BMC Bioinformatics

RESEARCH ARTICLE Open Access

Statistical learning techniques applied to epidemiology: a simulated case-control comparison study with logistic regression

John J Heine[1*], Walker H Land[2], Kathleen M Egan[1]

Abstract

Background: When investigating covariate interactions and group associations with standard regression analyses, the relationship between the response variable and exposure may be difficult to characterize. When the relationship is nonlinear, linear modeling techniques do not capture the nonlinear information content. Statistical learning (SL) techniques with kernels are capable of addressing nonlinear problems without making parametric assumptions. However, these techniques do not produce findings relevant for epidemiologic interpretations. A simulated case-control study was used to contrast the information embedding characteristics and separation boundaries produced by a specific SL technique with logistic regression (LR) modeling representing a parametric approach. The SL technique was comprised of a kernel mapping in combination with a perceptron neural network. Because the LR model has an important epidemiologic interpretation, the SL method was modified to produce the analogous interpretation and generate odds ratios for comparison.

Results: The SL approach is capable of generating odds ratios for main effects and risk factor interactions that better capture nonlinear relationships between exposure variables and outcome in comparison with LR.

Conclusions: The integration of SL methods in epidemiology may improve both the understanding and interpretation of complex exposure/disease relationships.

Background

The objectives of this work are to 1) demonstrate the benefits of applying statistical learning (SL) concepts to epidemiologic type problems using simulated data when nonlinearities are present, and 2) adapt the SL approach to produce findings relevant for epidemiologic interpretation. Statistical learning effectively describes statistical estimation with small samples [1]. The approach does not rely on prior knowledge of the mathematical form of the exposure/disease relationship, an assumption in parametric modeling. A more detailed account of SL theory is provided elsewhere [1,2].

A comparison of a kernel based SL technique with logistic regression (LR) modeling was developed using simulated case-control datasets with a focus on the separation boundary and information embedding characteristics of both approaches. Illustrations were developed to demonstrate how the kernel mapping addresses the nonlinearity without user imposition. Without loss of generality, a low-dimensional problem was used to demonstrate the central themes because the separation boundaries can be observed graphically, which is not the case for higher-dimensional problems. The comparison with LR serves three purposes. First, although LR modeling is widely used for epidemiologic applications, its separation boundary represents a latent characteristic that is often not considered directly. Secondly, the information embedding characteristic of LR is representative of parametric approaches. The possible benefits derived from applying a kernel based technique come with a tradeoff in comparison with parametric modeling of requiring training data for prospective analyses. Thirdly, the LR model has an important epidemiologic interpretation. Therefore, the SL approach was modified to conform to the LR model interpretation.

Epidemiologic research makes frequent use of LR modeling for determining relationships between covariates

* Correspondence: john.heine@moffitt.org
[1]H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, 33612, USA
Full list of author information is available at the end of the article

BioMed Central

© 2011 Heine et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

and group associations when the outcome is binary. We will refer to the group association as the binary disease status and refer to covariates as *risk factors or exposures*. Logistic regression has many attractive attributes in this setting. The model coefficients are related to odds ratios (ORs) by exponentiation, which convey relevant exposure/disease association relationships. The LR model is a generalized linear model [3,4]. Various methods have been investigated to generalize such relationships in epidemiologic research. Neural network (NNs) have been used in studies of immunodeficiency viral infection [5] and liver disease [6,7]. Other researchers modified the LR model to include non-parametric functions to study colon cancer [8]. Generalized models have also been used in various capacities to model lung function change [9], blood pressure [10], alcohol consumption [11], and heart disease [12].

We will consider a dataset assembled from a case-control study in which each observation contains information on the binary disease status and a set of associated exposures. These exposures can be assembled into one vector, $\mathbf{x}$, for each observation, which we label as the *input*. Hypothetically, there is some relation $f(\mathbf{x})$ that describes the separation boundary between the case and control groups to some specified degree, where the group status is the *output*. Otherwise, $\mathbf{x}$ would not show association with disease. In a multivariate setting, the separation, or decision boundary, is a hyper-surface that reduces to a hyper-plane when $\mathbf{x}$ and the disease status bear a linear relation. Error in predicting group status may occur from a number of sources including inferior model specification, complicated relationships between the exposure distributions and group status, random error, non-random measurement error, or some combination of these influences. In practice, decision models rarely, if ever, produce perfect class-separation when making predictions.

We will consider a model encompassing two-exposures for each observation [i.e., a two-dimensional input vector $\mathbf{x} = (x_1, x_2)$ for each observation] in which the solutions and covariate relationships can be viewed in a two-dimensional plane by design. For a linearly separable two-dimensional problem, the input/output separation boundary is a straight line. When this problem is nonlinear separable, the input/output separation boundary is a curve (one dimensional) of some form. In practice, $f(\mathbf{x})$ is rarely known. Interaction terms (or other functional forms) can be introduced within the LR model to capture the attributes of $f(\mathbf{x})$, which are discernable graphically in a two-dimensional problem. However, in higher dimensional problems, it may not be clear whether the modified LR model provides a correct fit of the data. The two-dimensional problem demonstrated herein is used for illustrative purposes though it is representative of higher dimensional problems that are difficult or impossible to observe and model by intuition.

Odds ratios and the area under the receiver operator characteristic (ROC) curve, designated as Az, are used for comparing group characteristics for different purposes. When model predictive capability is important, Az is often used as the measure of separation in two-class problems [13-15]. In epidemiologic research, ORs are used to gauge the magnitude of association between exposure and outcome. In contrast with the LR model, the SL approach does not produce a data representation that has a useful epidemiologic interpretation. Therefore, we present non-parametric probabilistic methods that can be used for converting SL outputs to more readily interpretable ORs. We also calculated the Az quantity for each model used in the comparison analysis because it is measure of how well the models fit the data. The relationship between ORs and ROC analysis has been previously described [16].

In this report, a SL technique comprised of the kernel mapping in combination with a perceptron NN [17] was compared with the LR modeling. Kernel mappings are used to capture the non-linear relationship between the input/output without prior knowledge of the form of $f(\mathbf{x})$. We simulated data from a case-control study, which is a study-design employed in our ongoing epidemiologic research [18,19]. The goals of this ongoing research are analogous to those of Phase I or Phase II clinical studies wherein the objective is to determine whether certain exposures or measurements are more (or less) likely to be associated with a targeted disorder [20], where the disorder in our work is breast cancer. There is no explicit intent to make predictions at the population level at this time, though our methods could be adapted for this purpose in the future.

Methods

An overview of the multiple steps used for this analysis is shown in Figure 1. Briefly, we simulated one training dataset that was used exclusively to determine all of the model parameters for both the LR and SL approaches and perform an initial evaluation. We then evaluated the fitted models with multiple independent simulated datasets (validation datasets) to estimate the variation in the model performance.

Simulated Case-Control Study

A simulated case-control dataset was generated with m = 200 observations in each of the case and control groups, which is a relatively small sample size by design. Both random variables (rvs) and their respective realizations are denoted by lower case letters, and vectors are similarly labeled with bold letters. To avoid using

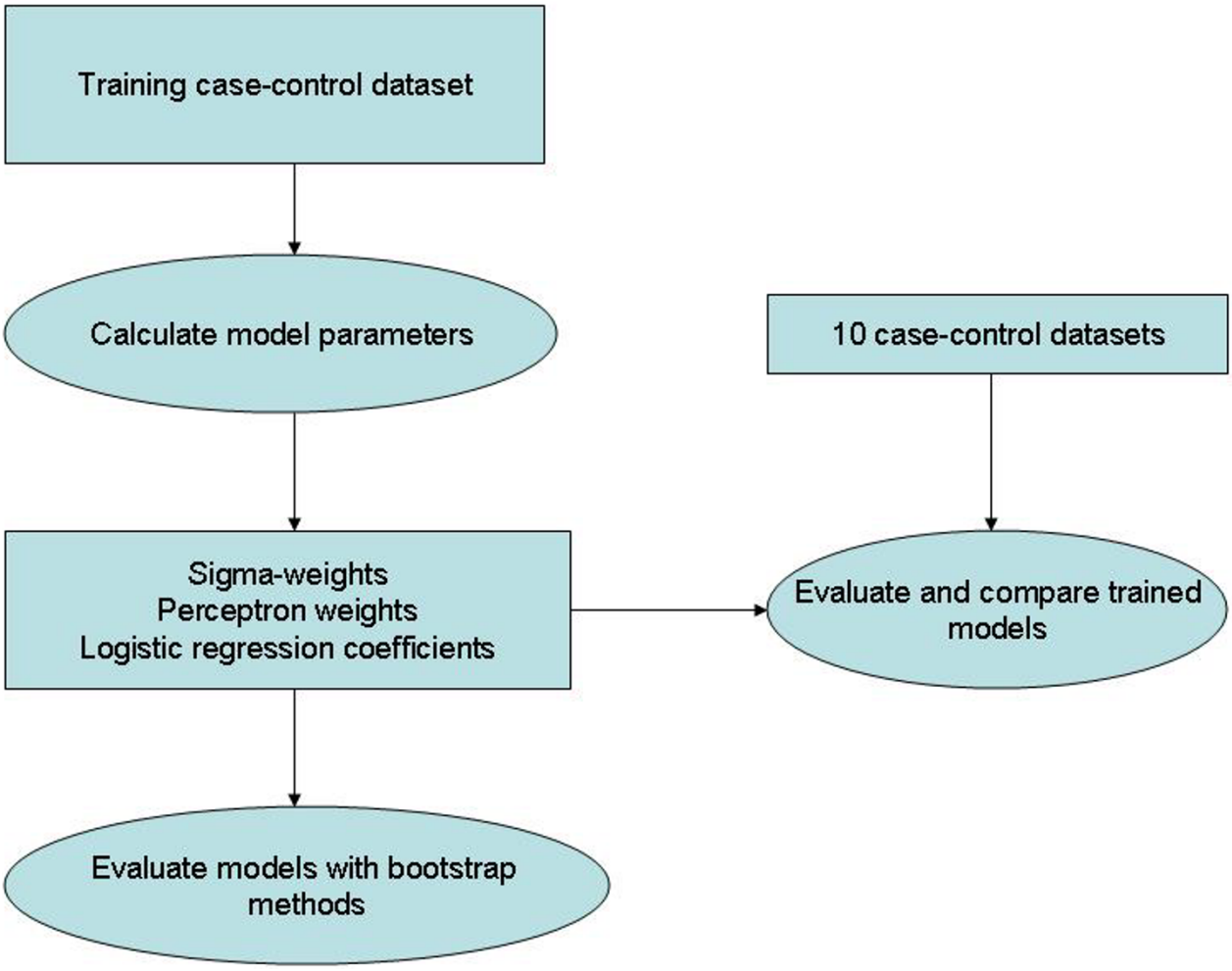

Figure 1 Training and evaluation scheme. This figure shows the analysis sequence. One simulated training dataset was used to estimate the parameters for all models and perform an initial evaluation with the fitted models. Ten independent simulated datasets were then used to evaluate the fitted models to eliminate training bias.

transpose notation, all vectors are defined as row vectors. Each observation (simulated study subject) has two risk factors denoted by x_1 and x_2 expressed as a vector $\mathbf{x} = (x_1, x_2)$. We used an *activation* function to randomly generate the disease status defined as

$$g(x_1) = \frac{1}{c_0} \times \left(\frac{x_1^2}{x_1^2 + (1 - x_1)^2} + \exp[-(a_0 x_1 - m_0)^2] \right), \quad (1)$$

where a_0, c_0, and m_0 are adjustable constants. This expression provides a flexible nonlinear boundary. The left term within the brackets is a sigmoidal function constructed from a parabola [21] and the right term gives a scalable spatially adjustable bulge. The disease status is dependent upon a given observation's $\mathbf{x}$ composition by this relation: $g(x_1) > x_2$. When this condition is met, the given observation is placed in the case group with its known risk factor vector $\mathbf{x} = (x_1, x_2)$. Otherwise, the observation is designated as control group member with the same vector $\mathbf{x} = (x_1, x_2)$. In this example, $g(x_1)$ assumes the position of the unknown function $f(\mathbf{x})$ discussed above. Equation (1) in combination with the defined case-control designation rule is an rv transformation for x_1 that creates a nonlinear separation boundary stochastically.

Simulated case-control datasets

We generated one case-control dataset for training (model fitting to determine all parameters) and ten additional validation datasets for evaluation purposes using the following prescription (11 datasets in total). To generate a given case-control dataset, 20,000 observations of (x_1, x_2) were generated randomly and processed with $g(x_1)$, which created the case-control designation. The first m observations from each group were used to form

a given case-control dataset resulting in 2m observations with equal numbers of cases and controls (m controls and m cases). The x_1 observations were uniformly distributed rvs with unit variance. The x_2 observations were generated by adding x_1 to a normally distributed rv, designated as z_1, with unit variance and mean = 5 giving $x_2 = (x_1+z_1)/10$. The empirical linear correlation between x_1 and x_2 after the $g(x_1)$ processing was estimated as R = 0.25.

Decision Models

The model construction, training methods, evaluation, and separation boundary analysis are described below in detail. Simulated case-control datasets were modeled with two LR models and three SL variants. Training (in which we estimate the model parameters) and model evaluations were performed with independent datasets to eliminate fitting bias in the comparison analysis. In the model comparison analysis, both predictive capability (i.e., Az) and ORs were compared. The training and evaluation sequences are shown in Figure 1.

Statistical learning overview

First, the kernel mapping was applied to the input vectors. The kernel-transformed data was then processed by a perceptron [17] using an algorithm described previously [22]. The perceptron can be used to solve a system of linear equations where each equation is of the form $y = \mathbf{r}\cdot\mathbf{w}+b$. In this expression, $\mathbf{w}$ is arbitrary weight vector, $\mathbf{r}$ is an arbitrary risk factor vector similar to $\mathbf{x}$ above, b is a constant, and y is a two-class binary variable representing the disease/no-disease status (i.e., y = 1, or y = -1). Hereafter, we refer to the kernel mapping and perceptron combination as the SL approach. The perceptron weight determination will converge when the problem is well approximated a linear-separable.

Kernel mapping

We will use a kernel mapping to express the input such that it is suitable for the perceptron processing. Under general circumstances, the researcher will find it difficult, if not impossible, to specify the mapping function that provides for a linear separation boundary. The kernel operates on the risk factor vectors and eliminates the need to determine the general mapping function denoted by $\varphi(\mathbf{x})$. We use $\varphi(\mathbf{x})$ for the mapping function, which is the transformation that renders the input/output relationship linear if chosen properly, because it conforms with the standard notation used in SL developments. As defined above, each observation has an associated risk factor vector, where $\mathbf{x}_j = (x_{1j}, x_{2j})$ designates the j^{th} training sample's vector, and $\mathbf{x} = (x_1, x_2)$ is used specifically to designate an arbitrary prospective observation's vector (not a training sample). Reproducing Kernel Hilbert Space theory states that a suitable kernel can be defined as the inner product of the mapping functions [23] expressed as

$$k(\mathbf{x}, \mathbf{x}_j) = \left\langle \phi(\mathbf{x}), \phi(\mathbf{x}_j) \right\rangle, \tag{2}$$

where $\mathbf{x}$ is a prospective observation (random) vector, with the same dimensionality as $\mathbf{x}_j$, and $\langle\cdot,\cdot\rangle$ is the inner product operation. The challenge changes from finding the mapping function to finding a valid kernel (there are many) as described previously [24]. The right side of Eq. (2) allows for the use of the left side without knowing the form of the right side. To define the specific kernel used here, we first define the distance measure between the vectors $\mathbf{x}$ and $\mathbf{x}_j$ given by

$$D(\mathbf{x}, \mathbf{x}_j) = \sqrt{\frac{s_1(x_1 - x_{1j})^2}{\sigma_1^2} + \frac{s_2(x_2 - x_{2j})^2}{\sigma_2^2}}. \tag{3}$$

The extension to higher dimensional vectors follows the same form by extending the sum within the radical to include more component terms. Each vector component difference has its own sigma-weight (σ_1 and σ_2) that was determined with training methods discussed below. These sigma-weights must be estimated properly because they impact the decision performance. We used three variations of Eq. (3). The s_1 and s_2 are for identifications purposes in this report only. Equation (3) was used with both component terms ($s_1 = s_2 = 1$) as above and with the individual component differences in isolation with ($s_1 = 1$, $s_2 = 0$) when the focus was on x_1 and ($s_1 = 0$, $s_2 = 1$) when the focus was on x_2. The kernel is then defined as

$$k(\mathbf{x}, \mathbf{x}_j) = c_{\mathbf{x}} \times \exp[-D(\mathbf{x}, \mathbf{x}_j)], \tag{4}$$

where $c_{\mathbf{x}}$ is a normalization constant. Equation (4) with Eq. (3) is from a class of universal kernels [25]. The kernel operation represents both a mapping of the input vectors [23] and also forms the basis for estimating probability density functions [26,27].

To determine the parameters for the SL approach, we used each individual training observation as a substitute for the prospective observation by cycling through the kernel processing. More specifically, each $\mathbf{x}_j$ training sample is processed with every other $\mathbf{x}_i$ training sample using Eqs. (3-4) to determine both the sigma-weights and the perceptron weight vector (i.e., $\mathbf{x}$ takes on all $\mathbf{x}_i$ for i = 0 through 2m). The i^{th} row of $\mathbf{K}$ results from the kernel operation of the i^{th} sample with each of the other 2m samples (including itself) indexed by j = 1 through 2m. The resulting kernel elements form 2m × 2m matrix, $\mathbf{K}$, with elements $k(\mathbf{x}_i, \mathbf{x}_j) = k_{ij}$. A given row in the $\mathbf{K}$ matrix can be considered as new feature set (or row vector) for the respective observation (patient),

which is the dimensionality expansion characteristic of the SL approach. The decision rule using the trained model (determined sigma-weights and perceprton weights) to make prospective predictions on the observation **x** is given by

$$y = \phi(\mathbf{x}) \cdot \mathbf{W} + b, \quad (5)$$

where y is the estimate of the binary disease status, b is an arbitrary (bias) constant, and **w** is generic weight vector. Expanding **w** in terms of the mapping function gives

$$\mathbf{w} = \sum_{j=1}^{2m} \alpha_j \phi(\mathbf{x}_j). \quad (6)$$

Using Eq. (5) in Eq. (6) and performing the inner product gives

$$y = \left\langle \phi(\mathbf{x}), \sum_{j=1}^{2m} \alpha_j \phi(\mathbf{x}_j) \right\rangle + b = \sum_{j=1}^{2m} \alpha_j k(\mathbf{x}, \mathbf{x}_j) + b, \quad (7)$$

which follows from the kernel inner product relation [23]. Equation (7) allows for the use of the kernel rather than the mapping function. For training, we let $\mathbf{x} = \mathbf{x}_i$ in Eq. (7) giving

$$y_i = \sum_{j=1}^{2m} \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + b, \quad (8)$$

where α_j are the components of the new weight vector α. The components of α are the preceptron weights that were determined with the training dataset using this linear combination to predict the i^{th} training observation's known case-control status designated by y_i.

Perceptron processing

We employed bootstrap methods [28] with the perceptron algorithm during the training analysis to estimate $\boldsymbol{\alpha}$ in Eq. (8). In the perceptron algorithm used here, the bias term, b, is not affected by the inputs [the kernel elements in Eq. (8)] but is an externally applied value (b = 1), left unchanged during the determination of the weight vector that fixes the position of the separation boundary (but does not affect the boundary orientation). When processing a prospective sample from a given validation dataset, the prospective observation's vector, **x**, is processed with the case-control training dataset consisting of 2m known risk factor vectors. The prospective observation's estimated output score, y_{est}, was generated using the Eq. (8) relationship from above

$$y_{est} = \sum_{j=1}^{2m} \alpha_j k(\mathbf{x}, \mathbf{x}_j) + b \quad (9)$$

with the previously determined α and b. Equation (9) demonstrates the information embedding characteristic of the kernel operation and illustrates how the mapping captures the underlying probability densities. A given kernel element (elements of **K**) can be interpreted as either 1) similarity measure between the prospective observation's vector **x** with the j^{th} training sample's vector $\mathbf{x}_j$, or 2) as one element of a multivariate kernel probability density estimation for **x**. Each new score (for the prospective **x**) is determined by making comparisons with the entire training set.

Each of the 2m validation observation scores for a given dataset (one of 10 datasets) was generated with the above equation by letting their risk factor vectors take the position of **x**. The dimensionality of the problem was fixed by the training methods. The number of observations in a given validation dataset is irrelevant for the mechanics of the processing. In addition to using both risk factors simultaneously, the perceptron was also trained using x_1 and x_2 separately with the same procedure without regenerating the sigma-weights, which created two additional SL variants used in the comparison. To standardize the associations for the three SL models, the y_{est} scores derived from Eq. (9) for a given model output were treated as single unit (both cases and control scores) and linearly mapped between [0-1]; we labeled these normalized output scores as z.

Logistic Regression

The LR model is expressed as

$$\Pr(\text{class} = 1 \mid \mathbf{x}) = p(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)}, \quad (10)$$

where Pr indicates probability. This model was used with x_1 and x_2 without interaction (referred to as the standard model with $\beta_3 = 0$) and with $x_1 \times x_2$ interaction (referred to as the interaction model). The respective parameter vectors (β_0, β_1, β_2) and (β_0, β_1, β_2, β_3) for each model were determined with the training dataset. We note, this model embeds information in the coefficients (on the order of the dimensionality) regardless of the number of observations on hand and is representative of parametric approaches.

Training and evaluation methods

Both the SL approach and LR model required training to estimate the various parameters. These models were trained with the same training dataset consisting of 2m observations. Figure 1 shows the training and evaluation flow schematic. The LR models were fitted with SAS (SAS Institute, NC) software. The SL approach required more involved training with bootstrap re-sampling [28]. Because the sigma-weights impact the performance of the perceptron output, the perceptron training was embedded within

the sigma- weight estimation. Perceptron weights were determined by drawing row vectors from the **K** (training) matrix at random with replacement. The Az was used as a guide for convergence. Because there are only two sigma-weights, a constrained search was used by varying both weights over a range of values. For each sigma-weight combination, the perceptron weights were determined, and the Az value was estimated resulting in an experimental set of values: $\{\sigma_{1i}, \sigma_{2i}, Az_i\}$ for the i^{th} combination. The sigma-weights were determined by the position of the maximum Az value (Az_{max}): $\sigma_1 = \sigma_{1i}$ and $\sigma_2 = \sigma_{2i}$ where $Az_i = Az_{max}$. Once the sigma-weights were established, the perceptron weights were regenerated (fine-tuned) by incrementally increasing the Az convergence criterion using a feedback loop. The perceptron weights that gave the highest Az before non-convergence were used in the validation processing along with $\{\sigma_1, \sigma_2\}$. When using x_1 and x_2 individually [$s_1 = 0$ or $s_2 = 0$ in Eq. (3)], we retrained the perceptron with the same Az criterion using the respective sigma-weights (determined above). In sum, the sigma-weight pair in combination with the perceptron weights that gave the highest Az for given SL variant were used in the model evaluation comparison.

The training dataset was used to evaluate the fitted models initially by generating 10 repetitions of 150 bootstrap datasets [28]. Each bootstrap dataset was processed by each of the models. For a given repetition, the distribution mean (Az_{150}) and standard deviation (σ_{150}) were calculated for each model. Averages of the Az_{150} and σ_{150} quantities were used to estimate the respective average performances and standard errors (SEs). For independent evaluation, 10 additional datasets were processed by each fitted model to estimate the average performance and SEs.

Separation boundary analysis

To compare the specific separation boundaries produced by the various models, it was necessary to apply a threshold to each model's output and estimate its performance. For consistency and to avoid user imposition, the same method was used to set the threshold for each model. In two class prediction problems (disease/no disease) used to assign class status, an operating point (decision threshold) must be selected from the model output, often derived from the ROC curve. This operating point represents a tradeoff between making two errors [13,14]. These are 1) the error of classifying cases as controls, defined in summary as the false negative fraction (FN), which is equivalent to 1-sensativty, where the sensitivity is the correctly identified proportion of cases, which is often referred to as the true positive fraction (TP), and 2) the error of classifying controls as cases denoted as the false positive fraction (FP) in summary. Plotting the ordered pairs, (FP, TP), for each threshold, which is a latent variable, approximates the continuous ROC curve. Choosing a threshold fixes the separation boundary. For the LR model, all samples with $p(\mathbf{x})$ scores $\geq p_t$ were classified as case group members, otherwise they were classified as control group members, where p_t is a fixed threshold. To determine the separation boundaries, the operating point for a given model was selected by choosing the sensitivity equivalent to its Az value. Because the FP variable is defined over this range [0-1], the Az value may also be interpreted as the model's mean (average) sensitivity (i.e., the value of the area under the ROC curve is also the mean value of the ROC function). For an arbitrary threshold value, p_t, the separation boundary for the standard LR model was found by solving Eq. (10) for x_2, giving

$$x_2 = -\frac{(\tau_0+\beta_0)}{\beta_2} - \frac{\beta_1}{\beta_2} x_1 \tag{11}$$

with $\tau_0 = \ln\left(\frac{1-p_t}{p_t}\right)$, which is a linear boundary. Including the LR interaction term gives

$$x_2 = -\frac{\tau_0+\beta_0+\beta_1 x_1}{\beta_2+\beta_3 x_1}, \tag{12}$$

We will find the value of p_t that gives a sensitivity equivalent to the Az (or the mean sensitivity) for the respective LR models to determine the separating boundaries and estimate the corresponding FP for comparison purposes. The same approach was applied to the SL output. This method used to set the thresholds eliminated user input because there are an unlimited number of thresholds to choose from, each representing a different tradeoff as described above. Our objective is to show the form of the various separation boundaries, therefore the method used to set the threshold is not important to the central demonstration.

Odds Ratio Transformation

The SL technique output [the perceptron output defined in Eq. (9)] was modified to conform to the LR model interpretation and generate ORs. Specifically, we estimated the empirical conditional probability function $p_r = Pr(class = 1|z)$ as the reference, where z is the SL method normalized output score We then estimated $p_1 = Pr(class = 1|z+\Delta z)$ in the same manner, where Δz is in positive increment in the respective z score. The ORs were calculated using this definition

$$OR = \frac{p_1}{1-p_1} \times \frac{1-p_r}{p_r}. \tag{13}$$

Equation (13) can be applied by using all of the risk factors in the model or any subset. When using more than one risk factor, it can be considered as multivariate

OR. In the Eq. (13) representation, p_r has the analogous interpretation as the LR model in Eq. (10), although it was derived numerically. Equation (13) was generated for each of the SL variants for one of the evaluation datasets. We note that using Eq. (13) with these specific definitions for p_r and p_1 parallels the development used to derive the interpretation for the LR model coefficients for continuous independent variables [29].

The components (p_1 and p_r) in Eq. (13) were constructed as approximations for continuous functions using non-parametric techniques. To estimate p_1 and p_r, first the histograms of normalized output scores for the m cases and m controls were analyzed separately. A kernel density estimation technique [27] was used to generate the empirical probability densities from the output score histograms using a Gaussian kernel. The kernel density technique is a non-parametric method used to estimate the underlying probability density function given samples drawn from a given population without assumption that generalizes the respective histogram (similar to the kernel mapping). This is a particularly useful technique when the dataset is sample-limited with missing bins in the histogram because it is essentially a sifting mechanism that can eliminate discontinuities. The estimated densities for the cases and controls are denoted by h_1 and h_0, respectively, giving $p_r = h_1/(h_1+h_0)$, which is a function of z. The p_1 function was estimated similarly by shifting p_r by Δz.

Results

Model Training

Model parameters were determined and each model was assessed with the training dataset. The coefficients for the standard LR model using x_1, and x_2 simultaneously without interaction and with $x_1 \times x_2$ interaction were: $(\beta_0, \beta_1, \beta_2)$ = (-7.251, -1.743, 14.33) and $(\beta_0, \beta_1, \beta_2, \beta_3)$ = (-13.73, 9.66, 26.03, -20.12), respectively. These coefficients are presented as log(ORs) [i.e, ln(OR)] per unit increase in the respective variables. These large values are due to the unit increase because both x_1 and x_2 span less than one unit. For the standard model, x_1 provides a shielding effect with respect to the disease status (e.g., the coefficient remains negative, implying an inverse association of the factor with disease status), whereas x_2 shows a relatively stronger positive magnitude of association in comparison with x_1. In contrast, in the interaction model, the x_1 and x_2 terms both show a positive association with the outcome while the interaction term has a negative coefficient. For this initial Az assessment, averages, standard deviations, and SEs derived with bootstrap methods [28] are given in Table 1. The Az quantities for the training x_1 and x_2 sample distributions were also generated for comparison purposes; these Az quantities were estimated by comparing the respective distributions without model-processing. The sigma-weight pair in combination with the perceptron weights that gave the highest Az were used in the comparison evaluation: (σ_1, σ_2) = (3.88, 2.47). The trained model Az findings are given in Table 1 for the three SL models.

Table 1 Training area under the receiver operator characteristic curve quantities

Method	Az	σ	SE
LR	0.791	0.028	0.008
LR_{int}	0.814	0.027	0.008
k	0.958	0.013	0.004
k_{x1}	0.867	0.023	0.007
k_{x2}	0.728	0.031	0.009
x_1	0.490	0.035	0.011
x_2	0.772	0.029	0.009

This table gives the area under the receiver operator characteristic curve (Az) quantities derived from the training dataset for the standard logistic regression model with x_1 and x_2 (LR), the logistic regression model with x_1 and x_2 with $x_1 \times x_2$ interaction (LR_{int}), the statistical learning (SL) techniques using a kernel mapping with x_1 and x_2 simultaneously (k), and partial SL-kernel models using x_1 (k_{x1}) and x_2 (k_{x2}) individually. This also gives the Az quantities for the x_1 and x_2 case-control training distribution samples estimated without using model processing. Az and σ are the respective means and standard deviations summarized from the bootstrap trials. SE is the standard error in Az.

Model Evaluation

The two trained LR models and the three trained SL variants were used to process the 10 validation case-control datasets (Figure 1). Summarized Az findings for all model outputs are listed in Table 2, which mirror those in Table 1. The SL approach provided the best performance. The predictive capacity of the LR model is captured in the x_2 term by noting its coefficient. The LR model gained marginal predictive capacity by adding the interaction term as indicated by the increased Az value. In contrast, the univariate SL variants show that x_1 in isolation contains considerable

Table 2 Evaluation area under the receiver operator characteristic curve quantities

Method	Az	σ	SE
LR	0.781	0.029	0.009
LR_{int}	0.798	0.029	0.009
k	0.947	0.015	0.004
k_{x1}	0.852	0.018	0.005
k_{x2}	0.734	0.023	0.008

This table gives the area under the receiver operator characteristic curve (Az) quantities for the standard logistic regression model using x_1 and x_2 (LR), the logistic regression model using x_1 and x_2 with $x_1 \times x_2$ interaction (LR_{int}), the statistical learning (SL) model using a kernel mapping with x_1 and x_2 simultaneously (k), and partial SL-kernel models using x_1 (k_{x1}) and x_2 (k_2) individually. Az, and σ are the respective means and standard deviations derived from processing the 10 validation datasets with the trained models. SE is the standard error in Az.

information content in comparison with x_2. Figure 2 shows the linear separation boundary for the standard LR model plotted with the case-control data points. The solid line is the LR separation boundary derived from Eq. (11) with Az ≈ 0.78, which gave FP ≈ 0. 42 with $p_t \approx 0.42$. The other curve (dashed line) in Figure 2 represents the ideal boundary that was derived with Eq. (1). Figure 3 shows the separation boundary for the LR interaction model (same format) derived from Eq. (12) with Az ≈ 0.80, which gave FP ≈ 0.40 with $p_t \approx 0.41$. Figure 4 shows the SL plot derived with Az ≈ 0.95, which gave FP ≈ 0.33 with z ≈ 0.49 (the solid line separation boundary). In this plot, samples were ordered along the horizontal axis according to the observation index. The first 200 points correspond to controls and the next 200 points correspond to the cases. The respective normalized output scores are plotted on the vertical axis with the control scores denoted by multiplication signs and the case scores by diamonds. These examples illustrate the information embedding characteristics of the kernel mapping.

Once the model parameters were determined for the LR models, the functional form of their separation boundaries were fixed. For example, changing the thresholds for either of the LR models will shift the boundaries (Figure 2 and Figure 3) and provide different decision performance (i.e., different sensitivity and FP) but will not alter the boundary forms. The boundary in Figure 4 illustrates that the kernel mapping transformed the input/output relation from the separation boundary shown in Figure 2 or Figure 3 to the separation shown in Figure 4.

Odds Ratio Transformation

Odds ratios were calculated using x_1 and x_2 simultaneously, as well as individually, by applying Eq. (13) to

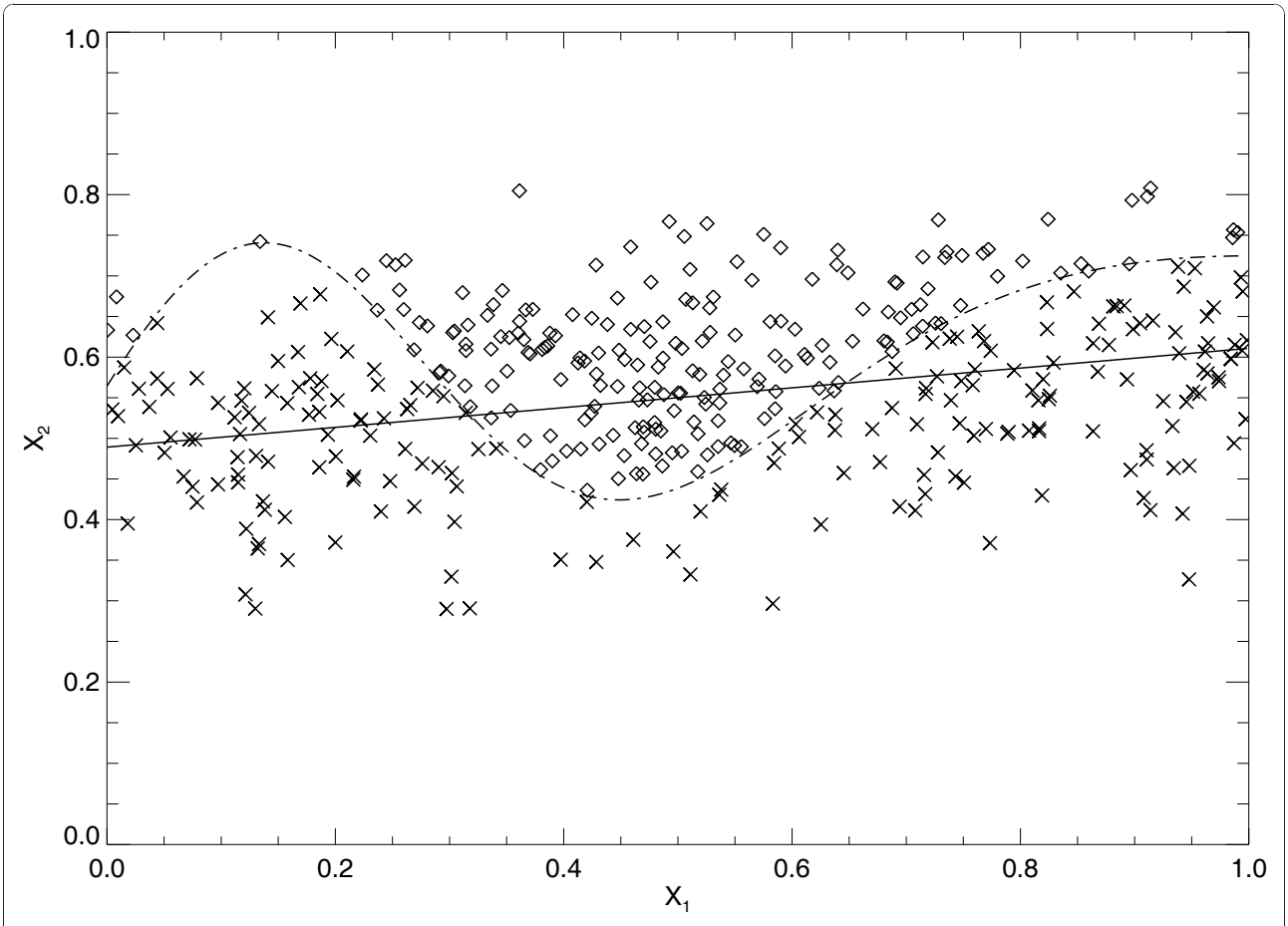

Figure 2 The x_1-x_2 scatter plot and logistic regression boundary. This figure shows the two risk factor scatter plot for cases (diamonds) and controls (multiplication signs). Each point represents a given sample's (x_1, x_2) risk vector plotted in component form. The solid line is the standard logistic regression (no-covariate interaction) model linear separation boundary for a fixed threshold and the curved dashed line is the Eq. (1) (ideal) separation boundary. The sensitivity = 0.78 and false positive fraction = 0.42.

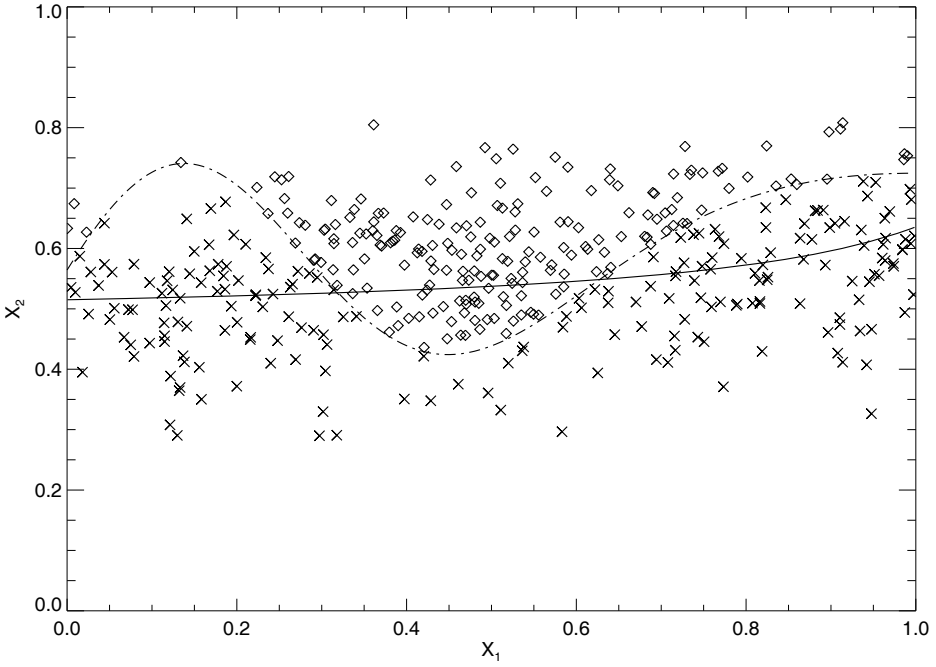

Figure 3 The x_1-x_2 scatter plot and the logistic regression with interaction boundary. This figure shows the two risk factor scatter plot for the cases (diamonds) and controls (multiplication signs) for the LR model with $x_1 \times x_2$ interaction. Each point represents a given sample's (x_1, x_2) risk vector plotted in component. The solid line is the LR model separation boundary (solid) and the curved dashed line is the Eq. (1) (ideal) separation boundary. In comparison with Figure 2, there is a slight curvature in the boundary on the right side. The sensitivity = 0.80 and false positive fraction = 0.40.

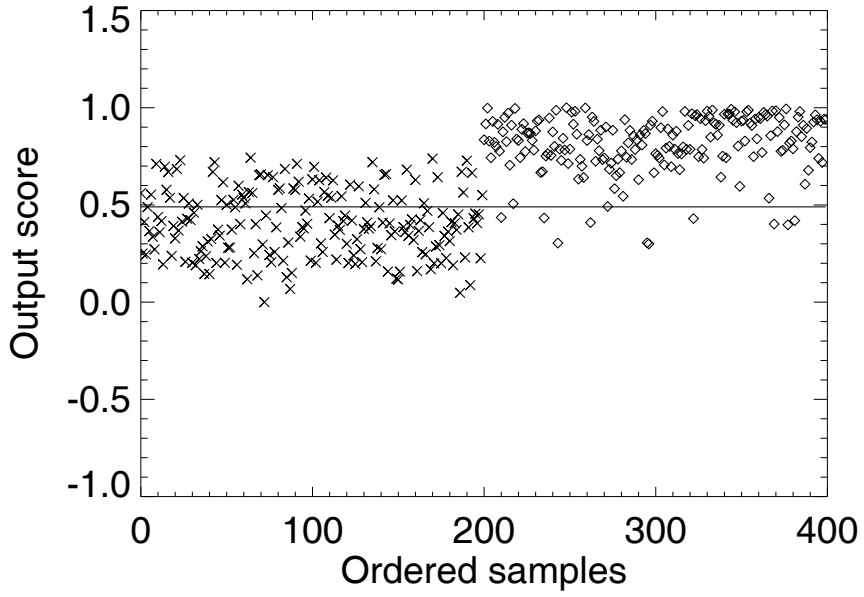

Figure 4 Statistical learning (SL) output and boundary. This figure shows the SL output separation boundary. Ordered samples are plotted along the horizontal-axis with the 200 control observations plotted first (multiplication signs on the left side) followed by the 200 case observations (diamonds on the right side). The SL output normalized z-scores for each sample are plotted on the vertical-axis. The separation boundary that gave 0.95 sensitivity is z = 0.49 (solid line) with a false positive fraction = 0.33.

each of the SL model's normalized output scores. For SL approach with both variables, the numerical estimate of p_r = Pr(class = 1|z) is shown in Figure 5 [same interpretation as Eq. (10)]. The ORs were then derived by letting p_1 = Pr(class = 1|z+Δz) with Δz = 0.10 (output-score increment units). The corresponding *continuous* log(OR) plot is shown in Figure 6, which can be considered as a multivariate OR showing the influence of both factors simultaneously. Similarly, the log(OR) plots for x_1 and x_2, individually, are shown in Figure 7 and Figure 8, respectively. In practice, the ORs can be rescaled. Because the problem was simulated, rescaling has little relevance. The focus of the analysis is the OR nonlinearity. These plots show the functional dependence of the ORs in comparison with the LR coefficients that are constants. When the log(ORs) derived from the SL outputs are constant, the Eq. (13) relations would approximate constant valued functions similar to the LR model coefficients, which are essentially average effects under the linear assumption.

Discussion

A two-dimensional problem was simulated to illustrate some advantages of applying SL techniques to epidemiological type datasets. Comparisons of the Az quantities among the various models (Table 2) demonstrates the capacity of the SL approach when addressing nonlinear problems in contrast with the LR results. The SL output scores were transformed into ORs using a kernel density estimation technique. This transformation provided the essential link between the SL output and the epidemiologic interpretation for both the multivariate OR relation, which is the combined disease/risk factor association for both (all) the covariates simultaneously including their interactions, as well as the individual risk factor associations. As demonstrated, the ORs exhibit (see figures 6-8) a nonlinear functional dependence with respect to the output score. When the input/output relationship is nonlinear, the LR coefficient does not describe the association properly due to the LR model linear separation boundary. We note that the LR output

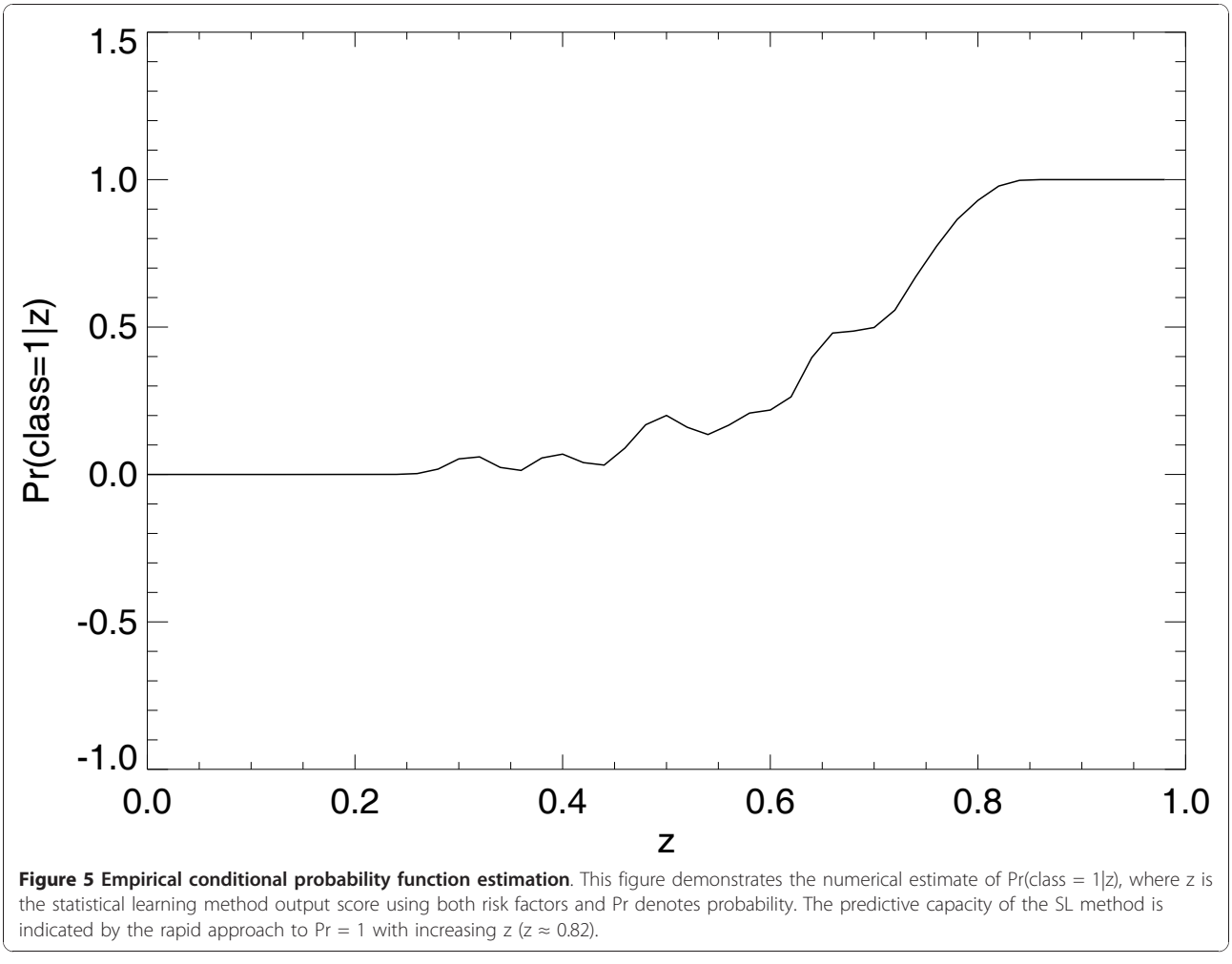

Figure 5 Empirical conditional probability function estimation. This figure demonstrates the numerical estimate of Pr(class = 1|z), where z is the statistical learning method output score using both risk factors and Pr denotes probability. The predictive capacity of the SL method is indicated by the rapid approach to Pr = 1 with increasing z (z ≈ 0.82).

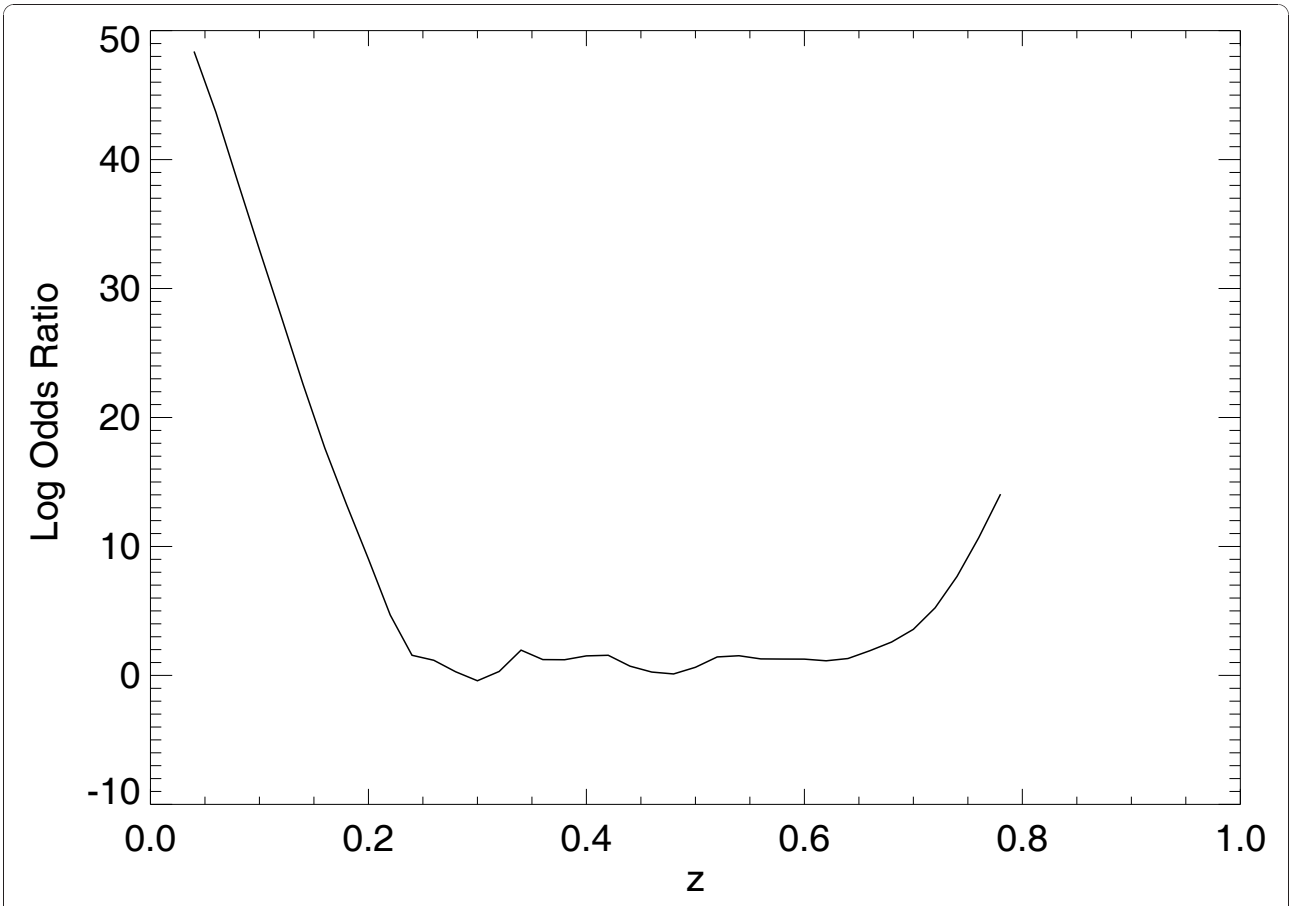

Figure 6 Multivariate logarithm of the odds ratio for the two-risk factor statistical learning method output. This figure demonstrates the log (odds ratio) [i.e., ln(odds ratio)] plot derived from the two-risk factor statistical learning method output, z, using the formulism illustrated in Figure 5 [$s_1 = 1$ and $s_2 = 1$ in Eq. (3)].

could be manipulated in the same fashion, but the relationship would not capture the correct interaction because of the linear model form.

Other researchers incorporated kernel density estimations in epidemiologic research for different applications [30-32]. Similar kernel density estimations techniques were used earlier to derive relative risks [31]. Duh et al [6] provided an epidemiologic interpretation of the NN weights when using an LR type activation function. In contrast with this related work using kernel density estimations, we applied the kernel density estimation to the SL model output after the kernel mapping. This approach used the decision model outputs as new risk factor quantities that captured the inherent nonlinearities.

The kernel mapping expands the dimensionality of the problem and uses the entire training dataset for prospective analysis. This expansion enables the SL system to learn the input/output relationship, which is captured in the kernel elements and the perceptron weight vector. Each kernel element in the Eq. (9) linear combination represents a similarity measure between the respective training sample and the prospective observation. This is in contrast with parametric modeling techniques that use relatively few model coefficients to summarize the training dataset attributes. The ability of the SL approach to learn the input pattern in exemplified by the Az result for x_1 when processed in isolation. The relatively large Az value resulting from the SL technique when including both exposure variables indicates the kernel mapping captured the nonlinear information content and transformed the original representation to a nearly linear separable representation. Generally, SL methods require more involved training than that of parametric modeling, an inevitable trade-off required to capture the nonlinearity. For higher-dimensional problems more sophisticated optimization techniques are required, such as those derived from differential evolution principles [33], to ensure the proper optimization is achieved and derived in an acceptable lengths of time.

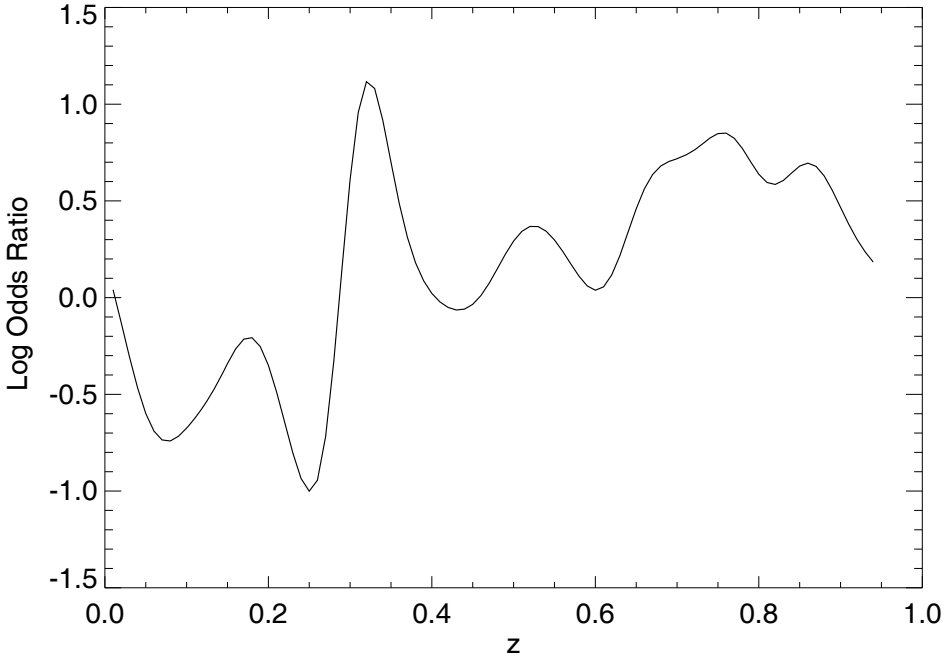

Figure 7 Logarithm of the odds ratio for the statistical learning method output for the first risk factor. This figure demonstrates the log (odds ratio) [i.e., ln(odds ratio)] plot derived from the statistical learning method output, z, using x_1 [$s_1 = 1$ and $s_2 = 0$ in Eq. (3)].

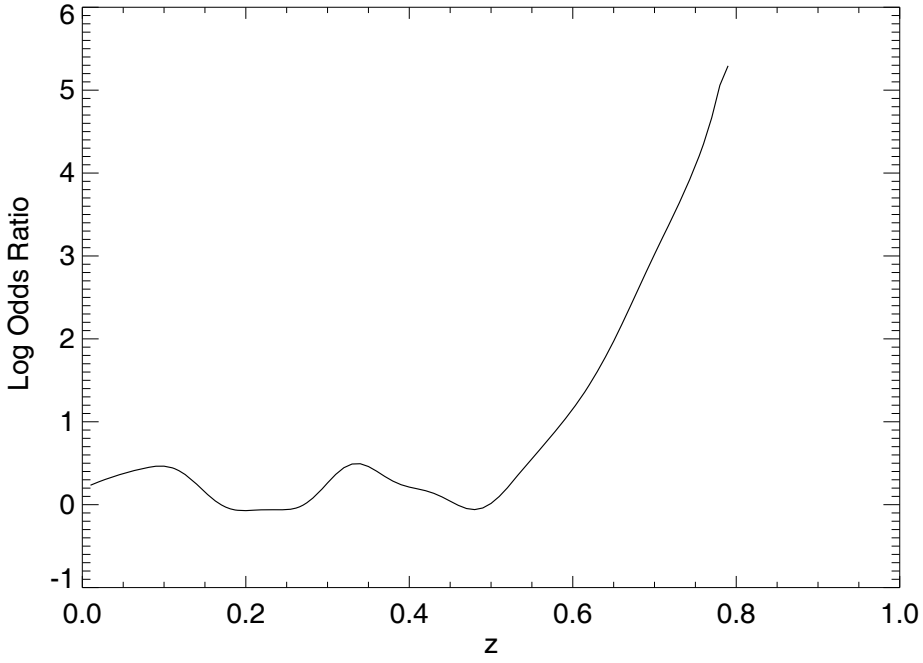

Figure 8 Logarithm of the odds ratio for the statistical learning method output for the second risk factor. This figure demonstrates the log (odds ratio) [i.e., ln(odds ratio)] plot derived from the statistical learning model output, z, using x_2. [$s_1 = 0$ and $s_2 = 1$ in Eq. (3)].

These simulations involved two risk factors and one outcome. However, we recognize that this scenario is seldom observed in real epidemiologic practice, in which more typically there are multiple covariates that may predict the outcome. Nevertheless, the simulations illustrated how SL techniques can potentially improve upon common methods currently applied in epidemiologic research when nonlinearities are present. The linear separation produced by the LR model was exemplified with a low-dimensional problem that contained all of the features of higher dimensional problems. The kernel mapping transformed the original relationship to a feature space where linear techniques are applicable without assuming interaction forms, although a valid kernel must be determined.

Conclusions

The work demonstrated the potential benefits derived from applying SL techniques to nonlinear epidemiologic type problems. Integrating SL techniques with epidemiologic research may aid researchers in defining complex exposure/disease relationships. These applications will require validation in population-based studies and further rigorous comparisons with existing methods.

Acknowledgements

The authors wish to thank Dr. Robert for his helpful insights in developing this work.

Author details

[1]H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, 33612, USA. [2]Binghamton University, Bioengineering Department, Binghamton, NY, USA.

Authors' contributions

JJH and WHL developed the statistical learning analysis methods. All authors contributed equally in the manuscript conception, experimental design, and composition. All authors read and approved the final manuscript.

Received: 16 June 2010 Accepted: 27 January 2011
Published: 27 January 2011

References

1. Vapnik VN: *Statistical Learning Theory* NY: John Wiley & Sons, Inc; 1998.
2. Vapnik VN: *The Nature of Statistical Learning Theory.* 2 edition. NY: Springer; 2000.
3. Myers RH, Montgomery DC: **A tutorial on generalized linear models.** *Journal of Quality Technology* 1997, **29**:274-291.
4. Nelder JA, Wedderburn RWM: **Generalized linear models.** *Journal of the Royal Statistical Society, Series A (General)* 1972, **135**:370-384.
5. Ioannidis JPA, McQueen PG, Goedert JJ, Kaslow RA: **Use of neural networks to model complex immunogenetic associations of disease: human leukocyte antigen impact on the progression of human immunodeficiency virus infection.** *American Journal of Epidemiology* 1998, **147**:464-471.
6. Duh MS, Walker AM, Ayanian JZ: **Epidemiologic interpretation of artificial neural networks.** *American Journal of Epidemiology* 1998, **147**:1112-1122.
7. Duh MS, Walker AM, Pagano M, Kronlund K: **Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example.** *American Journal of Epidemiology* 1998, **147**:407-413.
8. Zhao LP, Kristal AR, White E: **Estimating relative risk functions in case-control studies using a nonparametric logistic regression.** *American Journal of Epidemiology* 1996, **144**:598-609.
9. Cui J, de Klerk N, Abramson M, Del Monaco A, Benke G, Dennekamp M, Musk AW, Sim M: **Fractional polynomials and model selection in generalized estimating equations analysis, with an application to a longitudinal epidemiologic study in Australia.** *American Journal of Epidemiology* 2009, **169**:113-121.
10. Rosner B, Cook N, Portman R, Daniels S, Falkner B: **Determination of blood pressure percentiles in normal-weight children: some methodological issues.** *American Journal of Epidemiology* 2008, **167**:653-666.
11. Kimball AW, Friedman LA, Moore RD: **Nonlinear modeling of alcohol consumption for analysis of beverage type effects and beverage preference effects.** *American Journal of Epidemiology* 1992, **135**:1287-1292.
12. Abrahamowicz M, du Berger R, Grover SA: **Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality.** *American Journal of Epidemiology* 1997, **145**:714-729.
13. Faraggi D, Reiser B, Schisterman EF: **ROC curve analysis for biomarkers based on pooled assessments.** *Statistics in Medicine* 2003, **22**:2515-2527.
14. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143**:29-36.
15. Hanley JA, McNeil BJ: **A method of comparing the areas under receiver operating characteristic curves derived from the same cases.** *Radiology* 1983, **148**:839-843.
16. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P: **Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker.** *American Journal of Epidemiology* 2004, **159**:882-890.
17. Rosenblatt F: **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological Review* 1958, **65**:386-408.
18. Heine JJ, Carston MJ, Scott CG, Brandt KR, Wu FF, Pankratz VS, Sellers TA, Vachon CM: **An automated approach for estimation of breast density.** *Cancer Epidemiol Biomarkers Prev* 2008, **17**:3090-3097.
19. Manduca A, Carston MJ, Heine JJ, Scott CG, Pankratz VS, Brandt KR, Sellers TA, Vachon CM, Cerhan JR: **Texture features from mammographic images and risk of breast cancer.** *Cancer Epidemiol Biomarkers Prev* 2009, **18**:837-845.
20. Sackett DL, Haynes RB: **Evidence base of clinical diagnosis: the architecture of diagnostic research.** *British Medical Journal* 2002, **324**:539-541.
21. Elliott D: **Sigmoidal transformations and the trapezoidal rule.** *Journal of the Australian Mathematical Society B* 1998, **40(E)**:E77-E137.
22. Haykin S: *Neural Networks.* 2 edition. Upper Saddle River, NJ: Prentice Hall; 1999.
23. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis* Cambridge, UK Cambridge University Press; 2004.
24. Mercer J: **Functions of positive and negative type, and their connection with the theory of integral equations.** *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character* 1909, **209**:415-446.
25. Gretton A, Herbrich R, Smola A, Bousquet O, Scholkopf B: **Kernel methods for measuring independence.** *The Journal of Machine Learning Research* 2005, **6**:2075-2129.
26. Cacoullos T: **Estimation of a multivariate density.** *Annals of the Institute of Statistical Mathematics* 1966, **18**:179-189.
27. Parzen E: **On estimation of a probability density function and mode.** *Annals of Mathematical Statistics* 1962, **33**:1065-1076.
28. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap* Boca Raton, FL: Chapman & Hall; 1993.
29. Hosmer DW, Lemeshow S: *Applied Logistic Regression.* 2 edition. New York, NY: John Wiley & Sons, Inc; 2000.
30. Johnson GD, Eidson M, Schmit K, Ellis A, Kulldorff M: **Geographic prediction of human onset of West Nile virus using dead crow clusters: an evaluation of year 2002 data in New York State.** *American Journal of Epidemiology* 2006, **163**:171-180.
31. Kelsall JE, Diggle PJ: **Kernel estimation of relative risk.** *Bernoulli* 1995, **1**:3-16.

32. Yip PSF, Lau EHY, Lam KF, Huggins RM: **A chain multinomial model for estimating the real-time fatality rate of a disease, with an application to severe acute respiratory syndrome.** *American Journal of Epidemiology* 2005, **161**:700-706.
33. Price KV, Storn RM, Lampinen JA: *Differential Evolution: A Practical Approach to Global Optimization* Heidelberg: Springer; 2005.

doi:10.1186/1471-2105-12-37
Cite this article as: Heine *et al.*: Statistical learning techniques applied to epidemiology: a simulated case-control comparison study with logistic regression. *BMC Bioinformatics* 2011 **12**:37.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central