Statistics and Its Interface Volume 3 (2010) 445–453

Spline-based models for predictiveness curves and surfaces

Debashis Ghosh* and Michael Sabel

A biomarker is defined to be a biological characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. The use of biomarkers in cancer has been advocated for a variety of purposes, which include use as surrogate endpoints, early detection of disease, proxies for environmental exposure and risk prediction. We deal with the latter issue in this paper.

Several authors have proposed use of the predictiveness curve for assessing the capacity of a biomarker for risk prediction. For most situations, it is reasonable to assume monotonicity of the biomarker effects on disease risk. In this article, we propose the use of flexible modelling of the predictiveness curve and its bivariate analogue, the predictiveness surface, through the use of spline algorithms that incorporate the appropriate monotonicity constraints. Estimation proceeds through use of a two-step algorithm that represents the "smooth, then monotonize" approach. Subsampling procedures are used for inference. The methods are illustrated to data from a melanoma study.

Keywords and phrases: Active set algorithm, Isotonic regression, Nonregular asymptotics, Pool adjacent violators algorithm, Risk prediction, Thin-plate spline.

1. INTRODUCTION

There has been extensive work done on the development of methodology for diagnostic testing and screening. One primary scientific goal in this area is to determine the discriminatory power of a biomarker for detecting disease. 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.

With the difficulty in developing tests and finding biomarkers that can lead to early detection of aggressive disease, interest has more recently focused on risk prediction using biomarkers (Huang et al., 2007). The idea is that it might be more feasible to find one or more biomarkers that can stratify patients into risk subgroups that would lead to better clinical management of these patients. Rather than dealing with the distribution of biomarker measurements conditional on disease status as with the receiver operating characteristic (ROC) curve, the objective has been has been on studying the distribution of the risk scores themselves. It is this type of modelling that we focus on in this article.

In the case of one biomarker, Huang et al. (2007) have proposed a quantity termed the predictiveness curve which graphically displays the distribution of risk in the population, standardized to the baseline distribution of the biomarker. We more carefully define this quantity in Section 2. Huang et al. (2007) proposed flexibly parametric and nonparametric estimation and inference procedures for this quantity. One common assumption that is made in biomarker contexts is that risk of disease is a monotone function of level of the biomarker. This leads to simple clinical decision rules. For example, if a man's PSA is above 4 ng/mL, then he is predicted to have prostate cancer. However, it remains an open question as to how to extend monotonic estimation procedures to more than one biomarker. Two proposals in this direction are those of Mukarjee and Stern (1994) and Beran and Dümbgen (2009).

In this article, we propose two extensions for the modelling of the predictiveness curve. The first is to develop and compare flexible algorithms for estimation of covariate-adjusted predictiveness curves using smoothing splines (Ruppert et al., 2003). The second is to generalize the predictiveness curve to the bivariate setting. This leads to consideration of a predictiveness surface. While the predictiveness surface has been considered by Gilbert and Hudgens (2008) recently, they modelled the quantity in a causal inference setting in which they seeked to identify causal estimands for evaluating surrogacy. By contrast, we are interested in the use of the predictiveness curve for the same purposes as Huang et al. (2007); what we wish to explore is the use of constrained smoothing techniques for estimation. The structure of this article is as follows. In Section 2, we outline the data available and give a brief review on predictiveness curves. We then describe estimation and inference in a semiparametric model for the predictiveness curve using the approaches given in Ghosh (2007). In Section 3, we describe the bivariate analog of the predictiveness curve, termed the predictiveness surface. We then propose an estimator that is a bivariate generalization of the "smooth,

*Corresponding author.

then isotonize" procedure of Ghosh (2007). An illustration using data from a melanoma study is provided in Section 4. We conclude with some discussion in Section 5.

2. DATA AND MODEL SETUP

Let S_1 and S_2 denote two biomarkers, $\mathbf{Z}$ a p-dimensional vector of covariates and Y be an indicator of disease status (i.e., $Y = 1$ if subject has disease, $Y = 0$ otherwise). We observe the data $(Y_i, S_{1i}, S_{2i}, \mathbf{Z}_i)$, $i = 1, \ldots, n$, a random sample from $(Y, S_1, S_2, \mathbf{Z})$. Define $\mathbf{S} \equiv (S_1, S_2)$ and $\mathbf{S}_i \equiv (S_{i1}, S_{2i})$, $i = 1, \ldots, n$.

2.1 Predictiveness curves: A review

For the moment, assume that $\mathbf{S} \equiv S$ is one-dimensional. In this situation, Huang et al. (2007) define the predictiveness curve as a plot of $R(v)$ versus v, where v takes values in $(0, 1)$, and

$$R(v) = P[Y = 1|S = F^{-1}(v)],$$

where F is the cumulative distribution function (CDF) for Y. We will be assuming here and throughout that higher values of Y are associated with increased risk of disease. It might be the case that one would then have to take negative the value of Y to satisfy this convention, as was done by Huang et al. (2007) in the example corresponding to their Figure 1. The predictiveness curve describes the distribution of risk in the population. In particular, comparisons are made between the estimate of $R(v)$ with the estimated disease prevalence, $\rho = P(Y = 1)$. Note that by the law of iterated expections, the area under the predictiveness curve is $\int R(v)dv = \rho$. In the Appendix, we describe an alternative theory for the predictiveness curve based on the nonparametric maximum likelihood estimation (NPMLE) procedure.

We now briefly discuss the difference between the predictiveness curve presented here with the usual metrics of receiver operating characteristic curves, sensitivity and specificity that are typically used in diagnostic testing. The latter quantities can be estimated in both a cohort or a case-control setting. However, the distribution of the risk, or equivalently the risk scores, can only be estimated in a cohort setting. There is an analogy here with logistic regression. The estimators for the odds ratios associated with covariates do not depend on the study design when the distribution of the covariates is completely unconstrained (Prentice and Pyke, 1979). By contrast, for estimating the disease risk, one either needs data from a cohort study or external information on the prevalence.

Huang et al. (2007) proposed a simple-two step procedure for estimating the predictiveness curve. First, one estimates $P(Y = 1|S)$ using a regression model and then computes $\hat{R}(v) = P[Y = 1|\hat{F}^{-1}(v)]$, where $\hat{F}$ is the empirical distribution for S. They also suggested a covariate-specific predictiveness curve in the case where $\mathbf{Z}$ is discrete. This is done by computing the predictiveness curve stratified based on the values of $\mathbf{Z}$. In the case where $\mathbf{Z}$ contains both continuous and discrete components, however, such an approach is not feasible. For this situation, Huang et al. (2007) suggested using both logistic regression and Box-Cox-type transformation models for modelling $P(Y = 1|S, \mathbf{Z})$ and to then plug in $\hat{F}$ to obtain an estimate of the predictiveness curve. In their discussion, they left open the possibility of using more flexible estimation procedures such as B-splines. However, such an approach should also attempt to utilize the monotonicity asumption relating risk of disease to S. We now explore that issue.

2.2 A constrained smoothing spline algorithm

The first model we consider is the following:

$$\text{logit}\{P(Y = 1|S, \mathbf{Z})\} = g(S) + \alpha^T \mathbf{Z}, \tag{1}$$

where g is monotone increasing in S. Given estimates of g and α from the model, one would estimate the covariate-adjusted predictiveness curve by $\hat{R}(v|\mathbf{Z}) \equiv P(Y = 1|S = \hat{F}^{-1}(v), \mathbf{Z})$.

The first procedure for predictiveness curve estimation is an adaptation of the work of Ghosh (2007). The algorithm proceeds as follows:

1. At the first stage, estimate g and α jointly ignoring the constraint using the likelihood-based algorithm in Lin and Zhang (1999).
2. Based on the estimated g from step 1, project it onto the space of monotonic functions, using the pool adjacent violators algorithm as described by Robertson et al. (1988). To be specific, the algorithm finds $\mu \in C$ that minimizes
$$\sum_{l=1}^{r} (\hat{f}(S_l) - \mu_l)^2,$$
such that $f(S_{(1)}) \leq f(S_{(2)}) \leq \cdots \leq f(S_{(r)})$, where $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(r)}$ are the ordered distinct values of the biomarker.

Based on the estimates from the two-step algorithm, one can then estimate a covariate-adjusted version of the predictiveness curve as described above. The approach we are proposing is an extension of the method proposed for nonparametric models by Mammen et al. (2001). The algorithm is computationally quite feasible and can be fit using the spm and isoreg functions from R. An alternative estimation procedure would be to replace step 2 by a sorting step. While the advantage of the procedure is its conceptual simplicity, there is no objective function that the sorting optimizes. By contrast, isotonic regression optimizes the least squares problem given in the second step above.

For inferential purposes, there are two possible options. One is to use profile likelihood (Murphy and van der Vaart,

1997) for construction of confidence intervals. We note that the monotonic effect of g in the model will not converge at an $n^{1/2}$ rate (Banerjee and Wellner, 2001); this is an example of a nonregular estimation problem. An alternative approach, and one that we use here, is to employ subsampling techniques (Politis et al., 1999).

1. Sample without replacement $(S_1^*, \mathbf{Z}_1^*), \ldots, (S_{b_0}^*, \mathbf{Z}_{b_0}^*)$, from the controls and $(S_1^*, \mathbf{Z}_1^*), \ldots, (S_{b_1}^*, \mathbf{Z}_{b_1}^*)$ from the cases, where b_0 and b_1 are the numbers of the subsampled controls and cases in the dataset.
2. Perform the two-stage estimation procedure described above.
3. Repeat steps 1. and 2. several times.

Further discussion of the inference for this type of model can be found in Banerjee et al. (2006).

3. BIVARIATE EXTENSION

We now consider the situation where there is more than one biomarker. Take $\beta = 0$ in (3). We define the predictiveness surface $\mathbf{R}(v_1, v_2)$ as

$$(2) \qquad R(v_1, v_2) = P[Y = 1 | S_1 = F_1^{-1}(v_1), S_2 = F_2^{-1}(v_1)],$$

where F_1 and F_2 are the marginal distribution functions for S_1 and S_2. This quantity was originally proposed by Gilbert and Hudgens (2008) and termed a predictiveness surface. While they proposed it in the context of identifying and performing inference for causal estimands in a surrogacy problem, our goal is to use the surface to better understand the behavior of biomarkers for risk prediction purposes. In addition, there might be covariates we want to adjust for. This will lead to a covariate-adjusted extension of the predictiveness surface:

$$R(v_1, v_2 | \mathbf{Z}) = P[Y = 1 | S_1 = F_1^{-1}(v_1), S_2 = F_2^{-1}(v_1), \mathbf{Z}].$$

This quantity was not considered by Gilbert and Hudgens (2008) in their work.

We formulate the following class of regression models:

$$(3) \qquad \text{logit}\{P(Y = 1 | \mathbf{S}, \mathbf{Z})\} = f(\mathbf{S}) + \beta^T \mathbf{Z},$$

where β is a p-dimensional vector of unknown regression coefficients to be estimated, and f is an unspecified bivariate monotone function. Note that in the absence of $\mathbf{Z}$, (3) reduces to a nonparametric model.

The algorithm of Ghosh (2007) has an obvious bivariate extension. The algorithm is to perform estimation of (3) in an unconstrained manner and then to constrain the resulting unconstrained estimate of f to satisfy the monotonicity condition. In the absence of the monotonicity constraint, a natural method of estimation in (3) is with thin-plate splines (Green and Silverman, 1994, Ch. 7). This can be formulated as maximizing the following penalized log-likelihood:

$$(4) \qquad \sum_{i=1}^{n} y_i \left\{ f(\mathbf{S}_i) + \beta^T \mathbf{Z}_i \right\} + \log(1 - p_i) \\ - \frac{1}{2} \lambda \int_{a_1}^{b_1} \int_{a_2}^{b_2} \left\{ \left(\frac{\partial^2 f}{\partial s_1^2} \right)^2 + \left(\frac{\partial^2 f}{\partial s_1 \partial s_2} \right)^2 \\ + \left(\frac{\partial^2 f}{\partial s_2^2} \right)^2 \right\} ds_1 ds_2,$$

where $\lambda > 0$ is a smoothing parameter, $p_i = P(Y_i = 1 | \mathbf{S}_i, \mathbf{Z}_i)$. We take (a_1, b_1) and (a_2, b_2) to define the range for the observed values of S_1 and S_2.

We seek to utilize the equivalence between the thin-plate spline estimation problem (4) and the mixed model framework. However, there is not a direct correspondence between the two because the derivative constraint cannot be reformulated as a proper covariance function. To do so, we will utilize an approximate bivariate smoothing procedure, which is an extension of the procedure described in Section 13.5 of Ruppert et al. (2003) to accommodate non-continuous outcomes. The algorithm works as follows:

1. Determine the number of knots, M, by
$$M = \max\{20, \min(n/4, 150)\}.$$
2. Use the space filling algorithm (Nychka et al., 1998), applied to $\mathbf{S}_1, \ldots, \mathbf{S}_n$ to obtain the knots $\kappa_1, \ldots, \kappa_M \in R^2$.
3. Create the matrices $\mathbf{W} = [\mathbf{1} \;\; \mathbf{Z}_i \;\; \mathbf{S}_i]_{1 \le i \le n}$,
$$\mathbf{X}_K = [\|\mathbf{S}_i - \kappa_k\|^2 \log \|\mathbf{S}_i - \kappa_k\|]_{1 \le i \le n, 1 \le k \le M},$$
and
$$\Omega = [\|\kappa_k - \kappa_l\|^2 \log \|\kappa_k - \kappa_l\|]_{1 \le k, l \le M}.$$
4. Take the singular value decomposition of $\Omega = \mathbf{UAV}^T$, and compute $\Omega^{1/2} = \mathbf{UA}^{1/2}\mathbf{V}^T$.
5. Compute $\mathbf{X} = \mathbf{X}_K \Omega^{-1/2}$.
6. Use mixed model software to fit the generalized linear mixed model for logit(p) where $\mathbf{W}$ is the design matrix for the fixed effects and $\mathbf{X}$ is the design matrix for the random effects. The random effects are distributed as $N(\mathbf{0}, \sigma_u^2 \mathbf{I})$.

There are several points to note here. First, the estimation procedure is similar in spirit to weighted least squares estimation procedures. Second, we are representing f using both fixed and random effects. Based on the resulting solutions for the estimated fixed and random effects, we can estimate the parametric and nonparametric components of (3). Note that the smoothing parameter λ can be estimated automatically using restricted maximum likelihood (Lin and

Zhang, 1999). The procedure exploits the fact that λ is related to the variance of the random effects as $\lambda^{-1} = \sigma^2$. Another issue is the impact of the number of knots in step 1 and their locations in step 2 of the algorithm. An excellent discussion of how to select them can be found in Chapter 17 of Ruppert et al. (2003). They argue for a sufficiently large number of knots; they find that while changes in this quantity can have an effect on the estimate of f when M is small, it is less for larger values of M. The effect of the location of the knots on the estimate of f appears to still be a question that has no satisfactory resolution within this framework. As noted in Section 17.3 of Ruppert et al. (2003), one could perform an optimization based on a grid search of values for M and locations of the M knots, but this is currently computationally prohibitive.

We next describe constraining the unconstrained estimate of f to satisfy the monotonicity constraint. The sandwich isotonic block class (SIBC) algorithm of Qian and Eddy (1996) is applied to the unconstrained estimator of f, $\hat{f}$, described in the previous paragraph. The SIBC algorithm is a modification of the algorithm of Dykstra and Robertson (1982). Let $(S^*_{11}, \ldots, S^*_{1r})$ and $(S^*_{21}, \ldots, S^*_{2m})$ denote the r and m unique ordered values of S_1 and S_2. Define $\mathcal{H} = \{(S^*_{1i}, S^*_{2j}) : i = 1, \ldots, r; j = 1, \ldots, m\}$. Define $E(s,t) = \{(S^*_{1i}, S^*_{2j}) : i = 1, 2, \ldots, s, j = t, t+1, \ldots, m\}$. We define $g^*(\cdot,\cdot)$, the isotonic regression is defined as the solution to the following minimization problem:

$$\min_g \sum_{u,v} \{g(u,v) - \hat{f}(u,v)\}^2 w(u,v)$$

subject to $g(\cdot,\cdot)$ isotonic in both variables, where $w(u,v)$ is a weight function. Define $g^*_{s,t}(\cdot,\cdot)$ to be the isotonic regression on $E(s,t)$. Let U_x and L_x be the expanded upper and expanded lower sets of $g^*_{s,t+1}$:

$$U_x = \{(i,j) \in E(s,t) : i = s \text{ or } g^*_{s-1,t}(i,j) \geq x\};$$
$$L_x = \{(i,j) \in E(s,t) : j = t \text{ or } g^*_{s,t+1}(i,j) \leq x\}.$$

The SIBC algorithm works as follows:

1. Use isotonic regression to find $g^*_{1,t}(\cdot,\cdot)$ on $E(1,t)$, $t = m, m-1, \ldots, 1$.
2. For $i = 2, \ldots, r$,
   (a) Compute $g^*_{i,m}(\cdot,\cdot)$, the isotonic regression of $E(i,m)$;
   (b) For $j = m, m-1, \ldots, 1$
      i. $c_0 = \text{median}\{\hat{f}(i,j), g^*_{i-1,j}(i,j), g^*_{i,j+1}(i,j+1)\}$.
      ii. For $k = 0, 1, 2, \ldots$ compute $c_{k+1} = \max\{h(x_k) : g^*_{i-1,j}(s,t) < x_k\}$, where $h(x)$ is the average value of g^* on the intersection of U_x and L_x;
      iii. Define g^* by

$$g^*(s,t) = \begin{cases} g^*_{i-1,j}(s,t) \text{ if } g^*_{i-1,j}(s,t) < x; \\ x \text{ if } x \text{ is in } U_x \cap L_x; \\ g^*_{i,j+1}(s,t) \text{ if } g^*_{i,j+1}(s,t) > x. \end{cases}$$

While the notation for the algorithm may seem cumbersome, in fact the crucial steps involve the univariate pool adjacent violators algorithm (Robertson et al., 1988) and taking the median. There are several advantages of the SIBC algorithm relative to other bivariate isotonic regression algorithms. First, the algorithm involves univariate isotonic regressions. Second, the algorithm is guaranteed to converge in a finite number of iterations. Finally, the algorithm is computationally quite fast. A comparison of the SIBC procedure with the algorithms of Moonesinghe and Wright (1994) and Block et al. (1994) through simulation studies revealed the SIBC algorithm to be at least five times faster.

To summarize the two-step algorithm, at the first stage, we fit the unconstrained estimate given using the mixed model framework and at the second stage, we compute the bivariate isotonic estimator using the algorithm of Qian and Eddy (1996).

As before, the results of Banerjee and Wellner (2001) suggest that estimation of the nonparametric component will not converge asymptotically at the usual rate. We again use subsampling to perform inference. The algorithm is a slight modification of that presented in §2.2. The procedure yields an empirical distribution for f and β from which we can construct confidence intervals.

4. NUMERICAL EXAMPLES

4.1 Melanoma data

The data analyzed here come from a prospective database maintained by the second author. They are on patients with cutaneous melanoma who underwent sentinel lymph node (SLN) biopsy at the University of Michigan during a period from August 1997 to March 2004. There were two exclusion criteria. The first was that patients who were aged less than 16 years at the time of surgery were excluded because it is believed that the biology of pediatric melanomas is quite different from that of adult melanomas. Second, patients with multiple primary melanomas that went to the same lymph node basin were excluded from the dataset. It is thought that these patients will have a greater likelihood of being SLN-positive.

A variety of demographic and clinical covariates were collected. For illustration, we use the following variables: patient age at biopsy and gender, Breslow depth and mitotic rate. The dependent variable is SLN-positivity (0=SLN-negative; 1=SLN-positive). Previous analyses have focused on modelling the effects of mitotic rate and Breslow depth parametrically (Sondak et al., 2004). Preliminary descriptive analyses suggested the following transformations:

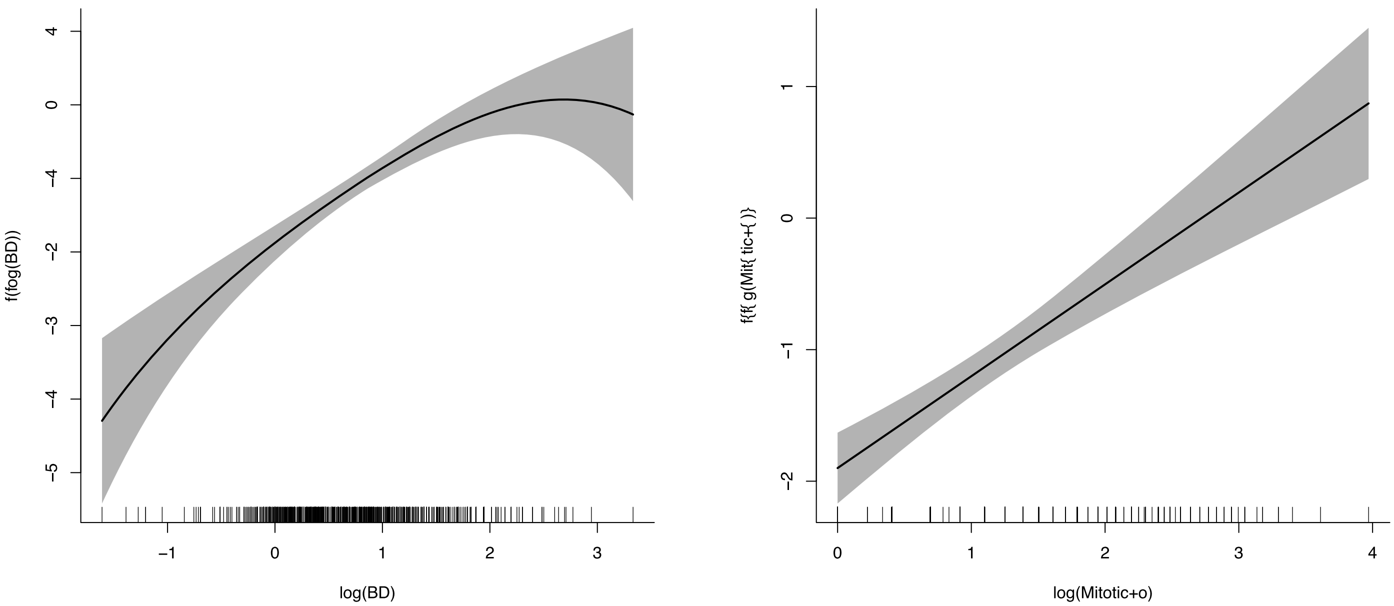

Figure 1. Left-hand figure: Plot of the effect of Breslow depth (on a natural log scale) on risk of SLN-positivity. Right-hand figure: Plot of the effect of mitotic rate (on a natural log scale after adding one for handling zero values) on risk of SLN-positivity. The tick marks denote the observed data points. While the solid line denotes the fitted value from the model, the shaded area denotes the 95% pointwise confidence intervals for the estimated function. Note that Breslow depth and mitotic rate have been transformed using the natural logarithm function.

1. The Breslow depth and mitotic rate were log-transformed in order to reduce skewness. In addition, one was added to the mitotic rate before taking the logarithmic transformation.
2. Age at biopsy had a negative association with risk of positive sentinel lymph nodes. Thus, the negative of age was used in the nonparametric modelling of the predictiveness curve and surface.

We begin by showing the plots of fits for disease risk as a function of Breslow depth and mitotic rate from step 1 of the algorithm of Ghosh (2007). These are given in Figure 1. Based on the plots, we see that there while risk of sentinel lymph node-positivity is increasing as a function of Breslow depth up to values of 2.5, the effect of mitotic rate is strongly monotonic. Thus, step 2 in the algorithm of Ghosh (2007) would not change the result for mitotic rate but would monotonize the tail region for Breslow depth.

One could calculate these curves separately for men and women. Qualitatively, the stratified regression curves and corresponding predictiveness curves are not different from those in the entire population (data not shown).

Next, we considered using age as a covariate to adjust for in the predictiveness curve calculations. Unlike Huang et al. (2007), we use a more flexible model (1) in which the effect of age is modelled parametrically and we model Breslow depth and mitotic rate nonparametrically. Note that this involves fitting two models, one with Breslow depth in the nonparametric component, the other with mitotic rate. The association between age and risk of SLN-positivity is negative adjusting for Breslow depth ($\hat{\alpha} = -0.02, 95\% CI = (-0.028, -0.015)$) and mitotic rate ($\hat{\alpha} = -0.02, 95\% CI = (-0.031, -0.012)$)

We now consider modelling the effects of Breslow depth and mitotic rate jointly. We fit a model in which the only term is a nonparametric function of Breslow depth and mitotic rate, again suitably transformed as previously described. We assume the function to be monotonic in both arguments and use the proposed methodology in the paper. The proposed algorithm is applied, and contours from the fitted function are presented in Figure 2. We also performed the same analysis using Breslow depth and age; this is also given in Figure 2. Comparing the two contour plots, we find that the monotonicity constraint makes the contours be straight lines, which is more easily interpretable than nonlinear contours. We also performed an analysis in which Breslow depth and mitotic rate where modelled nonparametrically but age was modelled parametrically. The covariate effect of age on risk of SLN-positivity and its associated 95% CI are approximately the same as in the previous analyses ($\hat{\alpha} = -0.025, 95\% CI = (-0.035, -0.015)$).

As before, it is of interest to determine if there a difference in these estimates by gender. We repeated the calculations for males and females separately. Again, the contours did not change much qualitatively.

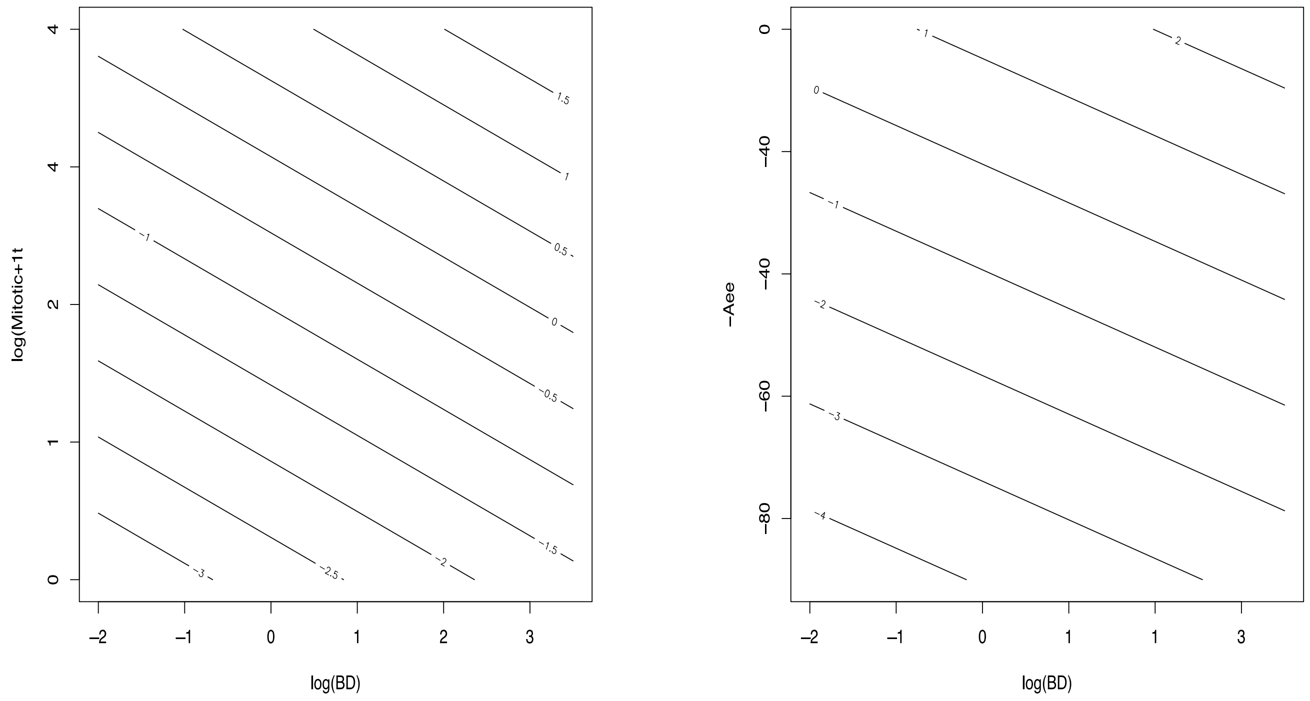

Figure 2. Contours for predictiveness curves for bivariate nonparametric monotonic models. Left-hand figure shows contours for the model with Breslow depth and mitotic rate. The right-hand figure shows contours for the model with Breslow depth and age. Recall that the negative value of age is being used so that disease risk is a monotonically increasing function of age. Also, Breslow depth and mitotic rate have been transformed as in Figure 1.

4.2 Simulation studies

We performed some limited simulation studies to assess the finite-sample properties of the proposed methodology. First, the situation with one-dimensional biomarker was considered, in which we compared the proposed method from Section 2 to the method of Huang et al. (2007). Data were generated from the simulation model presented in their Table 1 using the following model:

$$Pr(D=1|S) = \Phi(-0.486 + 0.793S^{0.4}),$$

where Φ denotes the standard normal cumulative distribution. We considered sample sizes of $n = 100$ and $n = 2,000$; 5,000 simulation samples were generated for each scenario. The subsampling was done using 50% of the samples. The results are summarized in Table 1. We find that the proposed procedure has better finite-sample properties for smaller sample sizes. For larger sample sizes, the proposed method gives similar answers to the method of Huang et al. (2007). Intuitively, this seems reasonable, as the constraint has less effect for larger sample sizes; equivalently, the difference between the unconstrained and constrained estimators goes to zero as the sample size gets larger. We also performed simulation studies about the predictiveness surface in the case of two-dimensional biomarkers. While no analog of the Huang et al. (2007) method exists for two dimensions, the proposed procedures from Section 3 exhibited satisfactory finite-sample performance (data not shown).

5. DISCUSSION

In this article, we have developed some computationally feasible spline-based algorithms for estimation of the predictiveness curve, whose use has been recently advocated for risk prediction by Huang et al. (2007). They left open the problem of how to model the quantity using spline-based procedures. We have developed procedures for both the predictiveness curve and surface that employ splines in conjunction with isotonic regression-based adjustments. As seen in the example from Section 4, incorporating the monotonicity constraint leads to very interpretable predictivness curve estimates.

While the asymptotic properties of the proposed methods were not studied, we make two observations. First, we would expect the asymptotics to follow those of Banerjee and Wellner (2001), which implies that the convergence of the profile likelihood ratio statistic to be nonregular. Second, our intuition suggests that imposing a monotonicity constraint matters more in small samples than in large samples. As the sample approaches infinity, most regular nonparametric estimation procedures will be able to identify

Table 1. Simulation results for the predictiveness case for the case of one-dimensional biomarker. Huang refers to method of Huang et al. (2007); results are taken from their Table 1. Proposed refers to methodology in Section 2.

	$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$	0.100	0.194	0.313	0.491	0.800
Bias					
% bias in $\hat{R}(v)$					
$n = 100$, Huang	−1.636	−3.058	−0.969	−0.493	−0.749
$n = 100$, Proposed	−0.981	−1.523	−0.858	−0.232	−0.613
$n = 2000$, Huang	−0.279	−0.240	−0.152	−0.095	−0.033
$n = 2000$, Proposed	−0.275	−0.238	−0.154	−0.097	−0.031
95% CI Coverage Probabilities					
$n = 100$, Huang	86.53	92.09	92.91	92.87	89.99
$n = 100$, Proposed	96.10	94.38	95.01	94.78	94.65
$n = 2000$, Huang	94.24	94.77	94.57	94.39	94.25
$n = 2000$, Proposed	94.45	94.51	94.79	95.16	94.58

the correct relationship. If the true relationship is monotonic, then the probability of any nonparametric estimator satisfying the constraint will have probability approaching one.

While a bivariate version of the predictiveness curve has been studied here, in practice there will be situations in which more than two biomarkers will be considered. For this situation, it would be desirable to have computationally feasible procedures that can simultaneously incorporate the nonparametric and monotonicity simultaneously. This is currently under investigation.

R scripts implementing the proposed methodology are available from the first author upon request.

ACKNOWLEDGEMENTS

The first author would like to acknowledge the support of NIH Grant R01 CA129402.

APPENDIX

Nonparametric estimation of predictiveness curve for univariate biomarkers

To keep ideas concrete, we will assume that there are no covariates $\mathbf{Z}$. We again consider data on (D_i, S_i), $i = 1, \ldots, n$. The log-likelihood for the sample is given by

$$l(F) = \sum_{i=1}^{n} d_i \log G(s_i) + (1 - d_i) \log\{1 - G(s_i)\} + \log f(s_i) \tag{A.1}$$

where $G(s) = P(D = 1|S)$ and f denotes the density of S. We now consider nonparametric maximization of (A.1). Note that because of the contstraint on G being monotone increasing, the theory from Prentice and Pyke (1979) is not directly applicable here.

A precise characterization of the maximizer $\hat{G}$ in this situation is found in Groenenboom and Wellner (1992, pp. 38–40). Let $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(n)}$ denote the observed order statistics for $(S_1, \ldots, S_n)$, and let $s_{(i)}$ $(i = 1, \ldots, n)$ denote the corresponding value of d. Define $s_{(0)} = 0$ and $d_{(0)} = 0$. The nonparametric maximum likelihood estimator (NPMLE) of G corresponds to the point $\tilde{x} \equiv (\tilde{x}_1, \ldots, \tilde{x}_n)$ that maximizes

$$h(x_1, \ldots, x_n) = \sum_{i=1}^{n} \{s_{(i)} \log x_i + (1 - s_{(i)}) \log(1 - x_i)\}$$

over $(x_1, \ldots, x_n) \in R^n$ subject to the constraint

$$0 \leq x_1 \leq \cdots \leq x_n \leq 1.$$

We derive the NPMLE of G, $\hat{G}_*$, through the relationship $\tilde{x}_i = \hat{G}_*(s_{(i)})$, $i = 0, \ldots, n$. Note that the NPMLE of G is defined only up to the set of observed times. The solution to this optimization problem can be characterized in one of two ways. The first is using the so-called "max-min formula" (Groenenboom and Wellner, 1992, p. 40):

$$\tilde{x}_m = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{i \leq j \leq k} d_{(j)}}{k - i + 1},$$

$m = 0, \ldots, n$. A second representation of the maximizer is more graphical in nature. One plots the points $\{i, \sum_{j \leq i} d_{(j)}\}$ $(i = 0, \ldots, n)$ and draws the greatest convex minorant of these points, defined as the function H^* such that

$$H^*(t) = \sup\left\{H(t) : H(i) \leq \sum_{j \leq i} d_{(j)}, H(0) = 0, H \text{ is convex}\right\}.$$

Then $\tilde{x}_i$ is the left derivative of H^* at $i = 0, \ldots, n$. Based on the estimator $\hat{G}$, the predictiveness curve at v is given by

$$\hat{R}(v) = \hat{G}_*(\tilde{F}^{-1}(v)),$$

where $\tilde{F}$ is the empirical estimator of the quantile function for S. Using the arguments in Chapter 5 of Groenenboom and Wellner (1992), we can prove the following asymptotic result:

Lemma 1. *Let z_0 be such that $0 < F_S(z_0) < 1$ and $0 < G(z_0) < 1$. Assume that F and G are differentiable at z_0 with strictly positive derivatives $f(z_0)$ and $g(z_0)$, respectively. Then $n^{1/3}\{\hat{G}_*(z_0) - G(z_0)\}$ converges in distribution to the random variable $C\mathcal{Z}$, where*

$$C = \left[\frac{4G(z_0)\{1-G(z_0)\}g(z_0)}{f(z_0)}\right]^{1/3}$$

and $\mathcal{Z} \equiv argmin\{W(t) + t^2\}$, and W is two-sided Brownian motion starting from zero.

Remark 1. The result presented in Lemma 1 differs considerably from those in Huang et al. (2007). In particular, they derive asymptotic normality results. By contrast, we are completely nonparametric except for the monotonicity constraint. This type of result is very common for isotonic regression estimators of the sort presented above.

Remark 2. An alternative approach would be to treat the values of G at the unique values of S as *a priori* parameters. In this case, the problem would reduce to a finite-dimensional model, and the resulting estimator would have a limiting distribution that is a mixture of chi-squared random variables (Sen and Silvapulle, 2002) and is different from the limiting distribution in Lemma 1. However, for this situation, the number of parameters in the model would depend on the particular dataset, which would appear to a very undesirable feature of the approach.

Remark 3. The result of Lemma 1 implies that $n^{1/3}(\hat{R}(v) - R(v))$ will also have the same form for the limiting distribution as $n^{1/3}(\hat{G}(s) - G(s))$. This is given heuristically by the following argument:

$$\begin{aligned} n^{1/3}\{\hat{R}(v) - R(v)\} &= n^{1/3}[\hat{G}\{\tilde{F}^{-1}(v)\} - G\{F^{-1}(v)\}] \\ &\approx n^{1/3}[\hat{G}\{F^{-1}(v)\} - G\{F^{-1}(v)\}] \\ &\to G'(F^{-1}(v))C\mathcal{Z}, \end{aligned}$$

where we assume that sufficient regularity conditions apply for $\hat{F}^{-1}$ to converge uniformly to F^{-1}. The above argument uses the delta-method.

Received 11 February 2010

REFERENCES

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Annals of Statistics* **29**, 1699–1731. MR1891743

Banerjee, M., Biswas, P. and Ghosh, D. (2006). Semiparametric binary regression under monotonicity constraints. *Scandinavian Journal of Statistics* **33**, 673–697. MR2300910

Beran, R. J. and Dümbgen, L. (2009). Least Squares and Shrinkage Estimation under Bimonotonicity Constraints. *Statistics and Computing*. Available at 10.1007/s11222-009-9124-0.

Block, H., Qian, S. and Sampson, A. (1994). Structure algorithms for partially ordered isotonic regression. *Journal of Computational and Graphical Statistics* **3**, 285–300. MR1292119

Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Annals of Statistics* **10**, 708–711. MR0663427

Ghosh, D. (2007). Incorporating monotonicity into the evaluation of a biomarker, *Biostatistics* **8**, 402–413.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**: 1146–1154.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. London: Chapman and Hall. MR1270012

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser: Boston. MR1180321

Huang, Y., Pepe, M. S. and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–1188. MR2414596

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B* **61**, 381–400. MR1680318

Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statistical Science* **16**, 232–248. MR1874153

Moonesinghe, R. and Wright, F. T. (1994). Likelihood ratio tests involving a bivariate trend in two-factor designs: the level probabilities. *Communications in Statistics – Computation and Simulation* **23**, 143–156.

Mukarjee, H. and Stern, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *JASA* **89**, 77–80. MR1266288

Murphy, S.A. and van der Vaart, A.W. (1997). Semiparametric likelihood ratio inference. *Annals of Statistics* **25**, 1471–1509. MR1463562

Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell, M. (1996). FUNFITS data analysis and statistical tools for estimating functions. In *Case Studies in Environmental Statistics* (Ed: D. Nychka, W. W. Piegorsch and L. H. Cox), Springer-Verlag, New York, pp. 159–179.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411. MR0556730

Qian, S. and Eddy, W. F. (1996). An algorithm for isotonic regression on ordered rectangular grids. *Journal of Computational and Graphical Statistics* **5**, 225–235. MR1411315

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: Wiley. MR0961262

Ruppert, D. Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press: Cambridge. MR1998720

Sen, P. K. and Silvapulle, M. J. (2002). An appraisal of some aspects of statistical inference under inequality constraints *Journal of Statistical Planning and Inference* **107**, 3–43. MR1927753

Sondak, V. K., Taylor, J. M., Sabel, M. S., Wang, Y., Lowe, L., Grover, A. C., Chang, A. E., Yahanda, A. M., Moon, J. and Johnson, T. M. (2004). Mitotic rate and younger age are predictors of sentinel lymph node positivity: lessons learned from the generation of a probabilistic model. *Annals of Surgery* **11**, 247–258.

Speed, T. (1991). Discussion to "BLUP is a good thing: The estimation of random effects" by Robinson, G. K., *Statistical Sciences* **6**, 50–51. MR1108815

Debashis Ghosh
Department of Statistics, Penn State University
University Park, PA 16802
USA
E-mail address: ghoshd@psu.edu

Michael Sabel
Department of Surgery, University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: msabel@umich.edu