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

Intro: Course Outline and Brief Intro to Bayesian Methods

Marina Vannucci

Rice University, USA

PASI-CIMAT
04/28-30/2010

Outline of the Course

- Brief Introduction to Bayesian Methods
- Part 1: Mixture Priors for Linear Settings
  - Linear regression models (univariate and multivariate responses)
  - Matlab code on simulated data
  - Other linear settings (categorical responses and survival outcomes)
  - Applications to high-throughput data from Bioinformatics
  - Models that incorporate biological information
- Part 2: Variable Selection for Mixture Models
  - Finite mixture models for sample clustering
  - Simulated data and applications to microarrays
- Part 3: Functional Data & Wavelets
  - A brief introduction to wavelets
  - Curve regression and classification
  - Applications to NIR spectral data from Chemometrics

Brief Introduction to Bayesian Methods

- Intuition: Combine inference from data with prior information
- Model: $y|\theta \sim f(y|\theta)$
- θ unknown (parameter, missing data, latent variable, ...)
- Bayesian point of view:

  - θ has a probability distribution reflecting our uncertainty about it
  - y is known, so we should condition on it

- Then $\theta \sim \pi(\theta)$ and inference is done using Bayes theorem

$$
\begin{aligned}
p(\theta|y) &= \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \\
p(\theta|y) &\propto f(y|\theta)\pi(\theta) \\
\text{posterior} &\propto \text{likelihood x prior}
\end{aligned}
$$

Bayes Theorem

$$p(\theta|y) \propto f(y|\theta)\pi(\theta)$$

- $p(\theta)$ is our uncertainty about θ before seeing the data
- $p(\theta|y)$ is our uncertainty about θ after seeing the data
- the integral in Bayes theorem is a normalizing constant that makes $p(\theta|y)$ integrate (sum) to 1
- The posterior distribution of θ can be summarized through:
  - point estimates (mean, median, mode)
  - interval estimates (HPD regions or lower/upper $\alpha/2$ percentiles)
  - hypothesis testing (often Bayes factors $\frac{p(y|M_1)}{p(y|M_2)} = \frac{p(M_1|y)/p(M_2|y)}{p(M_1)/P(M_2)}$)
- Prediction of a future observation z (independent of y given θ)

$$p(z|y) = \int p(z,\theta|y)d\theta = \int p(z|\theta)p(\theta|y)d\theta$$

Bayes and Decision Theory

- If we are willing to quantify the value of different consequences it is possible to use posterior probabilities as a basis for decision theory.
- Given a set of actions (rules) $a \in A$ and a loss function $L(\theta, a)$ we minimize the posterior expected loss

$$min_a E_{\theta|y} L(\theta, a) = min_a \int L(\theta, a) p(\theta|y) d\theta \rightarrow \hat{\theta}$$

- L_2-loss, $L(\theta, a) = (\theta - a)^2, a \in R \rightarrow \hat{\theta} = E(\theta|y)$
- L_1-loss, $L(\theta, a) = |\theta - a|, a \in R \rightarrow \hat{\theta} = median(\theta|y)$

- 0-1 loss (hyp. testing) $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, $a = \{0, 1\} = \{\text{accept } H_0, \text{ reject } H_0\}$,

$$L(\theta, a) = \begin{cases} a, & \text{if } \theta \in \Theta_0 \\ 1 - a, & \text{if } \theta \in \Theta_1 \end{cases}$$

$$E_{\theta|y} L(\theta, a) = \begin{cases} P(\theta \in \Theta_1 | y), & \text{if } a = 0 \\ P(\theta \in \Theta_0 | y), & \text{if } a = 1 \end{cases}$$

Bayes rule: accept $H_0 (a = 0)$ if $P(\theta \in \Theta_0 | y) > P(\theta \in \Theta_1 | y)$

Bayes in Practice

- How do I quantify my prior information?

  - conjugate choices
  - diffuse specifications

- How do I assess the effect of my prior beliefs?

  - sensitivity analyses across alternative specifications can reveal stability (or not) to prior models.

- How do I do integrals? i.e., $p(\theta|y)$ with non-standard or non-conjugate choices, prediction, marginal inference on single parameters as $p(\theta_i|y) = \int p(\theta|y)d\theta_{-i}$

  - Markov chain Monte Carlo methods offer a solution.

Markov Chain Monte Carlo Methods

MCMC methods achieve inference via Monte Carlo integration using simulated values generated from a Markov chain with $p(\theta|y)$ as stationary distribution.

- Gibbs sampling: given θ_0, at iteration t, sample all parameters from

$$\theta_i^{(t)} \sim p(\theta_i|\theta_{-i}^{(t-1)}, y) \text{ (full conditionals)}$$

Can prove that $\theta^{(t)} \to \theta \sim p(\theta|y)$ in distribution as $t \to \infty$
Samples are summarized for posterior inference via Monte Carlo integration (after convergence at t_0)

$$E(f(\theta)) = \frac{1}{m} \sum_{t_0+1}^{t_0+m} f(\theta^{(t)})$$

(posterior mean, posterior quantiles, kernel estimates)
Note: use prior distributions which are *conditionally* conjugate.

• Metropolis-Hastings: when full conditionals are not available in closed form, v is sampled from a proposal distribution $q(\cdot, \theta^{(t-1)})$ and accepted ($\theta^{(t)} = v$) with probability

$$min\left[1, \frac{p(\theta^{(t-1)}|y)q(v, \theta^{(t-1)})}{p(v|y)q(\theta^{(t-1)}, v)}\right]$$

Can prove that $\theta^{(t)} \rightarrow \theta \sim p(\theta|y)$ in distribution as $t \rightarrow \infty$

• Metropolis algorithm: M-H with a symmetric proposal,

$$q(u, v) = q(v, u)$$