BAYESIAN METHODS FOR VARIABLE SELECTION WITH APPLICATIONS TO HIGH-DIMENSIONAL DATA

Part 2: Variable Selection for Mixture Models

Marina Vannucci

Rice University, USA

PASI-CIMAT
04/28-30/2010
Part 2: Variable Selection for Mixture Models

- Finite mixture models for sample clustering
- Variable selection
- Simulated data
- Application to microarray
Objective

- Simultaneous variable selection and sample clustering
- Cluster structure of samples confined to a small subset of variables. Noisy variables mask the recovery of the clusters.

Proposed methodology:
- Use multivariate normal mixture model with an unknown number of components to determine cluster structure of the samples.
- Use stochastic search techniques to examine the space of variable subsets and identify most probable models.
- Also, infinite mixture models via Dirichlet process priors.

Genomic data: Identify disease subtypes and select the discriminating genes.
Finite Mixture Models

- Discriminating variables define a mixture of $G$ distributions

$$f(x_i | w, \theta) = \sum_{k=1}^{G} w_k f(x_i | \theta_k).$$

- We consider $f(x_i | \theta_k)$ multivariate normal with $\theta_k = (\mu_k, \Sigma_k)$.
- Cluster assignments: $y = (y_1, \ldots, y_n)'$, where $y_i = k$ if the $i^{th}$ observation comes from cluster $k$

$$p(y_i = k) = w_k.$$

Binder (1978); McLachlan and Basford (1988).
Variable Selection

- Need to select discriminating variables.
- Introduce latent $p$-vector $\gamma$ with binary entries
  \[
  \begin{cases}
  \gamma_j = 1 & \text{if variable } j \text{ defines a mixture distribution} \\
  \gamma_j = 0 & \text{otherwise.}
  \end{cases}
  \]
- The likelihood function is given by
  \[
  L(G, \gamma, w, \mu, \Sigma, \eta, \Omega|X, y) = \prod_{k=1}^{G} \left(2\pi\right)^{-\frac{pn_k}{2}} |\Sigma_k|^{-\frac{n_k}{2}} \ w_k^{n_k} \times \exp \left\{ -\frac{1}{2} \sum_{x_i \in C_k} (x_{(\gamma)i} - \mu_{(\gamma)k})^T \Sigma_{(\gamma)k}^{-1} (x_{(\gamma)i} - \mu_{(\gamma)k}) \right\} \times \phi(X_{(\gamma)c}|\eta_{(\gamma)c}, \Omega_{(\gamma)c}),
  \]
  where $C_k = \{x_i|y_i = k\}$ with cardinality $n_k$, $\phi(.)$ is multivariate normal density.
Prior Model

- Assume $\gamma_j$'s are independent Bernoulli variables.
- Number of components, $G$, can be assumed to follow a truncated Poisson or a discrete Uniform on $[2, \ldots, G_{\text{max}}]$.
- $w|G \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.
- $\{\mu_k(\gamma)|\Sigma_k(\gamma), G \sim \mathcal{N}(\mu_0(\gamma), h\Sigma_k(\gamma))\}$
- $\Sigma_k(\gamma)|G \sim \mathcal{IW}(\delta; Q_\gamma)$,

where $(\gamma)$ indicates the covariates with $\gamma_j = 1$.

We work with a marginalized likelihood.
Model Fitting

(1) Update $\gamma$ by Metropolis algorithm (add/delete and swap moves).
(2) Update $w$ from its full conditional (Dirichlet draw).
(3) Update $y$ from its full conditional (multinomial draw).
(4) Split one cluster into two, or merge two into one.
(5) Birth or death of an empty component.

Steps (4) and (5) via reversible jump MCMC (Green, 1995).
Posterior Inference for $y$

- Number of clusters, $G$, estimated by value most frequently visited by MCMC sampler.
- Estimate marginal posterior probabilities $p(y_i = k | X, G)$. Posterior allocation of sample $i$ estimated as

$$\hat{y}_i = \max_{1 \leq k \leq G} \{p(y_i = k | X, G)\}.$$
Posterior Inference for $\gamma$

- Select variables with largest marginal posterior probability

$$p(\gamma_j = 1 | X, G)$$

- Select variables that are in the “best” models

$$\hat{\gamma}^* = \arg\max_{1 \leq t \leq M} \left\{ p(\gamma^{(t)} | X, G, \hat{w}, \hat{y}) \right\},$$

with $\hat{y}$ the estimated sample allocations and $\hat{w} = \frac{1}{M} \sum_{t=1}^{M} w^{(t)}$.

Tadesse, Sha and Vannucci (JASA, 2005)
Infinite Mixture Models via Dirichlet Process Priors

- Integrating over $w$ and taking $G \to \infty$ we get
  
  \[
  p(y_i = k \text{ and } y_l = k \text{ for some } l \neq i | y_{-i}) = \frac{n_{-i,k}}{n - 1 + \alpha}
  \]
  
  \[
  p(y_i \neq y_l \text{ for all } l \neq i | y_{-i}) = \frac{\alpha}{n - 1 + \alpha}.
  \] (1)

- MCMC updates $\gamma$ via Metropolis and $y_i$ from full conditionals
  
  \[
  p(y_i = k \text{ and } y_l = k \text{ for some } l \neq i | y_{-i}, X, \gamma)
  \]
  
  \[
  p(y_i \neq y_l \text{ for all } l \neq i | y_{-i}, X, \gamma).
  \] (2)

- Inference on $y$ by MAP or by estimating $p(y_i = y_j | X)$. Same as before for $\gamma$

- Natural approach to clustering (samples from a DP can have a number of ties).

Kim, Tadesse and Vannucci (Biometrika, 2006)
Application to Simulated Data

- 15 samples, 4 multivariate normal densities, 20 variables
  \[ x_{ij} \sim \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) & \text{if } 1 \leq i \leq 4 \\ \mathcal{N}(\mu_2, \sigma_2^2) & \text{if } 5 \leq i \leq 7 \\ \mathcal{N}(\mu_3, \sigma_3^2) & \text{if } 8 \leq i \leq 13 \\ \mathcal{N}(\mu_4, \sigma_4^2) & \text{if } 14 \leq i \leq 15 \end{cases} \]

- Cluster sizes: 4-3-6-2

- Additional set of 980 noisy variables drawn from a standard normal density
Weakly informative priors for model parameters.
\((\delta = 3, \alpha = 1, h = 100, Q = kI)\)

Truncated Poisson prior for \(G\) with \(G_{\text{max}} = 10\).

MCMC with 100,000 iterations - starting model with 1 randomly selected \(\gamma_j\) set to 1.
Trace plot of number of clusters, $G$
Trace plot for number of included variables, $p_\gamma$
Marginal posterior probabilities, $p(\gamma_j = 1 | X, G = 4)$
Marginal posterior probabilities of sample allocations,
\[ p(y_i = k|X, G = 4), \ i = 1, \ldots, 15, \ k = 1, \ldots, 4 \]
Results

- $G = 4$ had stronger support
- All sample allocations corresponded to the true cluster structure
- There were 16 variables with marginal probability $> 0.7$ (15 were correct)
- Very little sensitivity to model parameters, with the exception of the covariance hyperparameters
Simultaneous Class Discovery and Gene Selection

- Endometrial cancer: Most common gynecologic malignancy in the US.
- 10 tumor and 4 normal tissues collected from hysterectomy specimens, examined with Affymetrix Hu6800 arrays.
- Probe sets with unreliable readings ($< 20$ and $> 16,000$) removed $\Rightarrow p = 762$.
- Gene expressions were log-transformed and scaled by their range.
- Specified weakly informative priors for model parameters.
- Used truncated Poisson prior for $G$ with $G_{\text{max}} = n$.
  
  $$p(\gamma_j) \sim \text{Bernoulli}(\varphi = 10/p).$$
- Ran four MCMC chains with widely different starting points:
  (a) 1; (b) 10; (c) 25; (d) 50 randomly selected $\gamma_j$’s set to 1.
- Posterior distribution of $G$
- Union of 4 chains – $p(\gamma_j = 1 | X, G = 3)$
We have identified 3 classes and a set of 31 genes that can distinguish subtypes of the disease.