BIOST 572 Presentation 1

JooYoon Han

Department of Biostatistics
University of Washington

April 15, 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

Gold Standard

- What is a Gold standard test?
  - Diagnostic test that is the best available under reasonable conditions
  - The most accurate test possible without restrictions
  - In medicine, the gold standard test is less accurate than the autopsy
- Gold standard ambiguity
  - "Sometimes" the best performing test available
  - "Other times" the best test available under reasonable conditions (Example: MRI)

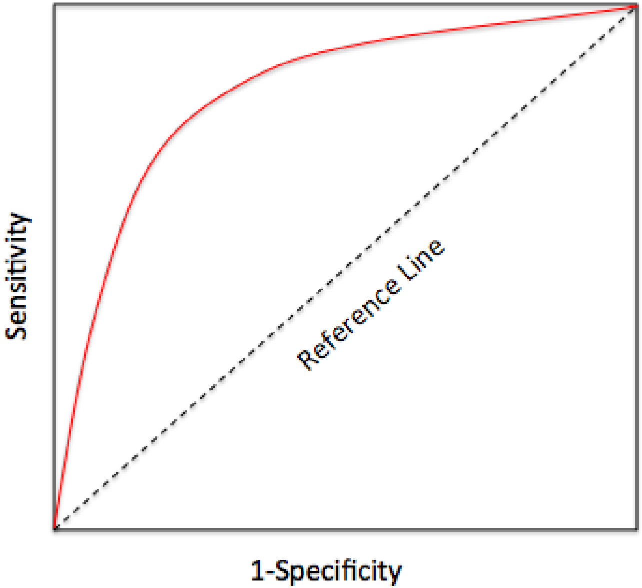
Absence of a Gold Standard

- Difficult to perform
- Expensive
- Impossible to perform on a living person
- This type of bias is called "Imperfect Gold Standard Bias"

Receiver Operating Characteristic (ROC)

- What is a ROC curve?
  - Method of describing the accuracy of a test apart from the decision thresholds
  - Plot of a test's true positive rate (or sensitivity) versus its false postive rate (or 1-specificity)
  - The most valuable tool for describing and comparing the accuracies of diagnostic tests
- Comparing the ROC curves
  - Best when the curve is near left upper end
  - Compare using Area Under the Curve (AUC) which is overall measure of test performance
  - Near 1: Excellence
  - Near 0.5: Fail

Receiver Operating Characteristic (ROC)



ROC curve for ordinal-scale tests

- Nonparametric ROC curve based on the discrete sensitivity and specificity
- Continuous ROC curve of a latent variable underlying the observable ordinal data

Previous Works

- Only a few published papers have dealt with the estimation of ROC curves of ordinal or continuous scale tests in the absence of a gold standard
- Henkelman, Kay, and Bronskill (1990)
  - Maximum likelihood estimation method for ROC curve of a 5-point rating scale using a multivariate normal mixture latent model
  - Limitation: Latent random variables from multiple ordinal-scale tests are assumed to follow MVN
- Hall and Zhou (2003)
  - Nonparametric estimator for the ROC curve of continuous-scale tests under the conditional independece assumption when the number of tests is more than two

Previous Works

- This paper will apply the ideas of Hall and Zhou (2003)
- Focus on a nonparametric maximum likelihood (ML) method under the conditional independence assumption

Setup

- N patients
- K diagnostic tests
- Scored on an ordinal scale from 1 to J
- Disease status is unknown for all N patients
- $T_1,...,T_K$: responses from K tests for a particular patient

Nonparametric ROC Curve

- Vary the threshold for a positive test
- Calculate J+1 pairs of true positive rates (TPR) and false positive rates (FPR)

Nonparametric ROC Curve

Specifically, for kth test

- Define a positive test as one with $T_k \geq j$, j=1,...,J+1
- $TPR_k(j) = P(T_k \geq j | D = 1)$
- $FPR_k(j) = P(T_k \geq j | D = 0)$
- $TPR_k(1) = FPR_k(1) = 1$
- $TPR_k(J+1) = FPR_k(J+1) = 0$

A discrete ROC curve is defined as a discrete function of $(FPR_k(j), TPR_k(j))$, j=1,...,J+1.

We obtain nonparamtric ROC curve by connecting coordinates with linear lines.

Nonparametric ROC Curve

Define

- $\phi_{0kj} = P(T_k = j | D = 0)$ and $\phi_{1kj} = P(T_k = j | D = 1)$
- $FPR_k(j) = \sum_{l=j}^{J} \phi_{0kl}$
- $TPR_k(j) = \sum_{l=j}^{J} \phi_{1kl}$
- ROC curve and AUC: functions of ϕ_{0kj} and ϕ_{1kj} because coordinates of the nonparametric ROC curve of T_k are $(FPR_k(j), TPR_k(j))$

Nonparametric ML method

We wish to find MLEs for these parameters and calculate MLEs for the ROC curve and its area under each of the K tests.

Define,

$$y_{ikj} = \begin{cases} 1 & \text{if } x = \textit{response of kth test is j for the ith patient} \\ 0 & \text{if } \textit{otherwise} \end{cases}$$

where i=1,...,N, k=1,...,K, and j=1,...,J

Test score vector for the ith patient is

$$\mathbf{y_i} = (y_{i11}, .., y_{i1J}, .., y_{iK1}, .., y_{iKJ})$$

Nonparametric ML method

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

Nonparametric ML method

Assume a Bernoulli distribution for D with $p_d = P(D = d)$ for d=0,1

- 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 $\phi_d = (\phi_{d11}, ..., \phi_{d1J}, ..., \phi_{dK1}, ..., \phi_{dKJ})$

Goal: Find the ML estimates for $p_1, \phi_0, and \phi_1 \Rightarrow EM\ algorithm$

EM Algorithm

- Complete data: ($\mathbf{y}$, D)
- θ=(p_1, ϕ_0, ϕ_1)
- $l_c(\theta) = \sum_{i=1}^{N}[D_i log p_1 g_1(\mathbf{y_i}) + (1 - D_i) log p_0 g_0(\mathbf{y_i})]$
- $\theta^{(t)}$: estimate of θ after tth iteration

EM Algorithm

- E step
  - Computes the conditional expectation of $l_c(\theta)$ given the observed data $\mathbf{y}$ and current parameter estimates $\theta = \theta^{(t)}$
- M step
  - Finds the updated estimate $\theta^{(t+1)}$ for θ by maximizing $E(l_c(\theta)|\mathbf{y}, \theta = \theta^{(t)})$

...details (Next time)

Next time

- Details for EM algorithm
- Some math proofs
- Simulation study