal-Inverse-Gamma,

$$p(\beta, \Sigma | \mathbf{z}, \theta, \lambda) = \mathrm{N}_{n+1}(\beta | \tilde{\mathbf{m}}, \Sigma \tilde{\mathbf{V}}) \mathrm{IG}(\Sigma | \tilde{\gamma}_1, \tilde{\gamma}_2),$$

where $\tilde{\mathbf{m}} = (\mathbf{K_0}' \mathbf{K_0} + \mathbf{D})^{-1} (\mathbf{K_0}' \mathbf{z})$, $\tilde{\mathbf{V}} = (\mathbf{K_0}' \mathbf{K_0} + \mathbf{D})^{-1}$, $\tilde{\gamma}_1 = \gamma_1 + n/2$, and $\tilde{\gamma}_2 = \gamma_2 + \frac{1}{2}(\mathbf{z}'\mathbf{z} - \tilde{\mathbf{m}}' \tilde{\mathbf{V}} \tilde{\mathbf{m}})$.

The conditional distribution for the precision parameter λ_i given the coefficient β_i is Gamma and is given by

$$p(\lambda_i | \beta_i) = \mathrm{Gamma}\left(m + \frac{1}{2}, c + \frac{1}{2\sigma^2}\beta_i^2\right), \; i = 2, \ldots, n+1.$$

Finally, the full conditional density for z_i is

$$p(z_i | z_{-i}, \beta, \sigma^2, \theta, \lambda) \propto \exp\left[-l(y_i, z_i) - \frac{1}{2\sigma^2}(z_i - \sum_{j=1}^{n} \beta_j K(\mathbf{x}_i, \mathbf{x}_j))^2\right].$$

MCMC Sampling

We make use of a Gibbs sampler that iterates through the following steps:

(i) update $\mathbf{z}$;

(ii) update $\mathbf{K}, \beta, \Sigma$;

(iii) update λ.

We update $z_i | \mathbf{z}_{-i}, \mathbf{y}, \mathbf{K}, \Sigma, \beta \; (i = 1, \ldots, n)$, where $\mathbf{z}_{-i}$ indicates the $\mathbf{z}$ vector with the ith element removed.

Leukemia Data

- Bone marrow or peripheral blood samples are taken from 72 patients with either myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL).
- Training data contains 38 samples, of which 27 are ALL and 11 are AML; Test Data consists of 34 samples, 20 ALL and 14 AML. Gene expression for 7000 genes.

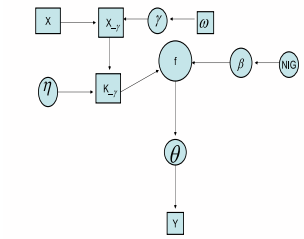
Leukemia Data

Model	modal misclassification error	error bound
RVM	2	(1,4)
BSVM	1	(0,3)
BNSVM	2	(1,6)
Probit	7	
SVM*	3	
RVM	3	

Gene Selection: Ghosh et al (2005, JASA)

- Gene selection is needed to improve the performance of the classifier.
- Introduce $\boldsymbol{\gamma}$, a $p \times 1$ vector of indicator.
- Where $\gamma_i = \begin{cases} 0 & \text{the gene is not selected} \\ 1 & \text{the gene is selected} \end{cases}$
- Prior: $\gamma_i \overset{iid}{\sim}$ Bernoulli(ω).
- Value of ω is chosen to be small to restrict the number of genes in the model.
- $\boldsymbol{K}_\gamma$ is the kernel matrix computed using only those genes whose corresponding elements of $\boldsymbol{\gamma}$ is 1 or using the X_γ matrix.

Hierarchical Model



Prediction

The classification rule :

$$\phi(\boldsymbol{x}_{new}) = arg\ \max_j P(Y_{new} = j | \boldsymbol{x}_{new}, Y_{old})$$

$$\begin{aligned} & P \quad (Y_{new} = j | \boldsymbol{x}_{new}, \boldsymbol{Y}_{old}) \\ = & \int_\gamma \int_\Theta P(Y_{new} = j | \boldsymbol{x}_{new}, \boldsymbol{Y}_{old}, \Theta, \gamma) \Pi(\Theta, \gamma | \text{ data}) d\Theta d\gamma \end{aligned}$$

Is the posterior predictive probability that the tumor belongs to the jth class.

Glioma Cancer

- Gliomas are most common primary brain tumors.
- It occurs at a rate of 12.8 per 100,000 people, and the problem is most common in children ages 3 to 12.
- In the United States, approximately 2,200 children younger than age 20 are diagnosed annually with brain tumors.
- 4 different types of Gliomas depending on the location of their origin.
- The classification of malignant gliomas remains controversial and effective therapies have been elusive.

Glioma Cancer

- All primary glioma tissues were acquired from the Brain Tumor Center tissue bank of the University of Texas M.D. Anderson Cancer Center.
- cDNA microarray with 597 genes.
- 4 types of gliomas GM (glioblastoma multiforme), OL (oligodendroglioma), AO (anaplastic oligodendroglioma), AA (anaplastic astrocytoma).
- A set of 25 patients available. No separate test set so performance is checked by leave one out crossvalidation.

Glioma Cancer

Table 1: Crossvalidation Error

Top Genes	NN	SVM	Wahba	RF	Model 1	Model 2	Model 3	Model 4
20	5	1	2	5	1	1	0	1
50	4	5	3	6	1	1		
100	7	5	4	8	3	2		
597	-	14	9	10	5	4		

Model 1: Bayesian Logit model with gene selection under BWSS.
Model 2: Bayesian SVM with gene selection under BWSS.
Model 3: Bayesian Logit model with Bayesian gene selection.
Model 4: Bayesian SVM with Bayesian gene selection.
On average around 20 genes are selected in the Model 3 and Model 4.

Summary

- RKHS based Bayesian multinomial logit model and Bayesian SVM are strong contenders in predicting the phenotype of a cancer based on its gene expression measurements.
- In both the examples our proposed 2 methods outperforms 3 other methods discussed methods.
- Dimension reduction is built in automatically, no additional projection required.

Comparison of Classifiers

characteristic	CART	MARS	k-NN	Neur. Net.	SVM
Natural handling data of mixed type	●	●	●	●	●
Handling of missing values	●	●	●	●	●
Robustness to outliers in feature space	●	●	●	●	●
Insensitive to monotone transformations of features	●	●	●	●	●
Computational scalability (large training sample size)	●	●	●	●	●
Ability to deal with irrelevant features	●	●	●	●	●
Ability to extract linear combinations of features	●	●	●	●	●
Interpretability	●	●	●	●	●
Predictive power	●	●	●	●	●

Green = Good; Yellow = Fair; Red: Poor

Courtesy: Matt Wand and Hastie, Tibshirani and Friedman (2001)

Spline Based Approaches

MARS models for microarrays

Splines and basis functions

- Given data (X_i, Y_i), $i = 1, \ldots, n$ we wish to estimate

$$Y = f(X) + \epsilon$$

- Splines are one-way to model f flexibly by writing $f(X) = B(X)\beta$ where $B(.)$ are called basis functions.
- Basis functions: there a lot choices available like truncated power basis, B-splines, thin plate splines etc; rich literature.
- Capture non-linear relationships between variables.

Splines and basis functions

- Truncated power basis of order p

$$f(X) = \beta_0 + \beta_1 X + \ldots + \beta_p X^p + \sum_{k=1}^{K} \beta_{k+p}(X - \kappa_k)_+^p$$

  - β's are the regression coefficients
  - κ are the knots
  - K is the number of knots.
- If $p = 1$, then basically join linear pieces at the knots
- Linear regression is just a special case
- Construction of splines involves specifying knots: both number and location.
- Conditional on K, this is just a linear model. Various methods to estimate β. Easiest: least squares (not optimal)

Scientific Questions

- Predict tumor type from gene expression profile
  - Treat gene expression measurements as predictors, tissue type as response
- Gene selection
  - Select most influential genes for the biological question under investigation
- More importantly gene-gene interactions
  - How different genes interact with each other; scale?
  - Provides valuable insights into gene-gene associations and their effect on cancer ontology.
- One unified model!

Statistical Goals

- Develop full probabilistic model-based approach to nonlinear classification
- Smooth classification/decision boundaries; might suggest some biology
- Use Bayesian model mixing for prediction or classification rather than a single model
- Advantage:
  - Model averaging: accuracy
  - By-product: Uncertainty (credible) intervals

Probabilistic Model based Classifiers

- We consider rule based classifiers that use primitives such as

  IF A THEN B

  - A relates to the conditions on the value of a set of predictors(genes) $\boldsymbol{X}$
  - B relates to change in $\mathbf{Pr(Y|X)}$ (log-odds ratio)
  - Provides explicit representation of classification scheme
  - Interpretable models unlike black box techniques (e.g. neural networks)
  - Alternatives: CART (Breimen et al., 1984); graphical order of rules
- Combine scientific interpretation with accurate prediction

Model

Assuming $\boldsymbol{Y_i}$, $\boldsymbol{i = 1, \ldots, n}$ are independent Bernoulli with,

$$\mathbf{Pr(Y_i = 1|X_i) = \mathcal{H}(\eta_i)}$$

- $\boldsymbol{\mathcal{H}(a) = 1/[1 + \exp(-a)]}$ (logistic link function)
- $\boldsymbol{X_i}$ = $\boldsymbol{i}$th row of gene expression matrix $\mathbf{X}$
- Linear model (naive)

$$\mathbf{Pr(Y_i = 1|X_i) = \mathcal{H}(X_i'\beta)}$$

- Non-linear model

$$\mathbf{Pr(Y_i = 1|X_i) = \mathcal{H}(f(X_i))}$$

- Key: Model $\boldsymbol{f}$ as X is high dimensional

Choices for f

- Kernel methods: $K_{ij} = K(x_i, x_j|\theta)$: Kernel matrix
  - Gaussian/Polynomial Kernels; RKHS; SVM's
  - See Mallick, Ghosh and Ghosh (2005, JRSSB)
- We will use basis function approach as,

$$f(X_i) = \sum_{j=1}^{k} \beta_j B(X_i, \theta_j)$$

k = number of basis; β = regression coefficients; θ=basis parameters

- Choices: wavelets, regression splines, artificial neural networks, radial bases, MARS
- Note: both Kernal and Basis function approaches are closely connected

MARS Background

- MARS: Multivariate Adaptive Regression Splines (Friedman, 1991)
  - Flexible regression modeling of high dimensional data
  - Particularly suited to non-linear data sets
- Originally designed for continuous responses
- Extended to deal with classification(categorical) problems (Kooperberg et al,. 1997)
- Extended in the Bayesian framework (BMARS, Denison et al., 1998)
- We extend it to deal with categorical data within a logistic regression framework

Bayesian MARS Model for Gene Interaction

MARS basis function,

$$f(X_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \prod_{l=1}^{z_j} (X_{id_{jl}} - \theta_{jl})_{q_{jl}},$$

- β's are spline coefficients
- z_j is the interaction level: 1 = main effect, 2 = bivariate interaction
- d_{jl} indices of which of the p genes enter the interaction
- k is the number of spline bases
- $q_{jl} \in \{+, -\}$ is the orientation of the spline
- θ_{jl} are knot locations
- All random!

Illustration

Simplified model with $k = 2$ bases and interaction order $z = \{1, 2\}$,

$$\hat{f} = 2.5 + 3.2(x_{20} - 2.5)_+ + 4.1(x_{10} - 1.2)_-(x_{30} + 3.4)_+$$

- Genes either enter the model as main effect or bivariate interaction
- Gene 20 enters the model as a linear term (main effect)
- Genes 10 and 30: bivariate interaction
- Easy to generalize to higher order interactions
- Incorporation of prior biological knowledge

Model

Assuming Y_i, $i = 1, \ldots, n$ are independent Bernoulli with,

$$\Pr(Y_i = 1|X_i) = \mathcal{H}(\eta_i)$$

- $\mathcal{H}(a) = 1/[1 + \exp(-a)]$ (logistic link function)
- X_i = ith row of gene expression matrix $\mathbf{X}$

Model

Assuming Y_i, $i = 1, \ldots, n$ are independent Bernoulli with,

$$\Pr(Y_i = 1|X_i) = \mathcal{H}(\eta_i)$$

- $\mathcal{H}(a) = 1/[1 + \exp(-a)]$ (logistic link function)
- X_i = ith row of gene expression matrix $\mathbf{X}$
- η_i (latent variables) is modeled as (Holmes and Mallick, 2003; JASA),

$$\eta_i = f(X_i) + \epsilon_i$$

- We model the unknown function f nonparametrically using basis functions as,

$$f(X_i) = \sum_{j=1}^{k} \beta_j B(X_i, \theta_j)$$

Model

MODEL

$$\Pr(Y_i = 1|X_i) = \mathcal{H}(\eta_i),$$
$$\eta_i = \sum_{j=1}^{k} \beta_j B(X_i, \theta_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

- η_i: Latent variables used to obtain conditional independence
  - Conditional on η_i's all parameters are independent of Y
  - Holmes and Mallick (2003, JASA)
- Eases computations considerably
  - Efficient sampling and good MCMC mixing
  - Calculations of marginal probabilities

Bayesian Formulation

MODEL: Matrix Notation

$$\Pr(Y = 1|X) = \mathcal{H}(\eta),$$
$$\eta = \Theta(X; \mathcal{M})\beta + \epsilon, \quad \epsilon \sim \text{MVN}(0, \sigma_\epsilon^2 I)$$

where

- $\Theta(X; \mathcal{M})$ is the MARS design matrix
- $\mathcal{M} = \{\theta, q, d, z, k\}$ the spline parameters

Conditionally,

$$\begin{aligned} p(\eta, \mathcal{M}, \beta, \sigma^2|Y) &= p(Y|\eta, \mathcal{M}, \beta, \sigma^2)p(\eta, \mathcal{M}, \beta, \sigma^2) \\ &= p(Y|\eta, \mathcal{M}, \beta, \sigma^2)p(\eta|\mathcal{M}, \beta, \sigma^2)\pi(\mathcal{M})\pi(\beta)\pi(\sigma^2) \end{aligned}$$

Priors

- Prior on regression coefficients

$$\begin{aligned} \beta|\lambda &= \mathbf{Normal}(0, \sigma^2 \mathbf{D}^{-1}); \quad \mathbf{D} = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n+1}) \\ \lambda_i &= \mathbf{Gamma}(\tau_{1i}, \tau_{2i}); \end{aligned}$$

  λ_i's are also smoothing parameters in the spline context

- Prior on spline parameters $\mathcal{M} = \{\theta, q, d, z, k\}$
  - Proper uniform priors on (θ, q, d)
  - $\pi(k) = \mathbf{Uniform}(1, \ldots, \infty)$,
    (improper: no apriori knowledge of number of splines (k))
- Inverse-Gamma prior on σ_ϵ^2

MCMC Computation

- Posteriors are not in explicit form
- Conventional fixed-dimension MCMC algorithms (Gibbs, Metropolis - Hastings) not applicable
- We use Reversible Jump MCMC (Green, 1995) since our model space is variable: we do not know the number of genes (splines) apriori
  - Birth: addition of spline
  - Move: change knot location
  - Death: delete a spline
- MCMC visits numerous models
- Efficient sampling using latent variables

Prediction and Model Choice

- Given x_{new}, marginal posterior distribution of the new disease state y_{new} is,

$$\Pr(y_{new} = 1|x_{new}) = \sum_{k=1}^{\infty} \int P(y_{new} = 1|x_{new}, \mathcal{M}_k) P(\mathcal{M}_k|Y) d\mathcal{M}_k$$

  Approximated by its Monte Carlo estimate,

$$\Pr(y_{new} = 1|x_{new}) = \frac{1}{m} \sum_{j=1}^{m} P(y_{new} = 1|x_{new}, \mathcal{M}^{(j)})$$

  m = number of MCMC samples

- Use misclassification error on to choose among models
- Test and training data

Example: Breast Cancer data

- 22 samples from breast cancer patients carrying mutations of BRCA1 or BRCA2 gene (Hedenfalk,2001); filtered a bit in Simon et al (2003):

  `http://linus.nci.nih.gov/BRB-ArrayTools.html`
- 3226 genes for each sample
- Classify BRCA1 vs. BRCA2 and sporadic
- Consider only main effects and bivariate interactions
- We identify sets of candidate genes which have most bearing on the tumor: MARS automatically ignores genes that have little effect on the response

Breast Cancer Data: Top interacting genes

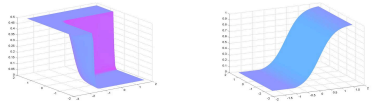
Top interacting genes entering MARS model

Gene 1 description	Gene 2 description	Frequency
glycogenin	ornithine decarboxylase	36.28
glycine cleavage system protein H	dishevelled 2	25.92
ring finger protein 14	ESTs	24.64
D123 gene product	polymyositis/scleroderma	23.92
fragile X mental retardation	ataxia-telangiectasia n	23.12
mitochondrial translational	ESTs	22.08
guanylate binding protein 2	ubiquitin-conjugating enzyme 2	21.64
transducin-like enhancer	hypothetical protein	19.40
⋮	⋮	⋮

Nonlinear Gene Interactions

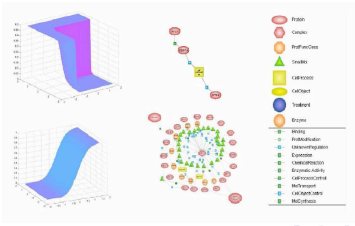
Posterior mean interaction function between two pairs of interacting genes

X and Y axis are the expression levels of interacting genes and vertical axis is the probability of disease



Interactions

Interaction functions of top 2 gene pairs along with the actual biological pathways



Breast Cancer Data: Top main effect genes

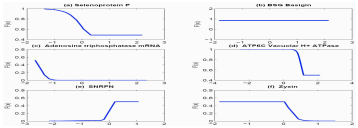
Top main effect genes entering MARS model

Image Clone ID	Gene description	Frequency
767817	polymerase (RNA) II (DNA directed) polypeptide F	71.92
307843	ESTs (*)	57.40
81331	"FATTY ACID-BINDING PROTEIN, EPIDERMAL"	49.20
843076	signal transducing adaptor molecule 1	47.92
825478	zinc finger protein 146	46.08
28012	O-linked N-acetylglucosamine	43.40
566867	heterochromatin-like protein 1 (*)	38.92
841617	ornithine decarboxylase antizyme 1 (*)	37.88
⋮	⋮	⋮

Main Effects

Posterior mean main effect functions of significant genes

X-axis = Gene expression; Y-axis: Probability of disease



Advantage of using a non-linear model : unearth a threshold expression level and its corresponding effect on the odds of having cancer

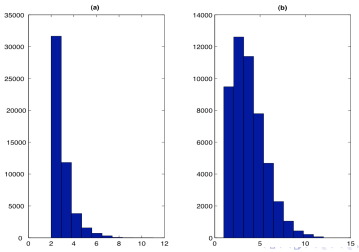
Missclassification Errors

Model Leave-one-out misclassification errors

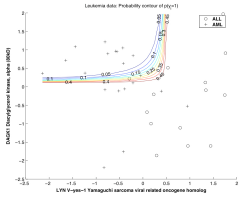
Data	Bayesian MARS	SVM
Leukemia (Golub)	3	4
Colon Cancer (Alon)	4	7
Breast Cancer (Hedenfalk)	0	4

SVM: Classical Support Vector Machine

Number of Genes



Classification Boundaries: Leukemia data



Advantage of using a non-linear model : unearth a threshold expression level and its corresponding effect on the odds of having cancer

Summary

- Nonlinear approach to model gene-gene interactions using Bayesian MARS
- Advantage: capture non-linear dependencies between genes
- Use MCMC based stochastic search algorithms to obtain models
- Identify significant genes of interest
- Potential extensions
  - Multicategory classification
  - Other forms of non-gaussian data
  - Gene regulatory networks