BIOST 572 Presentation 2

JooYoon Han

Department of Biostatistics
University of Washington

May 1, 2014

Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard

Xiao-Hua Zhou,[1,2,*] Pete Castelluccio,[3,] and Chuan Zhou[2,***]**

[1]HSR&D VA Puget Sound Health Care System, Seattle, Washington 98101, U.S.A.
[2]Department of Biostatistics, University of Washington, Box 357232, Seattle, Washington 98195, U.S.A.
[3]Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, Indiana 46202, U.S.A.
[*]*email:* azhou@u.washington.edu
[**]*email:* pfcastel@iupui.edu
[***]*email:* czhou@u.washington.edu

Biometrics 61, 600-609
June 2005

Setup

- N patients, K diagnostic tests with scale from 1 to J (ordinal)
- Disease status is unknown for all N patients
- T_1,...,T_K: responses from K tests for a particular patient
- $y_{ikj} = \begin{cases} 1 & \text{if } x = \textit{response of kth test is j for the ith patient} \\ 0 & \text{if } \textit{otherwise} \end{cases}$
- $\mathbf{y_i}$=$(y_{i11}, .., y_{i1J}, .., y_{iK1}, .., y_{iKJ})$: $K \times J$ test score vector for i^{th} patient

Setup

$$
\begin{aligned}
g_d(\mathbf{y_i}) &= P(\mathbf{y_i}|D_i = d) \\
&= \prod_{k=1}^{K}\prod_{j=1}^{J} P(T_k = j|D_i = d)^{y_{ikj}} (\textit{conditional indep of the } K \textit{ tests}) \\
&= \prod_{k=1}^{K}\prod_{j=1}^{J} [\phi_{dkj}]^{y_i kj}
\end{aligned}
$$

Setup

Assume D$\sim$Bernoulli(p_d), p_d=P(D=d)

- Likelihood contributed by the ith patient
  - $P(\mathbf{y_i}) = p_1 g_1(\mathbf{y_i}) + p_0 g_0(\mathbf{y_i})$
- Joint log likelihood
  - $l(p_1, \phi_0, \phi_1) = \sum_{i=1}^{N} log[p_0 g_0(\mathbf{y_i}) + p_1 g_1(\mathbf{y_i})]$
    where $p_0 = 1 - p_1$ and $\phi_d = (\phi_{d11}, ..., \phi_{d1J}, ..., \phi_{dK1}, ..., \phi_{dKJ})$

$\Rightarrow$EM algorithm: MLE for $p_1, \phi_0, and \phi_1$

EM Algorithm-Background

- Expectation Maximization Algorithm
- Iterative method for finding maximum likelihood estimates or maximum posteriori estimates of parameters
- "Broadly applicable algorithm for computing maximum likelihood estimates from incomplete data" Dempster et al.(1977)
- Useful for estimation of mixing proportions in finite mixture densities

EM Algorithm-Background

- X: Observed, Z:Unobserved
- $\theta^{(t)}$: (Current) parameter
- $q(z|x,\theta)$: Some distribution over the Z's

$$\sum_i log p(x|\theta) = \sum_i log \sum_z p(x,z|\theta)$$

$$= \sum_i log \sum_z \frac{p(x,z|\theta)}{q(z|x,\theta)} q(z|x,\theta)$$

$$= \sum_i log E_z[\frac{p(x,z|\theta}{q(z|x,\theta)}]$$

$$\geq \sum_i \sum_z q(z|x,\theta) log \frac{p(x,z|\theta)}{q(z|x,\theta)}$$

$$= \sum_i \sum_z [q(z|x,\theta) log p(x,z|\theta) - q(z|x,\theta) log q(z|x,\theta)]$$

EM Algorithm-Background

- $\sum_i \sum_z q(z|x,\theta) log \frac{p(x,z|\theta)}{q(z|x,\theta)}$: Lower-bound on the log likelihood
- E step: Calculate $E_{z|x,\theta}\left[log\ p(x,z|\theta)\right]$
- M step: Find $\theta^{(t+1)} = argmax_\theta\ E_{z|x,\theta}\left[log\ p(x,z|\theta)\right]$

EM Algorithm-Setup

- Observed data: ($\mathbf{y}$)
- Unobserved data: ($\mathbf{D}$)
- Complete data: ($\mathbf{y}$, $\mathbf{D}$)
- Parameter: $\theta = (p_1, \phi_0, \phi_1)$
- Estimate of θ after the t^{th} iteration: $\theta^{(t)}$

EM algorithm-E step

The E step computes the conditional expectation of $l_c(\theta)$ given the observed data $\mathbf{y}$ and current parameter estimates $\theta{=}\theta^{(t)}$

- log likelihood
  - Observed data: $l(\theta) = \sum_{i=1}^{N} log[p_0 g_0(\mathbf{y_i}) + p_1 g_1(\mathbf{y_i})]$
  - Complete data: $l_c(\theta) = \sum_{i=1}^{N}[(1 - D_i) log p_0 g_0(\mathbf{y_i}) + D_i log p_1 g_1(\mathbf{y_i})]$
- $\mathrm{E}(l_c(\theta)|\mathbf{y}, \theta = \theta^{(t)})$

$$
\begin{aligned}
&= E\left[\sum_{i=1}^{N}\sum_{d=0}^{1} D_i log p_d g_d(\mathbf{y_i})\right] \\
&= \sum_{i=1}^{N}\sum_{d=0}^{1} E[D_i] log p_d g_d(\mathbf{y_i}) \\
&= \sum_{i=1}^{N}\sum_{d=0}^{1} P(D_i = d|\mathbf{y_i}, \theta^{(t)}) log p_d g_d(\mathbf{y_i})
\end{aligned}
$$

EM algorithm-E step

Let $q_{id}^{(t)} = P(D_i = d|\mathbf{y_i}, p_1^{(t)}, \phi_0^{(t)}, \phi_1^{(t)})$

$$= \frac{p_d^{(t)} g_d^{(t)}(\mathbf{y_i})}{p_0^{(t)} g_0^{(t)}(\mathbf{y_i}) + p_1^{(t)} g_1^{(t)}(\mathbf{y_i})}$$

$$g_d^{(t)}(\mathbf{y_i}) = \prod_{k=1}^{K} \prod_{j=1}^{J} [\phi_{dkj}^{(t)}]^{y_{ijk}}$$

Then $\mathrm{E}(l_c(\theta)|\mathbf{y}, \theta = \theta^{(t)})$

$$= \sum_{i=1}^{N} \sum_{d=0}^{1} P(D_i = d|\mathbf{y_i}, \theta^{(t)}) log p_d g_d(\mathbf{y_i})$$

$$= \sum_{i=1}^{N} \sum_{d=0}^{1} q_{id}^{(t)} log p_d g_d(\mathbf{y_i})$$

EM algorithm-M step

The M step finds the updated estimate $\theta^{(t+1)}$ for θ by maximizing $\mathrm{E}(l_c(\theta)|\mathbf{y}, \theta = \theta^{(t)})$ from E step with respect to θ.

$$\frac{\partial E(l_c(\theta))}{\partial p_1^{(t)}} = \frac{1}{p_1^{(t)}} \sum_{i=1}^{N} q_{i1}^{(t)} - N = 0$$
$$\Longrightarrow p_1^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} q_{i1}^{(t)}$$

$$\frac{\partial E(l_c(\theta))}{\partial \phi_{dkj}^{(t)}} = \frac{\sum q_{id}^{(t)} y_{ikj}}{\phi_{dkj}^{(t)}} - \sum q_{id}^{(t)} = 0$$
$$\Longrightarrow \phi_{dkj}^{(t+1)} = \frac{\sum_{i=1}^{N} q_{id}^{(t)} y_{ikj}}{\sum_{i=1}^{N} q_{id}^{(t)}}$$

EM algorithm-M step

With these updated parameters, we can get $g_d^{(t+1)}$ *and* $q_{id}^{(t+1)}$

$\Longrightarrow$ How should we choose the initial parameter estimates?

EM algorithm-Initial Values

Recommendations from this paper

1) Avoid equal $\phi_{0kj} = \phi_{1kj}$ for all k and j

- We obtain $g_0^{(t=0)}(\mathbf{y_i}) = g_1^{(t=0)}(\mathbf{y_i})$
- $q_{i1}^{(t)}$ does not depend on data $\mathbf{y}$
- $p_1^{(t+1)} = p_1^{(t)} = \cdots = p_1^{(0)}$
- Iterative procedure will stop after just one iteration

EM algorithm-Initial Values

Recommendations from this paper

- 2) Try a set of reasonable initial parameter estimaes, and compare the local log-likelihood maxima obtained
- 3) Obtain reasonable initial values from similar studies with known disease status
- 4) Study the likelihood surface using exploratory and simulation techniques, such as the stochastic EM

AUC

With these updated parameters, we can get the Area Under the Curve (AUC)

- Positive test: $T_k \geq j$
- $FPR_k(j) = \sum_{l=j}^{J} P(T_k = j|D = 0) = \sum_{l=j}^{J} \phi_{0kl}$
- $TPR_k(j) = \sum_{l=j}^{J} P(T_k = j|D = 1) = \sum_{l=j}^{J} \phi_{1kl}$
- $A_k = \sum_{j=1}^{J-1} [\phi_{0kj} \sum_{l=j+1}^{J} \phi_{1kl}] + \frac{1}{2} \sum_{j=1}^{J} \phi_{0kj} \phi_{1kj}$

Simulation

- N=118
- K=7
- True prevalence p_1=0.5, 0.7, and 0.9
- Calculate Bias and MSE of estimators (p_1 and AUC)

Simulation-Fisher's Information Matrix

- Calculate MSE
- $E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial p_1^2}], E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial p_1 \partial \phi_{0kj}}], E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial p_1 \partial \phi_{1kj}}]$
- $E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial \phi_{0kj} \partial \phi_{0kj}}], E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial \phi_{0kj} \partial \phi_{1kj}}], E[-\frac{\partial^2 l(p_1,\phi_0,\phi_1)}{\partial \phi_{1kj} \partial \phi_{1kj}}]$

Next time

- Fisher's information matrix
- Simulation Study!