

Receiver operating characteristic curves and confidence bands for support vector machines

Daniel J. Lockett¹  | Eric B. Laber²  | Samer S. El-Kamary³ | Cheng Fan⁴ | Ravi Jhaveri⁵ | Charles M. Perou⁴ | Fatma M. Shebl⁶ | Michael R. Kosorok¹ 

¹ Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina

² Department of Statistics, North Carolina State University, Raleigh, North Carolina

³ Department of Epidemiology and Public Health, University of Maryland, Baltimore, Maryland

⁴ Department of Genetics, University of North Carolina, Chapel Hill, North Carolina

⁵ Department of Pediatrics, University of North Carolina, Chapel Hill, North Carolina

⁶ Department of Epidemiology, Yale University, New Haven, Connecticut

Correspondence

Daniel J. Lockett, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.

Email: daniel.lockett.work@gmail.com

Funding information

Division of Mathematical Sciences, Grant/Award Numbers: DMS-1513579, DMS-1555141, DMS-1557733; V Foundation for Cancer Research; National Cancer Institute, Grant/Award Number: P50-CA58223-09A1; Merck; Breast Cancer Research Foundation; National Institutes of Health, Grant/Award Numbers: 1R01DE024984, P01 CA142538, T32 CA201159, U01AI58372, U01HD39164; National Center for Advancing Translational Sciences, Grant/Award Number: UL1 TR001111

Abstract

Many problems that appear in biomedical decision-making, such as diagnosing disease and predicting response to treatment, can be expressed as binary classification problems. The support vector machine (SVM) is a popular classification technique that is robust to model misspecification and effectively handles high-dimensional data. The relative costs of false positives and false negatives can vary across application domains. The receiving operating characteristic (ROC) curve provides a visual representation of the trade-off between these two types of errors. Because the SVM does not produce a predicted probability, an ROC curve cannot be constructed in the traditional way of thresholding a predicted probability. However, a sequence of weighted SVMs can be used to construct an ROC curve. Although ROC curves constructed using weighted SVMs have great potential for allowing ROC curves analyses that cannot be done by thresholding predicted probabilities, their theoretical properties have heretofore been underdeveloped. We propose a method for constructing confidence bands for the SVM ROC curve and provide the theoretical justification for the SVM ROC curve by showing that the risk function of the estimated decision rule is uniformly consistent across the weight parameter. We demonstrate the proposed confidence band method using simulation studies. We present a predictive model for treatment response in breast cancer as an illustrative example.

KEYWORDS

classification, diagnostic medicine, machine learning, outcome weighted learning

1 | INTRODUCTION

Many important problems in biomedical decision-making can be expressed as binary classification problems. For

example, one may wish to identify infants infected with hepatitis C virus (HCV) from a sample of infants born to infected mothers (Shebl *et al.*, 2009), screen for prostate cancer using prostate-specific antigen (Etzioni *et al.*, 1999),

or predict which breast cancer patients will respond to treatment based on genetic characteristics (Fan *et al.*, 2011). In numerous applications, the costs of false positives and false negatives may differ, and classification methods must allow for unequal weighting of these errors. Classification problems are often based on high-dimensional data or data with complex structure, where it may be difficult to pose a parametric model for the conditional mean response. In such applications, it is important to have classification methods that allow for properly weighting false positives and false negatives within a framework that does not require restrictive modeling assumptions.

Understanding the accuracy of a classifier is important if a classifier is to be used to inform decisions in practice. Receiver operating characteristic (ROC) curve analyses are often used to assess and compare the accuracy of classifiers for a range of sensitivity and specificity values. (Zhou *et al.*, 2002; Pepe, 2003). Various methods for modeling and estimating ROC curves have been proposed, including parametric regression models (Pepe, 1997; McIntosh and Pepe, 2002) and semiparametric regression models (Pepe, 2000; Cai *et al.*, 2002; Cai and Dodd, 2008). Quantifying the uncertainty around the performance of a classifier is also necessary if a classifier is to be used in practice. Constructing confidence bands for the ROC curve is one way to do this. Methods for ROC curve confidence bands can involve estimating the biomarker distributions in the diseased and nondiseased samples using parametric models (Ma and Hall, 1993) or kernel density estimators (Jensen *et al.*, 2000; Claeskens *et al.*, 2003; Horváth *et al.*, 2008), or using empirical distribution functions in combination with the bootstrap (Campbell, 1994).

Machine learning techniques that output a continuous score or predicted probability allow for straightforward application of ROC curve methodology (see, eg, Spackman, 1989; Bradley, 1997; Provost and Fawcett, 1997, 1998). However, there are fewer examples of applying ROC curve methodology to classifiers that output only a class label, such as the support vector machine (SVM; Cortes and Vapnik, 1995). Platt (1999) proposed a method to extract class probabilities from the output of the SVM (see also Vapnik, 1998; Lin *et al.*, 2007) by fitting parametric models to the SVM class labels. Veropoulos *et al.* (1999) proposed fitting a sequence of weighted SVMs to construct an SVM ROC curve (see also Krzyżak *et al.*, 1996; Lin, 2002; Zhang, 2004; Steinwart and Christmann, 2008). However, there are no previously proposed methods for constructing a confidence band for the SVM ROC curve. The aforementioned confidence band methods assume a scalar biomarker and thus cannot be directly applied to the SVM ROC curve. Furthermore, the theoretical properties of the SVM ROC curve have heretofore been underdeveloped. We build on the work of Veropoulos *et al.*

(1999) by establishing a number of theoretical properties of the SVM ROC curve and developing a bootstrap method for constructing confidence bands for the SVM ROC curve.

There are numerous applications to motivate this work; however, we focus on one primary illustrative application, predicting which breast cancer patients will respond to treatment. Genomic data provide a wealth of information for this purpose. However, it is difficult to specify a parametric model for response given genomic features because of the high dimension of genomic data. Because the SVM provides nonparametric classification (Steinwart and Christmann, 2008), it is a natural choice for this problem. Furthermore, the costs of false positives and false negatives differ significantly in this application: a patient who is incorrectly predicted to respond may be subjected to unnecessary treatment, while a patient who is incorrectly predicted to not respond may not receive a potentially beneficial treatment. Thus, ROC curve analysis is needed in order to determine whether a predictive model for treatment response can be used in practice. A second illustrative example, the diagnosis of infant hepatitis C, is included in the Supporting Information.

There are a number of additional examples of weighted SVMs in the statistical literature. Shin *et al.* (2014) proposed fitting a sequence of weighted SVMs similar to the approach discussed in this paper; however, their goal was to construct estimates of class probabilities for the purposes of dimension reduction, rather than to estimate an ROC curve and construct confidence bands. Jiang *et al.* (2008) proposed a method to construct confidence intervals for the prediction error of the SVM; however, applying their approach in the current setting would result in pointwise confidence intervals for sensitivity and specificity, rather than uniform confidence bands for the ROC curve.

In Section 2, we briefly review the method developed by Veropoulos *et al.* (1999). In Section 3, we show a number of theoretical results. Section 4 presents our proposed confidence band method. In Section 5, we present a series of simulation experiments to evaluate the operating characteristics of the proposed bootstrap confidence bands. In Section 6, we present an illustrative case study. We conclude and discuss future research in Section 7. Proofs, additional simulation results, an additional illustrative example, and R code to implement the proposed method are provided in the Web-based Supporting Information.

2 | WEIGHTED SUPPORT VECTOR MACHINES

Veropoulos *et al.* (1999) proposed an approach to constructing an ROC curve using a sequence of weighted SVMs. Here, we provide a summary of this approach. Assume

that the available data are (A_i, \mathbf{X}_i) , $i = 1, \dots, n$, which comprise n independent and identically distributed copies of (A, \mathbf{X}) , where $A \in \{-1, 1\}$ is a class label (eg, in diagnostic medicine, $A = 1$ corresponds to a diseased individual and $A = -1$ corresponds to a nondiseased individual) and $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ are covariates. The goal is to estimate a classifier that correctly identifies a patient's class label based on that patient's covariates. Veropoulos *et al.* (1999) introduced the idea of using a sequence of weighted SVMs to control sensitivity and specificity in this context. Consider minimizing the expected weighted misclassification, where each misclassification event is weighted by the cost function $C_a(\alpha) = \{1 + (2\alpha - 1)a\}/2 = \alpha 1(a = 1) + (1 - \alpha)1(a = -1)$, where $C_a(\alpha)$ is the cost of misclassification when the true class label is $A = a$. In diagnostic medicine, with $A = 1$ corresponding to disease and $A = -1$ corresponding to nondisease, α determines the relative weight placed on the sensitivity and specificity of the test. When $\alpha = 1/2$, sensitivity and specificity are given equal weight as in the zero-one misclassification error. Let \mathcal{D} denote a class of functions from \mathcal{X} into $\{-1, 1\}$. The optimal classifier with respect to cost function $C_a(\alpha)$ within \mathcal{D} is

$$\tilde{D}_\alpha = \arg \min_{D \in \mathcal{D}} \mathbb{E}[1\{D(\mathbf{X}) \neq A\}C_A(\alpha)]. \quad (1)$$

For fixed $\alpha \in (0, 1)$ and a classifier D , the plug-in estimator of the weighted misclassification error is $\mathbb{E}_n 1\{D(\mathbf{X}) \neq A\}C_A(\alpha)$, where \mathbb{E}_n is the empirical measure of the observed data. Note that any classifier $D(\mathbf{X})$ can be represented as $\text{sign}\{f(\mathbf{X})\}$ for some decision function $f: \mathcal{X} \rightarrow \mathbb{R}$; we will assume that the decision function is smooth and thus f belongs to a class of smooth functions, \mathcal{F} . For example, we can let \mathcal{F} be the space of linear functions, the space of polynomial functions, or the reproducing kernel Hilbert space (RKHS) associated with the Gaussian kernel (Steinwart and Christmann, 2008). The weighted misclassification error associated with decision function f is $\mathbb{E}[1\{Af(\mathbf{X}) < 0\}C_A(\alpha)]$. Minimizing the empirical analog of the weighted misclassification, that is, the empirical risk, is difficult due to the discontinuity of the indicator function. Using the hinge loss, $\phi(u) = \max(0, 1 - u)$, as a surrogate loss function (Bartlett *et al.*, 2006), an estimator for the optimal decision function is

$$\hat{f}_{\alpha, n}^{\lambda_n} = \arg \min_{f \in \mathcal{F}} [\mathbb{E}_n \phi\{Af(\mathbf{X})\}C_A(\alpha) + \lambda_n \|f\|^2], \quad (2)$$

where $\|\cdot\|$ is a norm on \mathcal{F} and λ_n is a penalty parameter. We discuss how to choose a value of λ_n in Section 5. In the following, we write \hat{f}_α in place of $\hat{f}_{\alpha, n}^{\lambda_n}$ to simplify notation. The problem of estimating the optimal classifier

in (2) can be solved using the SVM introduced by Cortes and Vapnik (1995).

We estimate the optimal classifier, \tilde{D}_α , using $\hat{D}_\alpha(\mathbf{X}) = \text{sign}\{\hat{f}_\alpha(\mathbf{X})\}$. For any $\alpha \in (0, 1)$, we can estimate the sensitivity and specificity of the estimated classifier using the empirical estimators $\hat{se}(\hat{f}_\alpha) = \mathbb{E}_n 1(A = 1)1[\text{sign}\{\hat{D}_\alpha(\mathbf{X})\} = 1]/\mathbb{E}_n 1(A = 1)$ and $\hat{sp}(\hat{f}_\alpha) = \mathbb{E}_n 1(A = -1)1[\text{sign}\{\hat{D}_\alpha(\mathbf{X})\} = -1]/\mathbb{E}_n 1(A = -1)$. Plotting $\hat{se}(\hat{f}_\alpha)$ against $1 - \hat{sp}(\hat{f}_\alpha)$ as functions of α yields a non-parametric estimator of the optimal ROC curve. The ROC curve encodes a continuum of classifiers indexed by α ; to select a single classifier, there are a number of methods for defining an optimal value, say α^* , for α . For example, one could choose the α^* that leads to the point on the ROC curve closest to $(0, 1)$ in Euclidean distance, the α^* that maximizes the sum of estimated sensitivity and specificity, or the α^* that maximizes the estimated sensitivity for a fixed minimum specificity estimate (López-Ratón *et al.*, 2014). The choice of α^* will depend on the clinical application of interest. We classify an individual presenting with covariates $\mathbf{X} = \mathbf{x}$ as $\hat{D}_{\alpha^*}(\mathbf{x})$.

3 | THEORETICAL RESULTS

For any $\alpha \in (0, 1)$, the estimated classifier is the sign of \hat{f}_α , the minimizer of the empirical hinge loss in a class \mathcal{F} as defined in (2). For any function, f , define $\mathcal{R}_\alpha(f) = \mathbb{E}[1\{\text{sign}\{f(\mathbf{X})\} \neq A\}C_A(\alpha)]$ to be the risk of f , and define the ϕ risk of f to be $\mathcal{R}_{\alpha, \phi}(f) = \mathbb{E}[\phi\{Af(\mathbf{X})\}C_A(\alpha)]$. Let $\mathcal{R}_\alpha^* = \inf_f \mathcal{R}_\alpha(f)$ and $\mathcal{R}_{\alpha, \phi}^* = \inf_f \mathcal{R}_{\alpha, \phi}(f)$. Furthermore, define $\tilde{f}_\alpha = \arg \min_{f \in \mathcal{F}} \mathcal{R}_\alpha(f)$ and $f_\alpha^* = \arg \min_f \mathcal{R}_\alpha(f)$, that is, \tilde{f}_α minimizes the risk over \mathcal{F} and f_α^* minimizes the risk over all measurable functions mapping \mathcal{X} into \mathbb{R} . Define $f_{\alpha, \phi}^*$ as in Theorem 4. Throughout, we assume that $f_{\alpha, \phi}^* \in \mathcal{F}$, that is, that the function that minimizes the ϕ risk is contained within the chosen class. If this is not the case, the consistency results given here will not hold; however, the estimated decision function will still yield a reasonable approximation to \tilde{f}_α due to the identity $\mathcal{R}_\alpha(f) \leq \mathcal{R}_{\alpha, \phi}(f)$. When $\alpha = 0$, the optimal classifier assigns -1 uniformly and when $\alpha = 1$, the optimal classifier assigns 1 uniformly. Focusing on $\alpha \in (0, 1)$ will enable us to avoid these trivial extremes. Nonetheless, many of our results hold for all $\alpha \in [0, 1]$. We will make this distinction explicit as needed. Throughout, we assume that all requisite expectations exist.

The following result gives a bound on the excess risk in terms of the excess ϕ risk. The proof is similar to that of Theorem 3.2 of Zhao *et al.* (2012) and uses Theorem 1 and Example 4 of Bartlett *et al.* (2006). We omit the proof here.

This result will be used later to show uniform consistency of the risk of the estimated decision function.

Lemma 1. *For any measurable $f : \mathcal{X} \rightarrow \mathbb{R}$ and any distribution P of (\mathbf{X}, A) , $\mathcal{R}_\alpha(f) - \mathcal{R}_\alpha^* \leq \mathcal{R}_{\alpha,\phi}(f) - \mathcal{R}_{\alpha,\phi}^*$.*

This result implies that the difference between the ϕ risk of the estimated decision function and the optimal ϕ risk is no smaller than the difference between the risk of the estimated decision function and the optimal risk. Therefore, we can consider the ϕ risk when proving convergence results.

Next, we establish a number of consistency results for the risk of the estimated decision function. We begin with Fisher consistency. This result implies that estimation using either the hinge loss or the zero-one loss will yield the true optimal classifier given an infinite sample, providing justification for using the proposed surrogate loss function. The proof follows from an extension to the proof of Proposition 3.1 of Zhao *et al.* (2012) and is in Web Appendix B.

Theorem 1. *For any $\alpha \in [0, 1]$, if $f_{\alpha,\phi}^*$ minimizes $\mathcal{R}_{\alpha,\phi}$, then $D_\alpha^*(\mathbf{x}) = \text{sign}\{f_{\alpha,\phi}^*(\mathbf{x})\}$ for almost all $\mathbf{x} \in \mathcal{X}$.*

The following result establishes consistency of the risk of the estimated decision function when estimation takes place within an RKHS. We then extend this consistency by showing that it is uniform in α . The proof of the following result closely follows the proof of Theorem 3.3 of Zhao *et al.* (2012) and is in Web Appendix B.

Theorem 2. *Let $\alpha \in [0, 1]$ be fixed and let λ_n be a sequence of positive, real numbers such that $\lambda_n \rightarrow 0$ and $n\lambda_n \rightarrow \infty$. Let \mathcal{H}_k be an RKHS with kernel function k and let $\bar{\mathcal{H}}_k$ denote the closure of \mathcal{H}_k . Then, for any distribution P of (\mathbf{X}, A) , we have that $|\mathcal{R}_\alpha(\hat{f}_\alpha) - \inf_{f \in \bar{\mathcal{H}}_k} \mathcal{R}_\alpha(f)| \xrightarrow{P} 0$ as $n \rightarrow \infty$.*

We next strengthen the consistency stated above by showing that the convergence is uniform in α when estimation uses a linear, quadratic, polynomial, or Gaussian kernel (see Steinwart and Christmann, 2008, for a discussion of kernel functions used with the SVM). The following lemma indicates that the estimated decision function lies in a Glivenko-Cantelli (GC) class (Kosorok, 2008) indexed by α , which will help us to extend the consistency stated above to uniform consistency in α . The proof is in Web Appendix B.

Lemma 2. *Let \hat{f}_α be estimated using a linear, quadratic, polynomial, or Gaussian kernel function. Then, $\{\hat{f}_\alpha : \alpha \in [0, 1]\}$ is contained in a GC class.*

Given that \hat{f}_α and $-\hat{f}_\alpha$ are contained in a GC class, we have by Corollary 9.27 (iii) of Kosorok (2008), that $\phi(\hat{f}_\alpha)$ and $\phi(-\hat{f}_\alpha)$ are contained in a GC class because ϕ is continuous. By Corollary 9.27 (ii) of Kosorok (2008), $1(A = 1)\phi(\hat{f}_\alpha)$ and $1(A = -1)\phi(-\hat{f}_\alpha)$ are contained in a GC class and thus, $L_{\alpha,\phi}(\hat{f}_\alpha)$ is contained in a GC class by Corollary 9.27 (i) of Kosorok (2008), where $L_{\alpha,\phi}(f) = \phi(Af)C_A(\alpha)$. It follows that $\sup_{\alpha \in [0,1]} |\hat{\mathcal{R}}_{\alpha,\phi}(\hat{f}_\alpha) - \mathcal{R}_{\alpha,\phi}(\hat{f}_\alpha)| \xrightarrow{P} 0$, where $\hat{\mathcal{R}}_{\alpha,\phi}(f) = \mathbb{E}_n \phi\{Af(\mathbf{X})\}C_A(\alpha)$. This uniform convergence result is used in the proof of Theorem 3, which is given in Web Appendix B.

Theorem 3. *Assume that \hat{f}_α is estimated using a linear, quadratic, polynomial, or Gaussian kernel. For any sequence λ_n of positive, real numbers satisfying $\lambda_n \rightarrow 0$ and $n\lambda_n \rightarrow \infty$ and any distribution P of (\mathbf{X}, A) ,*

$$\sup_{\alpha \in [0,1]} \left| \mathcal{R}_\alpha(\hat{f}_\alpha) - \inf_{f \in \mathcal{H}_k} \mathcal{R}_\alpha(f) \right| \xrightarrow{P} 0 \quad (3)$$

as $n \rightarrow \infty$, where \mathcal{H}_k is the RKHS associated with \hat{f}_α .

Note that we do not allow the sequence λ_n to depend on α , which is reflected in the implementation in Section 5. Additional results, including continuity of the risk function, are given in Web Appendix A.

4 | CONFIDENCE BANDS

In this section, we present a method for constructing confidence bands for the ROC curve of \hat{f}_α , which provide an indication of how well the estimated classifier will perform in future samples. A number of methods have been proposed for constructing confidence bands for ROC curves (Ma and Hall, 1993; Campbell, 1994; Jensen *et al.*, 2000; Claeskens *et al.*, 2003; Horváth *et al.*, 2008) in the setting where a subject is classified as positive when a single biomarker is larger than some threshold and the ROC curve is constructed by varying the threshold across the range of possible biomarker values. If classification is based on a single biomarker (or some other score, such as a predicted probability), then confidence bands can be constructed by approximating the biomarker distribution in the diseased and nondiseased populations. In the current setting, however, classification is based on $\hat{f}_\alpha(\mathbf{X})$. Thus, the biomarker itself varies across α , and constructing confidence bands requires asymptotic results for \hat{f}_α viewed as a stochastic process indexed by $\alpha \in [0, 1]$. Our approach to constructing confidence bands requires the following result, which characterizes the asymptotic distribution of the estimated sensitivity and specificity of \hat{f}_α , along with

the consistency results given in Section 3. A proof of the following result is provided in Web Appendix B.

Theorem 4. Let $se(\hat{f}_\alpha)$ be the true sensitivity, $\widehat{se}(\hat{f}_\alpha)$ be the estimated sensitivity, $sp(\hat{f}_\alpha)$ be the true specificity, and $\widehat{sp}(\hat{f}_\alpha)$ be the estimated specificity of \hat{f}_α , where \hat{f}_α is defined in (2), and assume that \mathcal{F} is a space of linear or polynomial functions. Define $f_{\alpha,\phi}^* = \arg \min_f \mathbb{E}[\phi\{Af(\mathbf{X})\}C_A(\alpha)]$, where the minimization is taken over all measurable functions mapping \mathcal{X} into \mathbb{R} , and assume that $f_{\alpha,\phi}^* \in \mathcal{F}$. Then,

$$\sqrt{n} \begin{Bmatrix} \widehat{se}(\hat{f}_\alpha) - se(\hat{f}_\alpha) \\ \widehat{sp}(\hat{f}_\alpha) - sp(\hat{f}_\alpha) \end{Bmatrix} \rightsquigarrow \begin{Bmatrix} \mathbb{G}_1(\alpha) \\ \mathbb{G}_2(\alpha) \end{Bmatrix}$$

as $n \rightarrow \infty$, where $\mathbb{G}_1(\alpha)$ and $\mathbb{G}_2(\alpha)$ are mean zero Gaussian processes with covariances and cross-covariance given in Web Appendix A.

Let $f_{pf}(\hat{f}_\alpha) = 1 - sp(\hat{f}_\alpha)$ be the false positive fraction for the decision function \hat{f}_α . Define $f_{pf}^{-1}(\cdot)$ such that $f_{pf}^{-1}\{f_{pf}(\hat{f}_\alpha)\} = \alpha$, that is, $f_{pf}^{-1}(u)$ is the weight α such that $1 - sp(\hat{f}_\alpha) = u$. Let $0 < \delta < 1/2$ be fixed. A quantile bootstrap algorithm for constructing an asymptotically correct $(1 - \gamma)100\%$ confidence band for the ROC curve, $se\{f_{pf}^{-1}(u)\}$, $\delta < u < 1$, is as follows:

- (1) Set a large number of bootstrap replications, B , a grid $\delta = z_1 < \dots < z_K = 1$, and a grid $0 = \alpha_1 < \dots < \alpha_M = 1$.
- (2) For $m = 1, \dots, M$, compute the estimated ROC curve, $\widehat{R}(\alpha_m) = \{1 - \widehat{sp}(\hat{f}_{\alpha_m}), \widehat{se}(\hat{f}_{\alpha_m})\}$.
- (3) For $k = 1, \dots, K$, compute $\widehat{y}(z_k)$ by linearly interpolating $\widehat{R}(\alpha_m)$.
- (4) For $b = 1, \dots, B$:
 - (a) Generate a weight vector $W_{b,n,i} = \xi_{b,i}/\bar{\xi}_b$, where $\xi_{b,1}, \dots, \xi_{b,n}$ are independent standard exponential random variables and $\bar{\xi}_b = n^{-1} \sum_{i=1}^n \xi_{b,i}$.
 - (b) For $m = 1, \dots, M$, set

$$\widehat{se}_b(\hat{f}_\alpha) = \mathbb{E}_n(W_{b,n}1(A = 1)1[\text{sign}\{\hat{f}_\alpha(\mathbf{X})\} = 1]) / \mathbb{E}_n\{W_{b,n}1(A = 1)\},$$

$$\widehat{sp}_b(\hat{f}_\alpha) = \mathbb{E}_n(W_{b,n}1(A = -1)1[\text{sign}\{\hat{f}_\alpha(\mathbf{X})\} = -1]) / \mathbb{E}_n\{W_{b,n}1(A = -1)\},$$

$$\text{and } \widehat{R}_b(\alpha_m) = \{1 - \widehat{sp}_b(\hat{f}_{\alpha_m}), \widehat{se}_b(\hat{f}_{\alpha_m})\}.$$

- (c) For $k = 1, \dots, K$, compute $\widehat{y}_b(z_k)$ by linearly interpolating $\widehat{R}_b(\alpha_m)$.

- (5) Let $\widehat{y}^p(z_k)$ be the p th quantile of $\{\widehat{y}_b(z_k) : b = 1, \dots, B\}$ and let p^* be the largest $p \in [0, 1]$ such that $\widehat{y}^{p^*/2}(z_k) \leq \widehat{y}_b(z_k) \leq \widehat{y}^{1-p^*/2}(z_k)$ for all $k = 1, \dots, K$ for at least $(1 - \gamma)B$ bootstrap samples.
- (6) Set $y_\ell(z_k) = \widehat{y}^{p^*/2}(z_k)$ and $y_u(z_k) = \widehat{y}^{1-p^*/2}(z_k)$.

The fact that the ROC curve may be discontinuous at 0 necessitates starting the confidence bands at some $\delta > 0$. One can also use alternate choices for the weights, for example, a multinomial weight vector $W_{b,n} = (W_{b,n,1}, \dots, W_{b,n,n})^T$ with probabilities $(1/n, \dots, 1/n)$ and n trials. Let $\overset{P}{\underset{W}{\rightsquigarrow}}$ denote convergence in probability over W , as defined in Section 2.2.3 and chapter 10 of Kosorok (2008). The following result states the consistency of the bootstrap.

Corollary 1. Let $\widehat{se}_W(\hat{f}_\alpha) = \mathbb{E}_n(W1(A = 1)1[\text{sign}\{\hat{f}_\alpha(\mathbf{X})\} = 1]) / \mathbb{E}_n\{W1(A = 1)\}$ and define $\widehat{sp}_W(\hat{f}_\alpha)$ similarly. Let $\widehat{R}_W(\alpha) = \{1 - \widehat{sp}_W(\hat{f}_\alpha), \widehat{se}_W(\hat{f}_\alpha)\}$ and let $\widehat{R}(\alpha)$ be as defined above. Then, for any $0 < \delta < 1/2$, $\widehat{R}_W(\alpha) \overset{P}{\underset{W}{\rightsquigarrow}} \widehat{R}(\alpha)$ in $\ell^\infty([\delta, 1])$.

Proof. By Lemmas 12.7 and 12.8 of Kosorok (2008), taking the inverse of a bounded, monotone function is Hadamard differentiable under mild regularity conditions. The result now follows by Theorem 4 above and Theorems 2.6 and 2.9 of Kosorok (2008). \square

Thus, $\{y_\ell(z_k), y_u(z_k)\}$ will cover $\widehat{y}(z_k)$ across $k = 1, \dots, K$ with probability $1 - \gamma$ for large enough n and B . In addition to the linear and polynomial SVM, this procedure will work for any classifier such that the estimated decision function is in a Vapnik-Chervonenkis (VC) class, such as a logistic regression classifier.

5 | SIMULATION EXPERIMENTS

To investigate the performance of classification using a weighted SVM and the resulting ROC curves and confidence bands, we use the following generative model. Let \mathbf{X} be generated according to $\mathbf{X} \sim N_p(\mu\mathbf{Z}, \sigma^2I)$, where \mathbf{Z} is equal to a vector of ones with probability q and a vector of negative ones with probability $1 - q$ and I is a $p \times p$ identity matrix. Thus, \mathbf{X} is a mixture of multivariate normal distributions with mixing probability q . Let $\pi(\mathbf{X}) = \text{expit}(\mathbf{X}^T\beta)$ for a $p \times 1$ vector β , where $\text{expit}(u) = \exp(u) / \{1 + \exp(u)\}$. Given \mathbf{X} , we let A be equal to 1 with probability $\pi(\mathbf{X})$ and -1 with probability $1 - \pi(\mathbf{X})$. Because $\pi(\mathbf{X})$ depends on \mathbf{X} only through a linear function of \mathbf{X} , we refer to this model below as the linear generative model. We also consider a generalization of

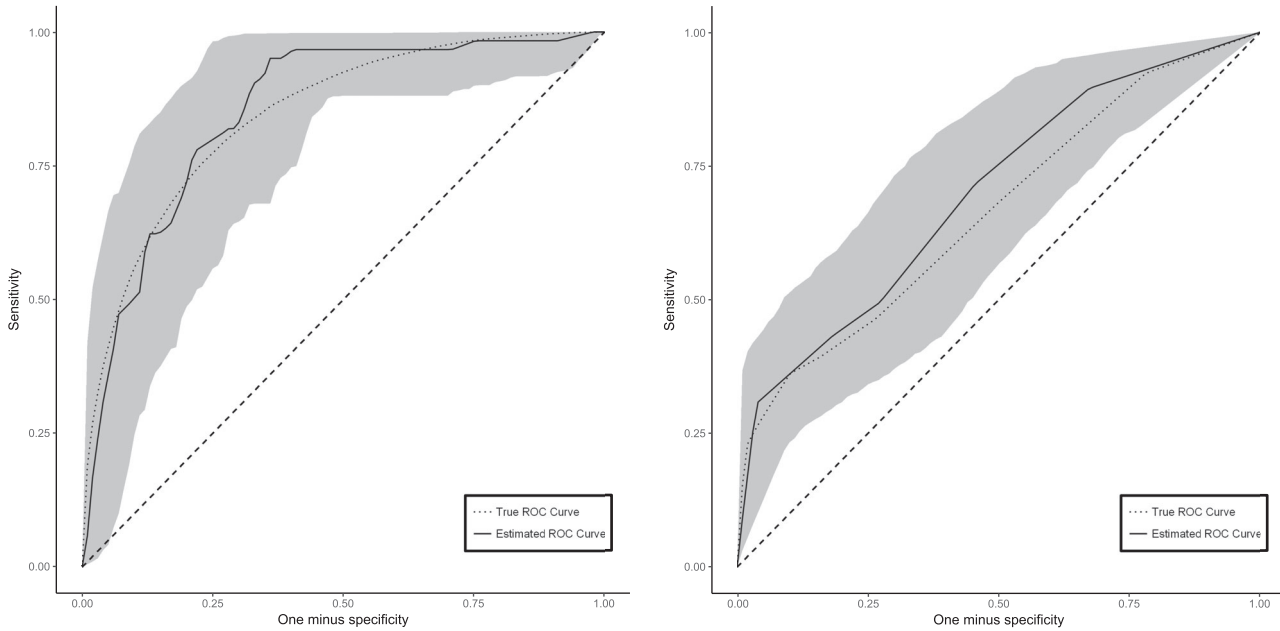


FIGURE 1 ROC curves and 95% confidence bands for $n = 500$, $p = 2$, and $q = 0.25$, when the true model is linear (left) and nonlinear (right).

the above model, where $\pi(\mathbf{X}) = \text{expit}(\mathbf{X}\boldsymbol{\beta} + X_1^2 + X_2^2 + 4X_1X_2)$, which we refer to below as the nonlinear generative model.

We implemented the weighted SVM in R software (R Core Team, 2016) using the kernlab package (Karatzoglou *et al.*, 2004). Each simulated data set is divided into training and testing sets with 70% of the data used for training the SVM and 30% used to estimate sensitivity and specificity. The penalty parameter, λ_n , is estimated using cross-validation within the training data for $\alpha = 0.5$, and the resulting tuning parameters are used to fit the weighted SVM for all α on a grid over $(0, 1)$. Tuning once at $\alpha = 0.5$ is needed to reduce the computational burden compared to tuning separately for each α . We applied the proposed method to data sets simulated according to the linear and nonlinear generative models with $n = 250, 500$, $p = 2, 5, 10$, $q = 0.05, 0.25$, $\sigma = 0.75$, and $\mu = 0.25$. When $p = 2, 5$, we use $\boldsymbol{\beta} = (2, 1)^\top$ and $\boldsymbol{\beta} = (2, 1, \dots, 1)^\top$, respectively. When $p = 10$, we use $\boldsymbol{\beta} = (2, 1, 1, 1, 1, 0, \dots, 0)^\top$, that is, noise variables are introduced for the case where $p = 10$.

Figure 1 below contains bootstrap confidence bands for one simulated replication of the linear SVM applied to the linear and nonlinear generative models when $n = 500$, $p = 2$, and $q = 0.25$. The true ROC curve, approximated using a large testing set of size 100 000 is also plotted. The plots in Figure 1 demonstrate that the proposed quantile bootstrap produces confidence bands that capture the true ROC curve and are sufficiently narrow as to provide useful inference about the future performance of an estimated SVM classifier.

TABLE 1 Estimated coverage probabilities and area between 90% confidence band curves

n	p	q	Coverage probability		Area between curves	
			Linear model	Nonlinear model	Linear model	Nonlinear model
250	2	0.05	0.85	0.86	0.31	0.37
		0.25	0.96	0.93	0.31	0.36
	5	0.05	0.83	0.93	0.28	0.38
		0.25	0.88	0.92	0.25	0.36
500	2	0.05	0.95	0.93	0.24	0.27
		0.25	0.92	0.96	0.23	0.27
	5	0.05	0.88	0.95	0.22	0.28
		0.25	0.96	0.93	0.18	0.26
10	0.05	0.86	0.93	0.23	0.29	
	0.25	0.96	0.94	0.20	0.27	

Table 1 contains the proportion of 100 Monte Carlo replications for which the true ROC curve was fully contained within the 90% confidence bands, along with area between the upper and lower confidence bands. Independent testing sets of size 100 000 were used to approximate the true ROC curve. Confidence bands for each replication were based on 1000 bootstrap samples.

We observe that, across n , p , and q , the proposed quantile bootstrap method provides approximately 90% coverage with the area between curves decreasing for larger sample sizes.

TABLE 2 Estimated coverage probabilities, area between 90% confidence band curves, and area under the curve when covariates are not all normal

n	p	q	Coverage probability	Area between the curves	Area under the curve
250	5	0.05	0.95	0.27	0.87
		0.25	0.96	0.26	0.88
10	5	0.05	0.94	0.28	0.87
		0.25	0.96	0.27	0.88
500	5	0.05	0.97	0.20	0.87
		0.25	0.98	0.20	0.88
		0.05	0.99	0.19	0.89
		0.25	0.96	0.18	0.90

To examine the performance of the proposed method when covariates are not all normally distributed, we used the following generative model. As before, let $(X_3, \dots, X_p) \sim N_{p-2}(\mu\mathbf{Z}, \sigma^2\mathbf{I})$ with \mathbf{Z} equal to a vector of ones with probability q and negatives ones with probability $1 - q$. Let X_1 be a Bernoulli random variable equal to 1 with probability $\text{expit}(X_3)$ and let X_2 be a Poisson random variable with mean $\text{exp}(X_3/4)$. As before, let $\pi(\mathbf{X}) = \text{expit}(\mathbf{X}^T\beta)$, and let A be equal to 1 with probability $\pi(\mathbf{X})$. We used $\beta = (1, 1, 2, 1, 0)$ for $p = 5$ and $\beta = (1, 1, 2, 1, 1, 0.5, 0.5, 0, 0, 0)$ for $p = 10$. We let $\mu = 0.25$ and $\sigma = 0.75$. Table 2 contains estimated coverage probabilities for 90% confidence bands, area between the confidence band curves, and area under the ROC curve for the linear SVM, averaged across 100 Monte Carlo replications. Each confidence band is based on 1000 bootstrap samples, and 70% of the data is used to train the weighted SVM in each replication.

The proposed confidence band method achieves greater than 90% coverage, with area between the curves decreasing with larger sample size. Area under the curve indicates that the weighted SVM can be used for classification in the presence of covariates that are not normally distributed. Additional simulation results are given in Web Appendix C in the Supporting Information.

6 | APPLICATION TO DATA

We apply the weighted SVM to the problem of predicting treatment response among patients with breast cancer. Many breast cancer patients will receive chemotherapy prior to surgery, called neo-adjuvant therapy, with the goal of shrinking the tumor to allow for a less invasive surgery. Predictive models for treatment response have the potential to aid physicians making treatment decisions; patients who are likely to respond to neo-adjuvant therapy should

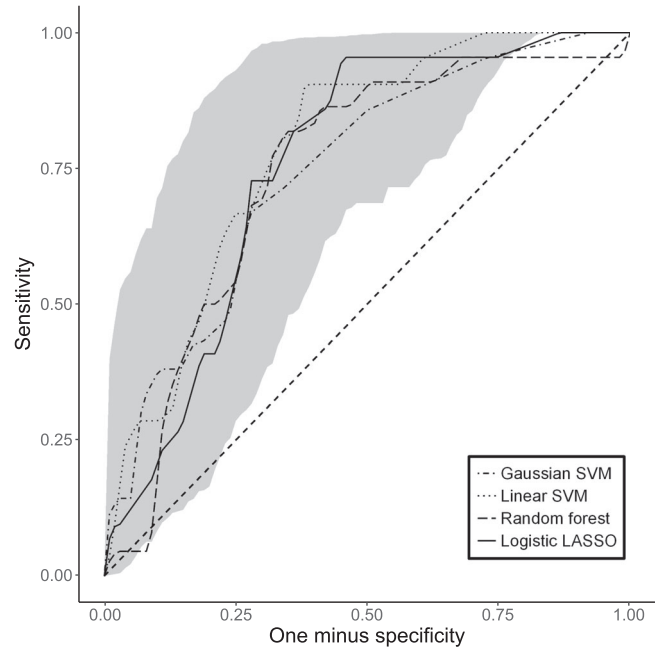


FIGURE 2 ROC curves for predicting response to treatment among breast cancer patients with 95% confidence bands for the linear SVM ROC curve

receive it, while patients who are unlikely to respond may achieve better outcomes with chemotherapy administered after surgery. The consequences of false positives and false negatives differ significantly for this classification problem. Although a patient who is incorrectly predicted to be a responder will be subjected to unnecessary treatment, a patient who is incorrectly predicted to be a nonresponder may be denied treatment that would have been beneficial. A low response rate can cause an unweighted classifier to produce high specificity at the expense of low sensitivity. This would result in a large proportion of patients who would have responded being incorrectly predicted as nonresponders, potentially missing out on beneficial treatment.

The data available for this analysis consist of 323 patients with complete data. For each patient, we calculated a collection of 512 gene expression signatures, called modules, each of which is a function of patient gene expression data (Fan *et al.*, 2011). We also observe a variety of clinical variables, for example, age and tumor stage. The data were divided into training and testing sets with 70% used to train the model. Figure 2 contains ROC curves for predicting response to treatment using the linear and Gaussian SVM, logistic regression with LASSO penalty (Tibshirani, 1996), and random forests (Breiman, 2001), along with confidence bands for the linear SVM. Each method performs equally well; each ROC curve falls within the linear SVM ROC curve confidence bands for much of the interval between 0 and 1, with the curves for

TABLE 3 Comparison of methods applied to breast cancer data

Method	AUC	\hat{se} (optimal)	\hat{sp} (optimal)	\hat{se} (unweighted)	\hat{sp} (unweighted)
Gaussian SVM	0.74	0.67	0.72	0.10	1.00
Linear SVM	0.79	0.90	0.63	0.19	0.97
Random forest	0.74	0.82	0.65	0.05	0.99
Logistic LASSO	0.75	0.73	0.72	0.00	1.00

the Gaussian SVM, random forest, and logistic regression falling outside the confidence bands close to 1. Table 3 contains AUC and optimal sensitivity and specificity for each method along with the sensitivity and specificity of the unweighted versions of each method. Although the linear SVM achieves the best AUC on these data, the wide confidence bands indicate large variability in the ROC curve for this classifier. Each method achieves a better balance between sensitivity and specificity after proper weighting. Unweighted classification results in close to perfect specificity at the expense of very low sensitivity for each method. This is likely due to the imbalance in the data (only 22% of patients in the sample respond).

We also apply the weighted SVM to a cohort study of mother-to-infant transmission of HCV (Shebl *et al.*, 2009). In this study, infants born to infected mothers were tested for HCV RNA using a polymerase chain reaction test and HCV antibodies using an enzyme-linked immunosorbent assay test. Mothers in the study were also tested for HCV RNA and antibodies during pregnancy. We applied the method of Veropoulos *et al.* (1999) to construct an SVM classifier for infant diagnosis of HCV using the results of diagnostic tests from the infants and mothers, and constructed confidence bands for the ROC curve using the proposed approach. Full results are given in Web Appendix D in the Supporting Information.

7 | CONCLUSION

A wide variety of problems in biomedical decision-making can be expressed as classification problems, such as diagnosing disease and predicting response to treatment. In some clinical applications, false positives may have different costs from false negatives; classification methods that can properly weight sensitivity and specificity and estimate an ROC curve are needed, along with inference methods for the ROC curve. Constructing an ROC curve using a sequence of weighted SVMs has been considered by Veropoulos *et al.* (1999). We have established the theoretical justification for the SVM ROC curve and provided a bootstrap method to construct confidence bands for the SVM ROC curve.

There is great potential for research in applying machine learning to diagnostic medicine and other biomedical decision-making problems. Methods of variable selection

for the weighted SVM (Dasgupta *et al.*, 2019) in this context would be an important step forward for this research. Other areas of future work may include developing methods to accommodate biomarker measurements that are taken at different time points from the same patient.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the following funding sources: NIH T32 CA201159, NIH P01 CA142538, National Center for Advancing Translational Sciences UL1 TR001111, NSF DMS-1555141, NSF DMS-1557733, NSF DMS-1513579, NIH 1R01DE024984, NIH U01HD39164, NIH U01AI58372, an investigator initiated grant from Merck, NCI Breast SPORE program (P50-CA58223-09A1), the Breast Cancer Research Foundation, and the V Foundation for Cancer Research


DATA AVAILABILITY STATEMENT

Research data are not shared as no new data were generated in this paper and the example data are not publicly available due to privacy restrictions.

ORCID

Daniel J. Luckett  <https://orcid.org/0000-0002-6358-9081>

Eric B. Laber  <https://orcid.org/0000-0003-2640-7696>

Michael R. Kosorok  <https://orcid.org/0000-0002-6070-9738>

REFERENCES

- Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006) Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 138–156.
- Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.
- Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.
- Cai, T. and Dodd, L.E. (2008) Regression analysis for the partial area under the ROC curve. *Statistica Sinica*, 18, 817–836.
- Cai, T. and Pepe, M.S. (2002) Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association*, 97, 1099–1107.
- Campbell, G. (1994) Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13, 499–508.
- Claeskens, G., Jing, B.-Y., Peng, L. and Zhou, W. (2003) Empirical likelihood confidence regions for comparison distributions and ROC curves. *Canadian Journal of Statistics*, 31, 173–190.

- Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273–297.
- Dasgupta, S., Goldberg, Y. and Kosorok, M.R. (2019) Feature elimination in kernel machines in moderately high dimensions. *The Annals of Statistics*, 47, 497–526.
- Etzioni, R., Pepe, M., Longton, G., Hu, C. and Goodman, G. (1999) Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19, 242–251.
- Fan, C., Prat, A., Parker, J.S., Liu, Y., Carey, L.A., Troester, M.A. and Perou, C.M. (2011) Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Medical Genomics*, 4, 3.
- Horváth, L., Horváth, Z. and Zhou, W. (2008) Confidence bands for ROC curves. *Journal of Statistical Planning and Inference*, 138, 1894–1904.
- Jensen, K., Müller, H.-H. and Schäfer, H. (2000) Regional confidence bands for ROC curves. *Statistics in Medicine*, 19, 493–509.
- Jiang, B., Zhang, X. and Cai, T. (2008) Estimating the confidence interval for prediction errors of support vector machine classifiers. *Journal of Machine Learning Research*, 9, 521–540.
- Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004) kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11, 1–20.
- Kosorok, M.R. (2008) *Introduction to Empirical Processes and Semiparametric Inference*. New York, NY: Springer Science & Business Media.
- Krzyżak, A., Linder, T. and Lugosi, G. (1996) Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. *IEEE Transactions on Neural Networks*, 7, 475–487.
- Lin, H.-T., Lin, C.-J. and Weng, R.C. (2007) A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68, 267–276.
- Lin, Y. (2002) Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6, 259–275.
- López-Ratón, M., Rodríguez-Álvarez, M.X., Cadarso-Suárez, C. and Gude-Sampedro, F. (2014) Optimalcutpoints: an R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61, 1–36.
- Ma, G. and Hall, W. (1993) Confidence bands for receiver operating characteristic curves. *Medical Decision Making*, 13, 191–197.
- McIntosh, M.W. and Pepe, M.S. (2002) Combining several screening tests: optimality of the risk score. *Biometrics*, 58, 657–664.
- Pepe, M.S. (1997) A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84, 595–608.
- Pepe, M.S. (2000) An interpretation for the ROC curve and inference using GLM procedures. *Biometrics*, 56, 352–359.
- Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.
- Platt, J. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10, 61–74.
- Provost, F. and Fawcett, T. (1998) Robust classification systems for imprecise environments. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 706–713.
- Provost, F.J. and Fawcett, T. (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 43–48.
- R Core Team (2016) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Shebl, F.M., El-Kamary, S.S., Saleh, D.A., Abdel-Hamid, M., Mikhail, N., Allam, A., et al., (2009) Prospective cohort study of mother-to-infant infection and clearance of hepatitis C in rural Egyptian villages. *Journal of Medical Virology*, 81, 1024–1031.
- Shin, S.J., Wu, Y., Zhang, H.H. and Liu, Y. (2014) Probability-enhanced sufficient dimension reduction for binary classification. *Biometrics*, 70, 546–555.
- Spackman, K.A. (1989) Signal detection theory: valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning*, 160–163.
- Steinwart, I. and Christmann, A. (2008) *Support Vector Machines*. New York, NY: Springer Science & Business Media.
- Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- Vapnik, V. (1998) *Statistical Learning Theory*. New York, NY: Wiley.
- Veropoulos, K., Campbell, C., and Cristianini, N. (1999) Controlling the sensitivity of support vector machines. *Proceedings of the International Joint Conference on AI*, 55–60.
- Zhang, T. (2004) Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32, 56–85.
- Zhao, Y., Zeng, D., Rush, A.J. and Kosorok, M.R. (2012) Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107, 1106–1118.
- Zhou, X.-H., McClish, D.K. and Obuchowski, N.A. (2002) *Statistical Methods in Diagnostic Medicine*. New York, NY: John Wiley & Sons.

SUPPORTING INFORMATION

Web Appendix A, containing additional discussion, Web Appendix B, containing proofs, Web Appendix C, containing additional simulation results, and Web Appendix D, containing an additional illustrative example, referenced in Sections 1, 5, and 6, along with R code to implement the proposed methods, are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Lockett DJ, Laber EB, El-Kamary SS, Fan C, Jhaveri R, Perou CM, Shebl FM, Kosorok MR. Receiver operating characteristic curves and confidence bands for support vector machines. *Biometrics*. 2020;1–9.
<https://doi.org/10.1111/biom.13365>