University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 38

Binary isotonic regression procedures, with application to cancer biomarkers

Debashis Ghosh* Moulinath Banerjee†

Pinaki Biswas‡

*University of Michigan, ghoshd@psu.edu

†University of Michigan, moulib@umich.edu

‡Univeristy of Michigan, pbiswas@umich.edu

http://biostats.bepress.com/umichbiostat/paper38

Binary isotonic regression procedures, with application to cancer biomarkers

Debashis Ghosh, Moulinath Banerjee, and Pinaki Biswas

Abstract

There is a lot of interest in the development and characterization of new biomarkers for screening large populations for disease. In much of the literature on diagnostic testing, increased levels of a biomarker correlate with increased disease risk. However, parametric forms are typically used to associate these quantities. In this article, we specify a monotonic relationship between biomarker levels with disease risk. This leads to consideration of a nonparametric regression model for a single biomarker. Estimation results using isotonic regression-type estimators and asymptotic results are given. We also discuss confidence set estimation in this setting and propose three procedures for computing confidence intervals. Methods for estimating the receiver operating characteristic (ROC) curve are also described. The finite-sample properties of the proposed methods are assessed using simulation studies and applied to data from a pancreatic cancer biomarker study.

Binary isotonic regression procedures, with application to cancer biomarkers

Debashis Ghosh[1], Moulinath Banerjee[2] and Pinaki Biswas[1]

[1]Department of Biostatistics and [2]Department of Statistics

University of Michigan

Ann Arbor, MI 48109

Summary

There is a lot of interest in the development and characterization of new biomarkers for screening large populations for disease. In much of the literature on diagnostic testing, increased levels of a biomarker correlate with increased disease risk. However, parametric forms are typically used to associate these quantities. In this article, we specify a monotonic relationship between biomarker levels with disease risk. This leads to consideration of a nonparametric regression model for a single biomarker. Estimation results using isotonic regression-type estimators and asymptotic results are given. We also discuss confidence set estimation in this setting and propose three procedures for computing confidence intervals. Methods for estimating the receiver operating characteristic (ROC) curve are also described. The finite-sample properties of the proposed methods are assessed using simulation studies and applied to data from a pancreatic cancer biomarker study.

Keywords: Diagnostic Test, Nonstandard asymptotics, Receiver Operating Characteristic (ROC) curve, Sensitivity.

1. Introduction

There has been extensive work done on the development of methodology for diagnostic testing and screening (Pepe, 2003). The primary scientific goal in this area is to determine the discriminatory power of a biomarker for detecting disease, or equivalently, to model the effect of biomarker on disease risk. As an example, we consider prostate cancer. Typically, prostate-specific antigen (PSA) has been used for detection of prostate cancer. If a man has a PSA measurement between 4 and 10 ng/mL, then this leads to a prostate needle biopsy. While PSA is able to detect prostate cancer when it is present, it also leads to numerous false positives. Thus, there is interest among oncologists in determining the effect of PSA on risk of prostate cancer.

For much of the work on statistical methodology on screening and diagnostic methods, it has been assumed that increasing levels of the biomarker are associated with increased disease risk. Typically, this effect has been specified in a parametric manner. We give two examples of this. The first is linear discriminant analysis, in which the distribution of the biomarker in the diseased and undiseased populations is assumed to be normally distributed. The second is logistic regression, in which the effect of the biomarker on the probability of developing disease is assumed to be linear on the logit scale. For these procedures, if the parametric form is misspecified, then this can lead to bias in the estimate of association between biomarker levels and disease risk.

It is desirable to have flexible models and procedures for studying the association between a biomarker and risk of disease. It seems intuitively reasonable to have the effect be monotonic but to leave it otherwise unspecified. There has been much research on monotonic regression in the statistical literature. Two-stage estimation procedures have been proposed by Mukerjee (1988) and Mammen (1991) in which nonparametric regression is followed by monotone empirical smoothing using a pool-adjacent-violators algorithm. Ramsay (1988), Kelly and Rice (1990) and He and Shi (1998) have proposed monotone B-spline estimation procedures that are guaranteed to provide monotonic estimates under nonnegativity constraints on the parameters. All of these procedures

involve some form of nonparametric smoothing, so the estimate will be sensitive to the choice of bandwidth. One theoretical problem with these monotonic regression procedures has been the lack of availability of asymptotic results concerning the proposed estimators.

Recently, there has been considerable progress in the characterization of asymptotic properties for nonparametric estimators subject to order constraints (Banerjee, 2000; Banerjee and Wellner, 2001). A key feature of these problems is that in contrast to many statistical problems, classical regularity conditions do not hold. Thus, the consistency and asymptotic normality results for maximum likelihood estimators in parametric models are not relevant. In this article, we develop statistical estimation procedures for nonparametric isotonic models of biomarkers on disease risk. The structure of this paper is as follows. In Section 2, we describe the observed data structures and probability model. We also relate the model to optimal screening rules using a classification point of view. In Section 3, we present the proposed estimation procedure and associated asymptotic results. In Section 4, we discuss the construction of confidence intervals for the nonparametric function. We also describe procedures for estimating receiver operating characteristic (ROC) curves in Section 5. The finite-sample properties of the estimators are assessed using simulation studies and are applied to data from a pancreatic cancer study in Section 6. We conclude with some discussion in Section 7.

2. Statistical Methodology

2.1. Preliminaries and Notation

Let D be a binary variable representing the indicator of disease and Z the biomarker. Without loss of generality, we assume Z to be nonnegative. To model the effects of the biomarker on disease risk, we formulate the following model:

$$Pr(D = 1 \mid Z) = G(Z), \tag{1}$$

where G is assumed to be a monotonic increasing and continuously differentiable on $[0, \infty)$ with $G(0) = 0$ and $\lim_{z\to\infty} G(z) = 1$. We are interested in making inferences on G.

Let $(D_1, Z_1), (D_2, Z_2), \ldots, (D_n, Z_n)$ be n i.i.d. observations from the above model. The joint density of (D, Z) is given by

$$p(d, z) = \{G(z)\}^d \{1 - G(z)\}^{1-d} h(z), \tag{2}$$

where $h(\cdot)$ is the density function of Z. The likelihood function for the data, up to a multiplicative constant not involving h, is given by

$$L_n(\{d_i, z_i\}_{i=1}^n) = \Pi_{i=1}^n \{G(z_i)\}^{d_i} \{1 - G(z_i)\}^{1-d_i}. \tag{3}$$

Before developing the estimation procedures in (1) using maximization of the non-parametric likelihood (3), we first relate (1) to the literature on screening rules. In particular, we show that (1) satisfies a certain optimality criterion useful for classification.

2.2. Optimal screening rules

Consider a diagnostic test based on thresholding Z. One relevant quantity is the false positive rate based on a cutoff c, defined to be $FP(c) = P(Z > c|D = 0)$. Another is the true positive rate $TP(c) \equiv P(Z > c|D = 1)$. The true and false positive rates can be summarized by the receiver operating characteristic (ROC) curve, which is a graphical presentation of $\{FP(c), TP(c) : -\infty < c < \infty\}$. The ROC curve shows the tradeoff between increasing true positive and false positive rates. Tests that have $\{FP(c), TP(c)\}$ values close to (0,1) indicate perfect discriminators, while those with $\{FP(c), TP(c)\}$ values close to the 45 degree line in the $(0, 1) \times (0, 1)$ plane are tests that are unable to discriminate between the diseased and healthy populations. An alternative terminology used for summarizing the operating characteristics of a test is

sensitivity (equal to the true positive rate) and specificity (equal to one minus the false positive rate).

Suppose we fix a false positive rate, say $f_0 \equiv FP(c_0)$ and compare the performance of two tests. Then the test that gives a higher true positive rate will be the better test. Suppose that given f_0, we wish to find the screening rule with the highest true positive rate. It follows from Neyman-Pearson theory (Lehmann, 1997, §3.2) that the classification rule based on

$$\mathrm{LR}(Z) \equiv \frac{Pr(Z|D=1)}{Pr(Z|D=0)} > c(f_0) \tag{4}$$

optimizes the true positive rate for a given false positive rate f_0, where $c(f_0)$ is chosen such that $Pr\{LR(Z) > c(f_0)\} = f_0$, and $f_0 \in [0,1]$. Such rules satisfying this property have been referred to by McIntosh and Pepe (2002) as uniformly most sensitive (UMS) screening tests based on Z. Using Bayes' rule to invert expression (4), we have that

$$\begin{aligned} \frac{Pr(Z|D=1)}{Pr(Z|D=0)} &= \frac{Pr(D=1|Z)Pr(Z)}{Pr(D=0|Z)Pr(Z)} \\ &= \frac{Pr(D=1|Z)}{1-Pr(D=1|Z)}. \end{aligned}$$

UMS rules can thus also be constructed based on the risk score $Pr(D=1|Z)$. This corresponds to the specification (1) we have utilized in this paper. This gives a justification of our model from a classification point of view.

Two methods that have been used for classification are linear discriminant analysis and logistic regression. These methods, as well as (1) in this paper, can be motivated as generating UMS screening rules. Our model (1) requires fewer assumptions than linear discriminant analysis and logistic regression. Linear discriminant analysis (LDA) requires normality of the measurements in both diseased and undiseased populations, while logistic regression (LR) assumes that the effect of the biomarker on disease risk is linear on the logit scale. Intuitively, our method will perform better than these other two methods when the underlying assumptions of LDA and LR are violated.

3. Estimation Procedures and Asymptotic Results

We now consider estimation in (1) by maximization of (3). One nice feature of this model is the fact that the likelihood function (3) can be shown to be equivalent to a likelihood function arising from a current status censoring scheme. Current status data represent a failure time data structure in which what is observed is a monitoring time and an indicator of whether or not the event occurred before the monitoring time. Nonparametric theory for the current status model, which is very well-studied (Huang and Wellner, 1997), can therefore be used to analyze the model. Let T be a non–negative random variable with distribution function G; we think of T as the survival time of an individual. Let Y be the time at which the individual is observed. Let Y be independent of T and distributed with density h. Our data consists of (Y, I) where $I = 1$ if $T \leq Y$ and 0 otherwise. Then, it is easy to see that the joint distribution of the vector (I, Y) is identical to the joint distribution of (D, Z) given by (2). Thus, $\{(D_1, Z_1), \ldots, (D_n, Z_n)\}$ can be viewed as a random sample from a current status censoring model, with distribution function G and censoring density h. The NPMLE of the survival distribution in the current status model is well-studied in Groeneboom and Wellner (1992) while the likelihood ratio for testing G at one or multiple points is dealt with in Banerjee (2000) and Banerjee and Wellner (2001). Thus, many of the results from this area can be applied to our problem here.

We now characterize the unconstrained MLE of G, say $\hat{G}_n$. Let $Z_{(i)}$ denote the ith largest value of the biomarker and let $D_{(i)}$ denote the corresponding indicator $(i = 1, \ldots, n)$. For arbitrary points $P_0 \equiv (0, 0), P_1 \equiv (p_{1,1}, p_{1,2}), \ldots, P_k \equiv (p_{k,1}, p_{k,2})$ in R^2, we will denote by slogcm $\{P_i\}_{i=0}^k$ the vector of slopes (left derivatives) of the greatest convex minorant (GCM) of the piecewise linear curve that connects $P_0, P_1, \ldots, P_k$ in that order, computed at the points $\{p_{i,1}\}_{i=1}^k$. We can set $\hat{G}_n$ to be the right–continuous piecewise constant increasing function which satisfies $G(Z_{(i)}) = \hat{u}_i$ where

$$(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm}\,\{i, \sum_{j=0}^{i} d_{(j)}\}_{i=0}^n,$$

and $Z(0) = d_{(0)} = 0$. Note the MLE is uniquely determined only up to its values at the Z_i, $i = 1, \ldots, n$. A closed-form formula for the unconstrained estimator is given by the "max-min" formula

$$\hat{G}_n(Z_{(i)}) = \max_{j \le i} \min_{k \ge i} \frac{\sum_{j \le i \le k} d_{(i)}}{k - j + 1}. \tag{5}$$

We now state the asymptotic distribution of the MLE of $G(z_0)$ in model (1). This result can be proven using arguments paralleling those in the proof of Theorem 5.1 of Groenenboom and Wellner (1992, p. 89).

Theorem 1: *The unconstrained MLE $\hat{G}_n(z_0)$ has the following limiting distribution:*

$$n^{1/3} \left\{ \hat{G}_n(z_0) - G(z_0) \right\} \to_d \left[\frac{4\, g(z_0)\, G(z_0)\, \{1 - G(z_0)\}}{h(z_0)} \right]^{1/3} Z \equiv CZ,$$

where $g(z_0)$ is the derivative of G evaluated at z_0, Z is the location of the minimum of $W(t) + t^2$; here W is a standard two-sided Brownian motion starting from 0.

Theorem 1 yields a complicated form for the limiting distribution of the maximum likelihood estimator. By contrast, for most statistical estimation problems in which classical regularity conditions are satisfied, the maximum likelihood estimator converges at an $n^{1/2}$ rate. In addition, the limiting distribution of the NPMLE estimator, properly normalized, is much more complicated than the normal distribution found for regular estimation problems. We will later consider construction of Wald-type confidence intervals based on this result.

One goal in this paper will be to make inference about the null hypothesis $H_0 : G(z_0) = \theta_0$. To do this requires characterization of the constrained MLE of G subject to $G(z_0) = \theta_0$, which we denote $\hat{G}_n^0$. Define $a \wedge b$ and $a \vee b$ to be the minimum and maximum of two real numbers a and b. We construct the vector of slopes of the greatest convex minorant:

$$(\tilde{u}_1^0, \ldots, \tilde{u}_m^0) = \text{slogcm}\, \{i, \sum_{j=0}^{i} d_{(j)}\}_{i=0}^{m}$$

and

$$(\tilde{u}_{m+1}^0, \ldots, \tilde{u}_n^0) = \text{slogcm}\, \{i, \sum_{j=0}^{i} d_{(m+i)}\}_{i=0}^{n-m},$$

where m is the number of biomarker values no greater than z_0. The constrained estimate $\hat{G}_n^0$ is the right-continuous piecewise constant function that satisfies $G(Z_{(i)}) = \tilde{u}_i^0 \wedge \theta_0$ for $1 \leq i \leq m$ and $G(Z_{(i)}) = \tilde{u}_i^0 \vee \theta_0$ for $m+1 \leq i \leq n$. Similar to before, max-min formulae could be developed for characterization of the constrained MLE.

In order to state results about the asymptotic distribution of the likelihood ratio statistic for testing $H_0 : G(z_0) = \theta_0$, we will need some more notation. For positive constants a and b, define the process $X_{a,b}(z) \equiv a\,W(z) + b\,z^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Let $G_{a,b}(z)$ denote the GCM of $X_{a,b}(z)$. Let $g_{a,b}(z)$ be the right derivative of $G_{a,b}$; this can be shown to be a piecewise constant (increasing) function, with finitely many jumps in any compact interval. We construct $G_{a,b}^0(z)$, a random function, in the following manner:

(a) When $z < 0$, we restrict ourselves to the set $\{z < 0\}$ and compute $X_{a,b}(z)$. $G_{a,b}^0(z)$ is the GCM of $X_{a,b}(z)$, constrained so that its slope (right derivative) is non-positive;

(b) When $z > 0$, we restrict ourselves to the set $\{z < 0\}$ and compute $X_{a,b}(z)$. $G_{a,b}^0(z)$ is the GCM of $X_{a,b}(z)$, constrained so that its slope (right derivative) is non-negative.

We have that $G_{a,b}^0(z)$ will almost surely have a jump discontinuity at zero. Let $g_{a,b}^0(z)$ be the slope (right-derivative) of $G_{a,b}^0(z)$; this, like $g_{a,b}(z)$, is a piecewise constant (increasing) function, with finitely many jumps in any compact interval and differing (almost surely) from $g_{a,b}(z)$ on a finite interval containing 0. Thus, $g_{1,1}$ and $g_{1,1}^0$ are the unconstrained and constrained versions of the slope processes associated with the canonical process $X_{1,1}(z)$.

The following theorem describes the joint limit behavior of the unconstrained and constrained MLEs of G, the constraint being that imposed by the null hypothesis $H_0 : G(z_0) = \theta_0$. This can be proven using arguments similar to those in the proof of Theorem 2.6.1 of Banerjee and Wellner (2001):

Theorem 2: *Consider testing the null hypothesis* $H_0 : G(z_0) = \theta_0$ *with* $0 < z_0 < \infty$ *and* $0 < \theta_0 < 1$ *and assume* H_0 *holds. Let*

$$X_n(t) = n^{1/3} \{\hat{G}_n(z_0 + t\, n^{-1/3}) - \theta_0\} \;\; \textit{and} \;\; Y_n(t) = n^{1/3} \{\hat{G}^0_n(z_0 + t\, n^{-1/3}) - \theta_0\}\,.$$

Suppose that G *is continuously differentiable in a neighborhood of* z_0 *with* $g(z_0) > 0$ *and that* h *is continuous in a neighborhood of* z_0 *with* $h(z_0) > 0$. *Let*

$$a = \left[\frac{G(z_0)\{1 - G(z_0)\}}{h(z_0)}\right]^{1/2}$$

and $b = g(z_0)/2$. *Then*

$$\{X_n(t), Y_n(t)\} \to \{g_{a,b}(t), g^0_{a,b}(t)\} \equiv_d \left[a\,(b/a)^{1/3} g_{1,1}\left\{(b/a)^{2/3}t\right\}, a\,(b/a)^{1/3} g^0_{1,1}\left\{(b/a)^{2/3}t\right\}\right]\,.$$

finite dimensionally and also in the space $\mathcal{L}_p[-K, K] \times \mathcal{L}_p[-K, K]$ *for every* $K > 0$ *(*$p \geq 1$*), where* $\mathcal{L}_p[-K, K]$ *is the set of functions that are* L^p *integrable on* $[-K, K]$.

Based on this result, we can develop the asymptotic theory for the likelihood ratio test statistic $H_0 : G(z_0) = \theta_0$.

Theorem 3: *If* λ_n *denotes the likelihood ratio, i.e.*

$$\lambda_n = \frac{\Pi_{i=1}^n \{\hat{G}_n(z_i)\}^{d_i} \{1 - \hat{G}_n(z_i)\}^{1-d_i}}{\Pi_{i=1}^n \{\hat{G}^0_n(z_i)\}^{d_i} \{1 - \hat{G}^0_n(z_i)\}^{1-d_i}}\,,$$

then the limiting distribution of the likelihood ratio statistic for testing $H_0 : G(z_0) = \theta_0$ *is*

$$2 \log\, \lambda_n \to_d \mathcal{D} \equiv \int \left[\{g_{1,1}(z)\}^2 - \{g^0_{1,1}(z)\}^2\right]\, dz\,.$$

A heuristic proof of this theorem is given in the Appendix. Note that the limiting distribution of Theorem 3 is much different from that in regular statistical problems, where $2\lambda_n$ converges to a chi-squared distribution. The random variable $\mathcal{D}$ can be thought of as an analog of the χ^2 distribution to nonregular problems.

To visualize $\mathcal{D}$, we present a plot from a simple simulation study with model (1). We let $n = 100, Z \sim U(0, 1)$ and $G(z) = z^2$. We consider testing $H_0 : G(0.7) = 0.49$.

We generate 2000 samples of size 100 from the model and for each sample, compute the likelihood ratio statistic for testing H_0. The empirical distribution function of the likelihood ratio statistic is shown in red in Figure 1 and along with it, is displayed the empirical distribution from (a discrete approximation to) the theoretical limit. The two lines are seen to be in good agreement. The random variable $\mathcal{D}$ can be tabulated or simulated relatively easily. For further details, see Banerjee (2000) or Banerjee and Wellner (2001).

4. Confidence intervals for the distribution function

In this section we investigate the properties of confidence sets for the the value of the conditional probability function G at some point $z_0 > 0$. We denote $G(z_0)$ by θ. At first glance, it appears that one could use the nonparametric bootstrap (Efron and Tibshirani, 1986) and construct confidence sets for θ by using the empirical distribution of the NPMLE based on the bootstrapped datasets. However, because the statistical estimation problem here is nonregular, the use of the bootstrap is not valid. We refer the reader to §3.1 of Bühlmann and Yu (2002) for a discussion of this point. We discuss three methods for confidence set construction: (i) the Wald-based method; (ii) the subsampling based method; and (iii) the likelihood ratio based method.

4.1. Wald-based method

Recall the limiting distribution of $\hat{G}_n(z_0)$ from §3:

$$n^{1/3}\left\{\hat{G}_n(z_0) - G(z_0)\right\} \to_d \left[\frac{4\, g(z_0)\, G(z_0)\,\{1 - G(z_0)\}}{h(z_0)}\right]^{1/3} Z \equiv CZ.$$

A 95% confidence interval for $G(z_0)$ is then given by

$$\{G_n(z_0) - n^{-1/3}\,\hat{Q}_{.975}, G_n(z_0) + n^{-1/3}\,\hat{Q}_{.975}\},$$

where $\hat{Q}_{.975}$ is a consistent estimator of $Q_{.975}$, the 97.5th percentile of the limiting symmetric random variable CZ. But $Q_{.975}$ is simply $C \times .99818$ where .99818 is the

97.5th percentile of Z, where the quantiles of Z are from Groenenboom and Wellner (2001). Since C involves the unknown parameters $G(z_0)$, $h(z_0)$, and $g(z_0)$, we estimate C by

$$\widehat{C}_n = \left[\frac{4\hat{g}_n(z_0)\,\hat{G}_n(z_0)\,\{1-\hat{G}_n(z_0)\}}{\hat{h}_n(z_0)} \right]^{1/3},$$

where $\hat{g}_n$ and $\hat{h}_n$ are estimates of g and h. An asymptotic 95% confidence interval is then given by

$$\left\{ \hat{G}_n(z_0) - n^{-1/3}\,\widehat{C}_n \times .99818\ ,\ \hat{G}_n(z_0) + n^{-1/3}\,\widehat{C}_n \times .99818 \right\}.$$

The major drawback of these intervals is the need to estimate $g(z_0)$ and $h(z_0)$. We first consider $h(z_0)$. Since Z is observed for all individuals, nonparametric density estimation methods can be used to estimate $h(z_0)$. On the other hand, $g(z_0)$ is much more difficult to estimate consistently. Due to the data structure and model, we can only estimate G at $O_p(n^{1/3})$ support points, which means that we will never have sufficiently large sample sizes for estimating the derivative of G consistently. In the simulation studies and real data example presented in §6, we estimate the derivative of G using smoothing splines (Heckman and Ramsay, 2000). The smoothing parameter is chosen using generalized cross-validation (Craven and Wahba, 1979).

4.2. Subsampling-based method

The subsampling technique followed here is due to Politis, Romano and Wolf (1999) and is part of a general theory for obtaining confidence regions. The basic idea is to approximate the sampling distribution of a statistic, based on the values of the statistic computed over smaller subsets of the data. We start by calculating the unconstrained MLE $\hat{G}_n(z_0)$ for the observed dataset. This leads to the following algorithm:

1. Create a dataset $(D_1^*, Z_1^*), \ldots, (D_b^*, Z_b^*)$, where (D_j^*, Z_j^*) $(j = 1, \ldots, b)$ are a subset of the original data obtained by sampling without replacement, and b is the size of the subsampled dataset.

2. Calculate the unconstrained MLE $\hat{G}_n^*(z_0)$ for the dataset.

3. Repeat steps (1) and (2) several times.

By Theorem 2.2.1. of Politis et al. (1999), it follows that if $b, n \to \infty$ and $b/n \to 0$, then the conditional distribution of $n^{1/3}\{\hat{G}_n^*(z_0) - \hat{G}_n(z_0)\}$ converges to the unconditional distribution of $n^{1/3}\{\hat{G}_n(z_0) - G(z_0)\}$ with probability 1. This allows us to use the empirical distribution of $n^{1/3}\{\hat{G}_n^*(z_0) - \hat{G}_n(z_0)\}$ to construct confidence intervals.

While this appears to be a promising algorithm, a major issue is the choice of b. For the data example, we use a calibration algorithm, proposed in Delgado et al. (2001):

(a) Fix a selection of reasonable block sizes between limits b_{low} and b_{up}.

(b) Generate K pseudo sequences $(D_k^\star, Z_k^\star)_{k=1}^K$ which are i.i.d. $\hat{P}_n$; with $\hat{P}_n$ equal to the empirical distribution function this amounts to drawing K bootstrap samples from the actual data set.

(c) For each pseudo data set, construct a subsampling based confidence interval for $\hat{\theta}_n \equiv G_n(z_0)$ for each block size b. Let $I_{k,b}$ be equal to 1, if $\hat{\theta}_n$ lies in the k'th interval based on block size b and 0 otherwise.

(d) Compute $\hat{h}(b) = K^{-1} \sum_{i=1}^K I_{k,b}$.

(e) Find $\tilde{b}$ that minimizes $\mid \hat{h}(b) - (1-\alpha) \mid$ and use this as the block size to compute subsampling based confidence intervals based on the original data.

4.3. Likelihood ratio based method

Confidence sets of level $1-\alpha$ with $0 < \alpha < 1$ are obtained by inverting the acceptance region of the likelihood ratio test of size α; more precisely if $2\log\lambda_n$ is the likelihood ratio statistic evaluated under the null hypothesis $H_0 : G(z_0) = \theta_0$, then the set of all values of θ for which $2\log\lambda_n$ is not greater than d_α where d_α is the $1-\alpha$th percentile of D, gives us a limiting level $1-\alpha$ confidence set for θ.

Denote the confidence set of (approximate) level $1-\alpha$ based on a sample of size n from the binary regression problem by $C_{n,\alpha}$. Thus $C_{n,\alpha} = \{\theta : 2\log\,\lambda_n \leq d_\alpha\}$. Because we are inverting the likelihood ratio statistic, it achieves the correct coverage asymptotically. The finite-sample properties of this method are examined in Section 6.

5. Summaries using ROC curves

One summary of the estimated function $\hat{G}$ from (1) that is appealing to clinicians is the ROC curve. In our situation, given the estimator $\hat{G}$, we consider a grid of possible cutoff values $c^* = 0, 0.01, 0.02, \ldots, 1$ and calculate the following two quantities:

$$\widehat{TPR}(c^*) = \frac{\sum_{i=1}^n I\{\hat{G}(Z_i) > c^*, D_i = 1\}}{\sum_{i=1}^n I(D_i = 1)}$$

and

$$\widehat{FPR}(c^*) = \frac{\sum_{i=1}^n I\{\hat{G}(Z_i) > c^*, D_i = 0\}}{\sum_{i=1}^n I(D_i = 0)}.$$

A plot of $\{\widehat{TPR}(c^*), \widehat{FPR}(c^*) : c^* = 0, 0.01, 0.02, \ldots, 1\}$ then provides an ROC curve based on the estimated model (1).

Note that the ROC curve will be overoptimistic in that the predictions for $\widehat{TPR}$ and $\widehat{FPR}$ were based on the $\hat{G}$, which was estimated using the entire dataset. Thus, the ROC curves will be overoptimistic. An alternative is to construct a leave-one-out cross-validation estimate of the ROC curve. Let $\tilde{G}^{(i)}$ denote the estimate of G in (1) with the ith observation held out, $i = 1, \ldots, n$. Then a cross-validated estimate of the ROC curve is given by $\{\widetilde{TPR}(c^*), \widetilde{FPR}(c^*) : c^* = 0, 0.01, 0.02, \ldots, 1\}$, where

$$\widetilde{TPR}(c^*) = \frac{\sum_{i=1}^n I\{\tilde{G}^{(i)}(Z_i) > c^*, D_i = 1\}}{\sum_{i=1}^n I(D_i = 1)}$$

and

$$\widetilde{FPR}(c^*) = \frac{\sum_{i=1}^n I\{\tilde{G}^{(i)}(Z_i) > c^*, D_i = 0\}}{\sum_{i=1}^n I(D_i = 0)}.$$

Note that since our procedure does not require smoothing, it is fairly easy to calculate the isotonic regression-based estimators for each of the samples with one observation left out.

6. Numerical examples

6.1. Simulation studies

To compare the various confidence set methods described in §4, several simulation studies were conducted. We considered (1) in which Z has an exponential distribution with rate parameter 1.5 and $G(z) = z(3 + e^z)^{-1}$. The goal is to test $H_0 : G(1) = 0.475$. Sample sizes $n = 50, 100, 150, 200, 300, 500, 800, 1000$ and 1500 were considered. With the subsampling-based method, a prior series of simulations were performed in order to pick out the optimal sample size. As mentioned earlier, for the Wald intervals, numerical derivatives of a smoothing spline-based estimate of $\hat{G}(z)$ were used to estimate $g(z)$ from the algorithm described in Heckman and Ramsay (2000). The results are summarized in Table 1. There are several points to note from here. First, the isotonic regression based estimator of $G(z_0)$ is approximately unbiased for all sample sizes considered. Second, while the Wald confidence intervals tend to give the shortest confidence intervals based on length, their nominal coverage is far from 0.95. As the sample size increases, the coverage decreases. By contrast, the confidence intervals obtained using subsampling and inversion of the likelihood ratio test statistic yield relatively consistent estimates of the coverage probability. Comparing these two intervals, we find that the length of the likelihood ratio test statistic-based confidence interval tends to be slightly smaller than that from the subsampling procedure. Given this observation and the overall computational complexity of the latter, we advocate use of the likelihood ratio test statistic method for constructing confidence intervals in practice.

6.2. Pancreatic cancer data

We now apply the proposed methodology to data from a pancreatic study used in Pepe and Thompson (2000). They come from a study in which measurements on two serum biomarkers, CA125 and CA19-9, were collected on 90 pancreatic cancer patients

and 51 patient without the disease. To illustrate the methodology, we consider data on CA19-9 and CA125 separately.

CA19-9 (Carbohydrate Antigen 19-9) is present in the fetal epithelium of the stomach, intestine, liver and pancreas. In adults, traces can be found in the pancreas, liver, gallbladder and lung. While its primary use is for pancreatic cancer, it is also elevated in cancers of the liver, lung, breast, uterus and ovary (mucinous). CA19-9 may also be elevated in non-malignant liver diseases such as cirrhosis and hepatitis. A logarithmic transformation was taken to reduce skewness. The estimate of G at several points and the 95% pointwise confidence intervals from the three methods are summarized in Table 2. Based on the results, we find that the Wald-based CI generally gives shorter intervals than the other two methods. However, in light of the simulation results, the intervals from the Wald method might not achieve 95% coverage. One other point to note is that some of the intervals will contain values greater than one. This is because of the fact that we have not constrained estimates from (1) to be less than or equal to one. A practical solution is to truncate the interval at one.

Using the estimated function $\hat{G}$, one can construct an ROC curve from the methods in §5. For the purposes of comparison, a logistic regression model was used to model the effect of CA19-9 on pancreatic cancer, and the resulting parameter estimates were used to construct an ROC curve. Both the data-based and leave-one-out cross-validated ROC curves were estimated; these are presented in Figures 2a and 3a. While the leave-one-out cross-validated curves give lower true positive rates, it appears that the logistic regression is performing better. This suggests that the logistic regression model is a reasonable assumption for CA19-9.

The CA125 (Cancer Antigen 125) assay uses a monoclonal antibody that is relatively specific for surface antigen derived from a papillary serous cystadenocarcinoma. It is detectable in various adult tissues, such as the pleura, pericardium and eritoneum. It is used primarily as a marker for ovarian cancer, although elevated CA125 levels can also be seen in association with malignancies of breast, cervix, uterus, liver, pancreas,

stomach, colorectum and lung. Non-malignant elevations of CA 125 have been reported for a variety of inflammatory diseases, such as cirrhosis, hepatitis, pancreatitis, and pelvic inflammatory disease. As with CA19-9, logarithmic transformation was taken to adjust for skewness in the measurement. The estimated G and confidence intervals are presented in Table 3. We find the same types of conclusions as in Table 2. Plots of the data-based and leave-one-out cross-validated ROC curves are given in Figures 2b and 3b. We find that the isotonic regression is performing quite competitively relative to logistic regression; there is slight evidence to suggest that it is performing better for the leave-one-out ROC curve.

7. Discussion

In this article, we have proposed a nonparametric isotonic regression model for the analysis of biomarker data. This model is more flexible in that the effect of the biomarker on disease risk does not have a parametric form. By utilizing recently developed estimation methods and asymptotic results from the literature, we are able to characterize and provide the asymptotic distribution of the maximum likelihood estimator and the likelihood ratio test statistic for the conditional probability function. Note that unlike previous nonparametric estimators for monotone regression models, our estimator does not require smoothing and hence does not depend on a smoothing parameter. The simulation results indicated that in finite-samples, the Wald statistic-based confidence intervals tend to have poor coverage probabilities. Thus, we advocate the use of the likelihood ratio test statistic for constructing confidence intervals.

One limitation with (1) is that no other covariates are included. In screening settings, it might be important to account for potential confounders, such as demographic covariates. Thus, one might envision extensions of (1) in which the effect of the biomarker is monotone on disease risk and the other effect of other covariates has a parametric form. Inference and estimation with semiparametric monotone regression models has not been well-studied and is an area we are currently pursuing.

In the data analysis, we presented pancreatic cancer biomarker data in which the goal was to model the effects of the individual biomarkers, CA125 and CA19.9, separately on disease risk. An alternative objective is to combine the information on the two biomarkers in order to maximize screening accuracy. This is an area studied recently by Baker (2000), Pepe and Thompson (2000), and McIntosh and Pepe (2002). How to incorporate monotone restrictions in this problem remains fairly uninvestigated. Two procedures have been developed by Dykstra et al. (1999) and Wang and Taylor (2004). We are currently studying multivariate generalizations of (1).

Acknowledgments

The first author would like to thank Jeremy Taylor and Yue Wang for useful discussions. The second author acknowledges the support of grant DMS-0306235 from the National Science Foundation.

Appendix

Sketch of Proof of Theorem 2:

We can write $2\log\lambda_n$ as

$$\begin{aligned} 2\log\lambda_n &\equiv 2\sum_{i=1}^{n} d_i \log\hat{G}_n(z_i) + (1-d_i)\log\{1-\hat{G}_n(z_i)\} \\ &\quad -2\sum_{i=1}^{n} d_i \log\hat{G}_n^0(z_i) + (1-d_i)\log\{1-\hat{G}_n^0(z_i)\}. \end{aligned} \tag{6}$$

Rearranging terms in (6), we get

$$\begin{aligned} 2\log\lambda_n &= 2\left\{\sum_{i\in D} d_i \log\hat{G}_n(z_i) - \sum_{i\in D} d_i \log\hat{G}_n^0(z_i)\right\} \\ &\quad -2\left[\sum_{i\in D}(1-d_i)\log\{1-\hat{G}_n(z_i)\} - \sum_{i\in D}(1-d_i)\log\{1-\hat{G}_n^0(z_i)\}\right] \\ &= A_n + B_n \end{aligned} \tag{7}$$

where J_n is the set of indices where $\hat{G}_n$ and $\hat{G}_n^0$ differ. By Taylor series expansion and algebraic manipulation,

$$\begin{aligned} A_n &= 2\sum_{i\in J_n}\frac{d_i}{\theta_0}\left[\{\hat{G}_n(z_i)-\theta_0\}-\{\hat{G}_n^0(z_i)-\theta_0\}\right] \\ &\quad -2\sum_{i\in J_n}\frac{d_i}{2\theta_0^2}\left[\{\hat{G}_n(z_i)-\theta_0\}^2-\{\hat{G}_n^0(z_i)-\theta_0\}^2\right]+o_P(1) \qquad (8) \end{aligned}$$

and

$$\begin{aligned} B_n &= -2\sum_{i\in J_n}\frac{1-d_i}{1-\theta_0}\left[\{\hat{G}_n(z_i)-\theta_0\}-\{\hat{G}_n^0(z_i)-\theta_0\}\right] \\ &\quad +2\sum_{i\in J_n}\frac{1-d_i}{2(1-\theta_0)^2}\left[\{\hat{G}_n(z_i)-\theta_0\}^2-\{\hat{G}_n^0(z_i)-\theta_0\}^2\right]+o_P(1), \qquad (9) \end{aligned}$$

where $o_P(1)$ denotes a term that converges in probability to zero. Plugging in representations (8) and (9) of A_n and B_n into (7), we have

$$\begin{aligned} 2\log\lambda_n &= 2\sum_{i\in J_n}\frac{d_i-\theta_0}{\theta_0(1-\theta_0)}[\{\hat{G}_n(z_i)-\theta_0\}-\{\hat{G}_n^0(z_i)-\theta_0\}] \\ &\quad -2\sum_{i\in J_n}\left\{\frac{d_i}{2\theta_0^2}+\frac{(1-d_i)}{2(1-\theta_0)^2}\right\}\left[\{\hat{G}_n(z_i)-\theta_0\}^2-\{\hat{G}_n^0(z_i)-\theta_0\}^2\right]+o_P(1) \\ &= I_1-I_2+o_P(1). \qquad (10) \end{aligned}$$

Let D_n be the interval on which $\hat{G}_n$ and $\hat{G}_n^0$ differ. Define the processes $G_n(z)$ and $V_n(z)$ by

$$G_n(z)=\frac{1}{n}\sum_{i=1}^{n}1(Z_{(i)}\le z) \text{ and } V_n(z)=\frac{1}{n}\sum_{i=1}^{n}d_{(i)}1(Z_{(i)}\le z)\,.$$

Note that $\{G_n(Z_{(i)}),V_n(Z_{(i)})\}$ $(i=1,\ldots,n)$ characterize the points of the cusum diagram used to compute the NPMLE's $\hat{G}_n$ and $\hat{G}_n^0$. We can rewrite I_1 as

$$\begin{aligned} I_1 &\equiv \frac{2}{\theta_0\,(1-\theta_0)}n\int_{D_n}\{(\hat{G}_n(z)-\theta_0)-(\hat{G}_n^0(z)-\theta_0)\}\,d\{V_n(z)-\theta_0\,G_n(z)\}, \qquad (11) \\ &= \frac{2}{\theta_0\,(1-\theta_0)}\,n\int_{D_n}\{(\hat{G}_n(z)-\theta_0)^2-(\hat{G}_n^0(z)-\theta_0)^2\}\,d\,G_n(z) \qquad (12) \\ &= \frac{2}{\theta_0\,(1-\theta_0)}\,n\int_{D_n}\{(\hat{G}_n(z)-\theta_0)^2-(\hat{G}_n^0(z)-\theta_0)^2\}\,h(z)\,dz+o_P(1) \qquad (13) \\ &= \frac{2\,h(z_0)}{\theta_0\,(1-\theta_0)}\int_{\tilde{D}_n}\{X_n^2(t)-Y_n^2(t)\}\,dt+o_p(1) \qquad (14) \\ &= \frac{2}{a^2}\int_{\tilde{D}_n}\{X_n^2(t)-Y_n^2(t)\}dt+o_P(1). \qquad (15) \end{aligned}$$

Note that (12) follows from (11) using integration by parts. Also, (14) follows from (13) on changing to the local variable $t \equiv n^{1/3}(z - z_0)$. Also, $\tilde{D}_n \equiv n^{1/3}\,(D_n - t - 0)$ is the set where the processes X_n and Y_n differ. As a consequence of Theorem 2, the set $\tilde{D}_n$ is $O_p(1)$. By similar arguments,

$$I_2 = (1/a^2) \int_{\tilde{D}_n} \{X_n^2(t) - Y_n^2(t)\} dt + o_P(1). \tag{16}$$

Combining (15) and (16) and using the representation (10) for the likelihood ratio statistic, we have

$$2 \log \lambda_n = \frac{1}{a^2} \int_{\tilde{D}_n} \{X_n^2(t) - Y_n^2(t)\} dt\,.$$

Using the distributional convergence of $\{X_n(t), Y_n(t)\}$ to $\{g_{a,b}(t), g^0_{a,b}(t)\}$ as processes from Theorem 2, we conclude that

$$2 \log \lambda_n \rightarrow_d \frac{1}{a^2} \int \left[\{g_{a,b}(t)\}^2 - \{g^0_{a,b}(t)\}^2\right] dt\,.$$

Expressing $\{g_{a,b}(t), g^0_{a,b}(t)\}$ in terms of the slopes of the convex minorants of the canonical process from Theorem 2, we obtain

$$\begin{aligned} \frac{1}{a^2} \int \left[\{g_{a,b}(t)\}^2 - \{g^0_{a,b}(t)\}^2\right] dt &= \frac{1}{a^2}\, a^2\, (b/a)^{2/3} \int \left([g_{1,1}\{(b/a)^{2/3}\, t\}^2 - [g^0_{1,1}\{(b/a)^{2/3}\, t\}]^2\right) dt \\ &= \int \left[\{g_{1,1}(w)\}^2 - \{g^0_{1,1}(w)\}^2\right] dw \equiv \mathcal{D}\,, \end{aligned}$$

on changing variables: $w = (b/a)^{2/3} t$.

References

Baker, S. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082 – 1087.

Banerjee, M. (2000). *Likelihood ratio inference in regular and nonregular problems.* Ph.D. dissertation, University of Washington.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Annals of Statistics* **29**, 1699 – 1731.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Annals of Statistics* **30**, 927 – 961.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothness by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377 – 403.

Delgado, Miguel A., Rodriguez-Poo, Juan M. and Wolf, M. (2001). Subsampling Inference in Cube Root Asymptotics with an application to Manski's Maximum Score Estimator. *Economics Letters* **73**, 241 – 250.

Dykstra, R., Hewett, J. and Robertson, T. (1999). Nonparametric, isotonic discrimant procedures. *Biometrika* **86**, 429 – 438.

Efron, B. , and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statistical Science* **1**, 54 – 75.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Birkhäuser: Boston.

Groeneboom, P. and Wellner, J. A. (2001). Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics* **10**, 388 – 400.

He, X. and Shi, P. (1998). Monotone B-spline smoothing. *J. Am. Statist. Assoc.* **93**, 643 – 650.

Heckman, N. E. and Ramsay, J. O. (2000) Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28**, 241 – 258.

Huang, J. and Wellner, J. A. (1997). Interval Censored Survival Data: A Review of Recent Progress. Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Eds. D. Lin and T. Fleming. Springer-Verlag, New York.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and assessment of synergism. *Biometrics* **46**, 1071 – 1085.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses, 2nd Ed.* New York: Springer.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724 – 740.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657 – 664.

Mukerjee, H. (1988). Monotone nonparametric regression. *Ann. Statist.* **16**, 741 – 750.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123 – 140.

Politis, D.N., Romano, J.P. and Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wang, Y. and Taylor, J. (2004). Monotone Constrained Tensor-product B-spline with application to screening studies. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 23. Available at the following URL: http://www.bepress.com/umichbiostat/paper23

Table 1. *Summary of simulation results*

		Wald intervals			Subsampling intervals			LRT intervals		
n	*Bias*	LHM	RHM	CP	LHM	RHM	CP	LHM	RHM	CP
50	.302	0.237	0.713	0.867	0.179	0.774	0.948	0.276	0.720	0.936
100	-5.409	0.295	0.666	0.807	0.283	0.673	0.941	0.315	0.666	0.945
150	-0.156	0.327	0.628	0.752	0.310	0.646	0.952	0.335	0.643	0.945
200	-0.006	0.339	0.612	0.733	0.325	0.628	0.953	0.345	0.624	0.944
300	-4.969	0.370	0.591	0.682	0.348	0.610	0.954	0.366	0.605	0.949
500	0.215	0.383	0.567	0.667	0.371	0.581	0.956	0.378	0.585	0.960
800	-2.169	0.407	0.548	0.605	0.384	0.573	0.954	0.392	0.567	0.952
1000	-2.151	0.410	0.545	0.591	0.384	0.563	0.938	0.395	0.559	0.957
1500	-3.631	0.419	0.539	0.594	0.399	0.552	0.956	0.407	0.551	0.947

Note: Bias is the mean of the estimators of θ_0 minus θ_0 ($\times 10^4$); LHM is the mean of the left endpoint of the 95% confidence interval; RHM is the mean of the right endpoint of 95% confidence interval; CP is the nominal coverage probability of 95% confidence intervals.

Table 2. *Summary of CA19-9 results for pancreatic cancer data*

z	$G(z)$	Wald CI	Subsampling CI	LRT-based CI
1.5	0.233	(0.125,0.342)	(0.097,0.369)	(0.126,0.360)
2.0	0.233	(0.132,0.334)	(0.072,0.395)	(0.132,0.366)
2.5	0.233	(0.067,0.599)	(-0.034,0.501)	(0.132,0.503)
3.0	0.333	(0.174,0.492)	(0.101,0.566)	(0.178,0.552)
3.5	0.364	(0.330,1.000)	(-0.008,0.735)	(0.214,0.863)
4.0	0.714	(0.556,0.873)	(0.481,0.947)	(0.433,0.883)
4.5	0.778	(0.600,0.955)	(0.583,0.972)	(0.573,0.955)
5.0	1.000	(0.680,1.000)	(0.670,1.330)	(0.665,1.000)

Table 3. *Summary of CA125 results for pancreatic cancer data*

z	$G(z)$	Wald CI	Subsampling CI	LRT-based CI
2.0	0.250	(0.069,0.431)	(0.048,0.452)	(0.099,0.459)
2.3	0.350	(0.250,0.450)	(0.167,0.533)	(0.202,0.584)
2.5	0.571	(0.285,0.857)	(0.320,0.833)	(0.293,0.792)
3.0	0.778	(0.674,0.882)	(0.631,0.925)	(0.594,0.864)
3.5	0.794	(0.742,0.845)	(0.658,0.931)	(0.680,0.876)
4.0	0.794	(0.709,0.879)	(0.690,0.898)	(0.690,0.905)
5.0	0.800	(0.593,1.007)	(0.683,0.917)	(0.707,0.957)
6.0	1.000	(1.000,1.000)	(0.848,1.152)	(0.736,1.000)

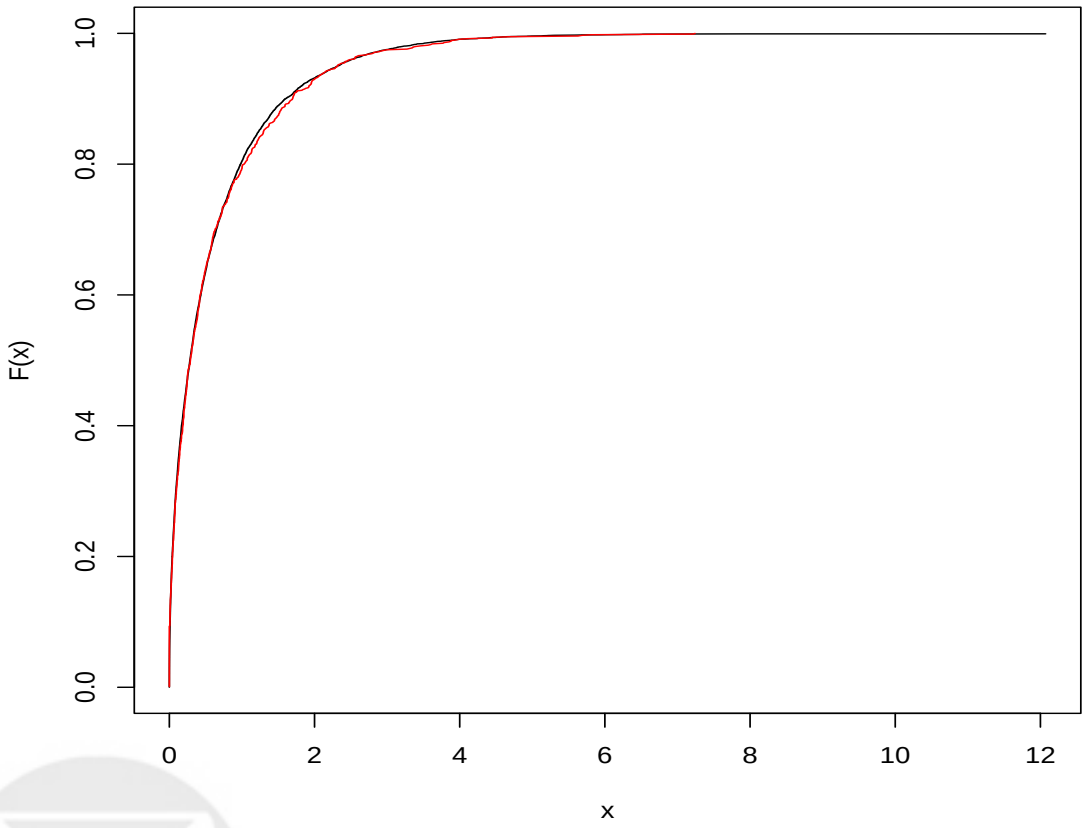

Figure 1: Empirical distribution of D (black line) relative to theoretical limit (red line)

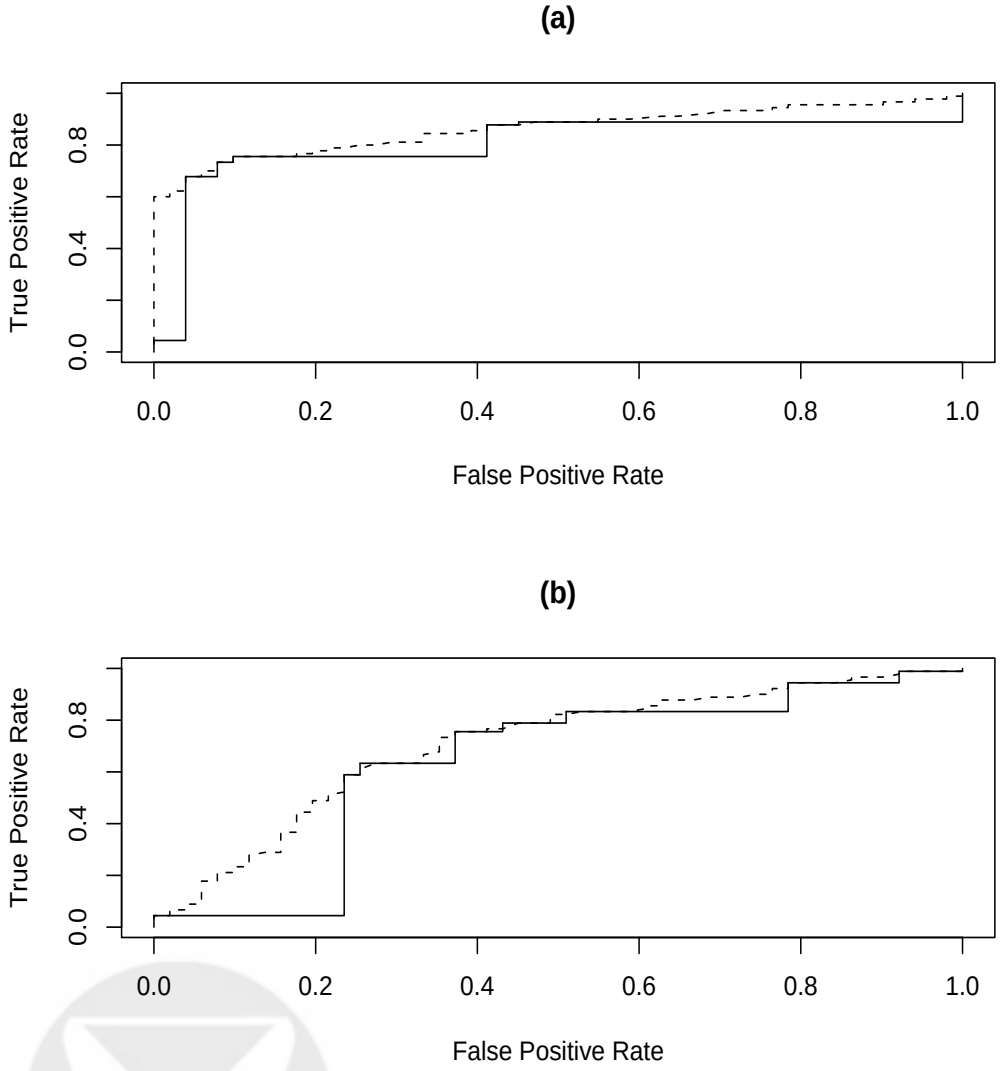

Figure 2: ROC curve estimated using pool-adjacent violators algorithm (solid line) and logistic regression (dotted line) for pancreatic cancer biomarkers: (a) CA19-9; (b) CA125.

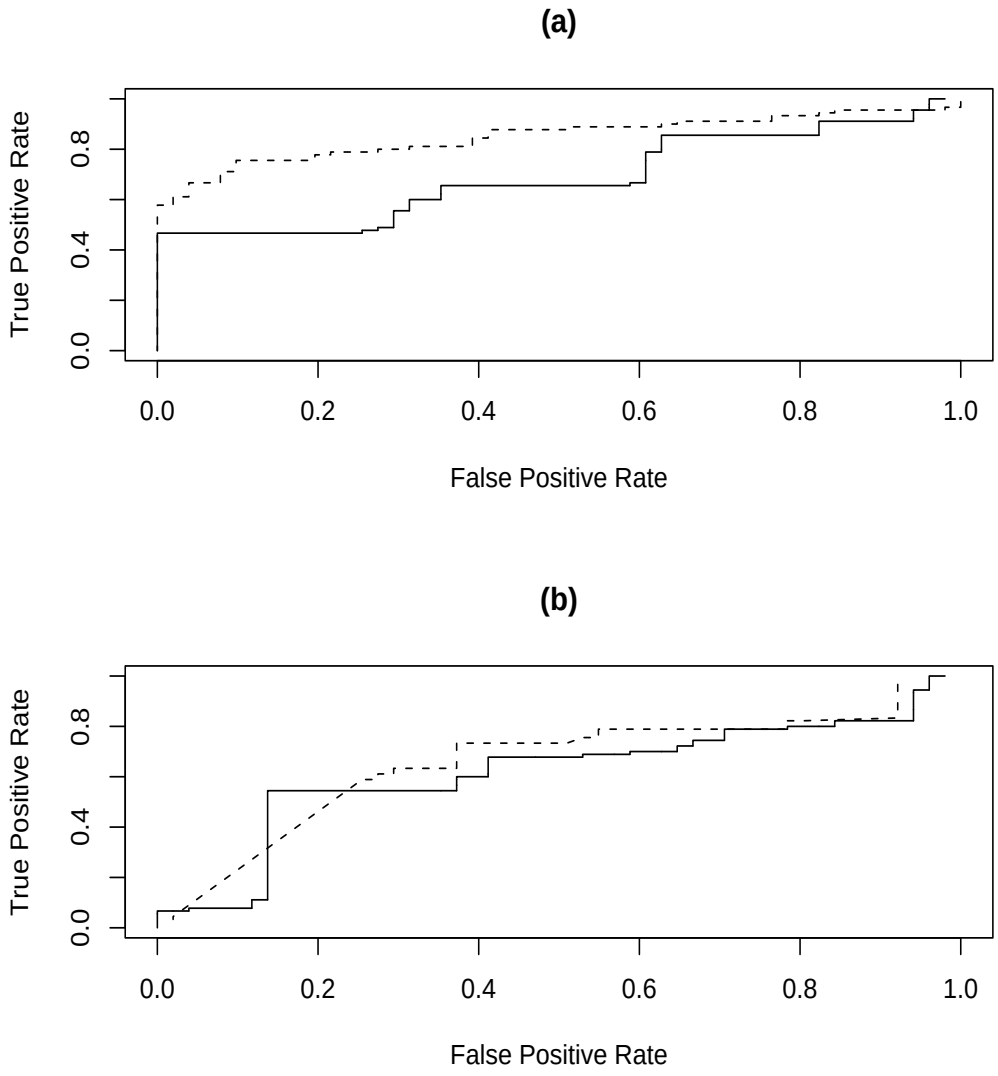

Figure 3: Leave-one-out cross-validated ROC curve estimated using pool-adjacent violators algorithm (solid line) and logistic regression (dotted line) for pancreatic cancer biomarkers: (a) CA19-9; (b) CA125.