Convex Surrogate Minimization in Classification

Cui Xiong[1], Jun Shao[1,2] and Lei Wang[3]

[1]*East China Normal University,* [2]*University of Wisconsin-Madison and* [3]*Nankai University*

Abstract: Convex optimization has become an increasingly important theme in applications. We consider the construction of a binary classification rule by minimizing the risk based on a convex loss as a surrogate to the 0-1 loss. Compared with the approach of directly estimating the conditional probability of the binary class label given a vector of covariates, our proposed convex surrogate minimization approach is computationally simpler and more efficient because of the convexity. We begin with a rigorous discussion of what type of convex surrogate is valid. When the conditional probability model for class label is parametric, we show that our proposed approach is either equivalent to the traditional maximum likelihood method or a substitute for computational saving. When the conditional probability model is semiparametric, we show how to apply convex surrogate minimization in conjuncture with kernel weighting, which results in an asymptotically valid classification rule. Some convergence rates are established and empirical simulation results are presented.

Key words and phrases: Convex optimization, binary classification, kernel weighting, 0-1 loss.

1 Introduction

In many social, economical, biological, and medical studies, an important analysis is to classify a subject into one of two classes based on a set of covariates observed from the subject. Let $Y = 1$ and -1 be the labels of

the two classes and X be the p-dimensional vector of covariates observed from the subject. In most applications, variability exists and, hence, X and Y are random. A classification rule or discriminant function is any Borel function T from the range of X to $\{1, -1\}$ such that the subject is classified to class $T(X)$ when X is observed. A natural measure to evaluate the performance of a rule T is the misclassification rate $P(Y \neq T(X))$, where $P(\cdot)$ is the probability with respect to the distribution of (Y, X).

If the distribution of (Y, X) is known, then we can construct an optimal classification rule T^* that has the smallest possible misclassification rate. It is known that the optimal rule is $T^*(X) = \text{sign}\{P(Y = 1|X) - P(Y = -1|X)\}$, where $\text{sign}(x)$ is the sign of x and $P(Y = y|X)$ is the conditional probability given X. In practice, the distribution of (Y, X) is usually unknown, and classification rules are constructed using observations from a training sample, $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, which is an independent sample identically distributed as (Y, X). A statistical issue is how to use the training sample to construct a classification rule that has a misclassification rate close to that of the optimal rule.

Throughout this paper we assume a model

$$P(Y = 1|X) = g(\beta_0^T X), \tag{1}$$

where $\beta_0 = (\beta_{01}, ..., \beta_{0p})^T$ is an unknown p-dimensional vector, a^T denotes the transpose of a vector a, g is either a known distribution function (the parametric case) or an unspecified distribution function (the semiparametric case), and $g(0) = 1/2$ is assumed. Without loss of generality, we assume throughout that the first component of X is 1 so β_{01} is an intercept and that $p > 1$ and $\beta_{0p} \neq 0$ so there is at least one useful non-constant covariate. Under the monotonicity condition on g and $g(0) = 1/2$, the optimal rule $T^*(X) = \text{sign}(\beta_0^T X)$.

Since we know the form of the optimal rule T^*, one way to construct a classification rule based on data from the training sample is to substitute the unknown β_0 in the optimal rule by an estimator $\hat{\beta}$ based on data $(X_1, Y_1), ..., (X_n, Y_n)$. Although we don't need to worry about g in the optimal

rule, whether or not g is known actually affects the estimation of β_0. When g is known so that model (1) is parametric, β_0 can be estimated using the parametric method of maximum likelihood and the estimator is asymptotically (as $n \to \infty$) normal and optimal. When both g and β_0 are unknown, model (1) is referred to as the single-index model and β_0 may be estimated by the method of sliced inverse regression (SIR) proposed in Li (1991, 1992) , sliced average variance estimation (SAVE) proposed by Cook and Weisberg (1991), directional regression (DR) proposed in Li and Wang (2007), and extended principal fitted components (EPFC) and likelihood acquired directions (LAD) proposed in Cook and Forzani (2009). Although we can only estimate $c\beta_0$ using these methods under the single-index model, it is enough for classification because $\text{sign}(\beta_0^T X) = \text{sign}(c\beta_0^T X)$ for any $c > 0$. These estimation-based methods, however, rely on either a correct specification of the function g in the parametric case or a linearity assumption on X, i.e., $E(b^T X|\beta_0^T X)$ is linear in $\beta_0^T X$ for any p-dimensional vector b, which is satisfied if X is elliptically symmetric (Li, 1991).

The purpose of this paper is to study a different approach for the construction of a classification rule, a method that directly minimizes an estimated misclassification rate. We focus on rules having the form $\text{sign}(\beta^T X)$ for a p-dimensional vector β (which may not be β_0 and may depend on data). Let $\ell(\alpha)$ be the 0-1 loss function that is 1 if $\alpha \leq 0$ and 0 otherwise. Then the misclassification rate of any $T(X) = \text{sign}(\beta^T X)$ is the risk under the 0-1 loss,

$$R(\beta) = E\{\ell(Y\beta^T X)\} = P(Y \neq \text{sign}(\beta^T X)). \tag{2}$$

Note that the expectation and probability are with respect to the distribution of (Y, X). Thus, R is a function of β as well as β_0 and g, but we omit β_0 and g for simplicity. If β is an estimator, i.e., a function of the training data $\mathcal{T}_n = \{(Y_1, X_1), ..., (Y_n, X_n)\}$, then the expectation is taken conditional on $\mathcal{T}_n$, and $R(\beta)$ is a function of β_0, g and data $\mathcal{T}_n$. Given the training data $\mathcal{T}_n$ and a fixed β, it is natural to estimate $R(\beta)$ by the sample average

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i \beta^T X_i).$$

Then, we wish to find a classification rule with a β that minimizes the estimated risk $\hat{R}(\beta)$. However, such a procedure is computationally intractable (Arora et al., 1997). It was suggested that we might replace the 0-1 loss $\ell(\alpha)$ by a convex surrogate $\varphi(\alpha)$ so that the minimization is tractable, where φ is differentiable, decreasing, and strictly convex (Zhang, 2004; Bartlett et al., 2006; Nguyen et al., 2009). Thus, we define $R_\varphi(\beta) = E\{\varphi(Y\beta^T X)\}$ to be the φ-risk, and hope that the minimizer of $R_\varphi(\beta)$ is a reasonable surrogate for the minimizer of $R(\beta)$ and that we can construct a rule by minimizing the sample average

$$\hat{R}_\varphi(\beta) = \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i\beta^T X_i). \tag{3}$$

This method will be called the convex surrogate minimization (CSM). Although our primary goal is classification, the minimizer of the risk function in (3) will be called the CSM estimator of $c\beta_0$, where $c > 0$ and β_0 is in (1).

Convexity has become an increasingly important theme in applied mathematics and engineering (Boyd and Vandenberghe, 2004). One area in which this trend has been most salient is machine learning, where computational efficiency is imperative and many of the most prominent methods make significant use of convexity; for example, support vector machines (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002) and boosting (Collins et al., 2002; Lebanon and Lafferty, 2002).

In Section 2, we establish a sufficient condition on the validity of a convex surrogate φ in CSM in the sense that $R_\varphi(\beta)$ and $R(\beta)$ share the same minimizer, which leads to a CSM classification rule that converges to the optimal rule. This sufficient condition relates φ to the function g in (1). We show in Section 3 that, when g is known, a CSM estimator is the same as the classical maximum likelihood estimator (MLE) if the parametric likelihood is convex; otherwise the CSM estimator is different from the MLE but is computationally more efficient because of the convex optimization.

Section 4 is devoted to the case of unknown g, the single-index semiparametric model. When g is unknown, a valid convex surrogate φ is difficult to find. We show how to use an approximate convex surrogate φ that is based on

a truncated quadratic loss function. The quadratic loss function, however, depends on the derivative function of g. To avoid the estimation of the derivative function of g, we apply kernel weighting that enables us to use an approximate convex surrogate depending on the derivative of g at 0 only. The derivative of g at 0 does not need to be estimated since $\text{sign}(\beta_0^T X) = \text{sign}(c\beta_0^T X)$ for any $c > 0$. This method produces a classification rule that converges to the optimal rule under reasonable conditions. It does not rely on the linearity condition and is computationally simpler than the dimension-reduction methods such as SIR or SAVE. We obtain explicitly the asymptotic distribution of the proposed CSM estimator and the convergence rate of the CSM classification rule. Some simulation results are presented to show the finite sample performance of the CSM estimator and other estimators under the semiparametric setting. Technical details are given in the Supplementary Material.

2 Convex Surrogates

A basic requirement for a convex surrogate φ is that the φ-risk $R_\varphi(\beta) = E\{\varphi(Y\beta^T X)\}$ has the same minimizer as the 0-1 risk $R(\beta) = E\{\ell(Y\beta^T X)\}$. If this is true for a convex φ, then the CSM based on φ and the training data may produce a classification rule that converges to the optimal rule when the train sample size $n \to \infty$. We study this issue in two steps.

2.1 Validity of a convex surrogate

For any given φ and β,

$$E\{\varphi(Y\beta^T X)|X = x\} = g(\beta_0^T x)\varphi(\beta^T x) + \{1 - g(\beta_0^T x)\}\varphi(-\beta^T x)$$

is the conditional φ-risk, where β_0 is the true parameter value in (1). With $\alpha_0 = \beta_0^T x$ and $\alpha = \beta^T x$, the conditional φ-risk in the previous express can be written as

$$C(\alpha) = g(\alpha_0)\varphi(\alpha) + \{1 - g(\alpha_0)\}\varphi(-\alpha), \tag{4}$$

which is called the generic conditional φ-risk by Bartlett et al. (2006). If a convex function φ has the property that α_0 is a unique minimizer of the generic conditional φ-risk $C(\alpha)$ defined by (4), then φ is called a valid convex surrogate. The following result shows the existence of a valid convex surrogate.

Theorem 1. *For any continuous g satisfying $0 < g(\alpha) = 1 - g(-\alpha) < 1$ for $\alpha \in (-\alpha_b, \alpha_b)$ and $g(\alpha) \neq g(\alpha_0)$ for any $\alpha \neq \alpha_0$, where $g(\alpha_b) = 1$ and $g(-\alpha_b) = 0$ (α_b may be infinity), there exists a decreasing valid convex surrogate that depends on g.*

To establish this result, we construct a decreasing convex function φ of the form

$$\varphi(\alpha) = \begin{cases} \zeta(\alpha) & 0 < \alpha < \alpha_b \\ \lambda(-\alpha) & -\alpha_b < \alpha \leq 0 \end{cases} \tag{5}$$

where α_b is the upper bound for α, λ is an increasing, convex and differentiable function, ζ is a decreasing and differentiable function with $\zeta(0) = \lambda(0)$, and both ζ and λ are defined on $[0, \alpha_b)$. Given λ, we hope to find a decreasing convex function ζ such that the solution of $C'(\alpha) = 0$ is $\alpha = \alpha_0$, where C' is the derivative of the function C in (4). From the definition of C in (4) and φ in (5),

$$C'(\alpha) = \begin{cases} g(\alpha_0)\zeta'(\alpha) + \{1 - g(\alpha_0)\}\lambda'(\alpha) & 0 < \alpha < \alpha_b \\ -g(\alpha_0)\lambda'(-\alpha) - \{1 - g(\alpha_0)\}\zeta'(-\alpha) & -\alpha_b < \alpha \leq 0 \end{cases}$$

If $C'(\alpha_0) = 0$, then

$$0 = \begin{cases} g(\alpha_0)\zeta'(\alpha_0) + \{1 - g(\alpha_0)\}\lambda'(\alpha_0) & 0 < \alpha_0 < \alpha_b \\ -g(\alpha_0)\lambda'(-\alpha_0) - \{1 - g(\alpha_0)\}\zeta'(-\alpha_0) & -\alpha_b < \alpha_0 \leq 0 \end{cases}$$

Note that $g(\alpha) = 1 - g(-\alpha)$. Thus, as long as λ is chosen such that $\frac{g-1}{g}\lambda'$ is integrable, the relationship between ζ and λ is

$$\zeta'(\alpha) = \frac{g(\alpha) - 1}{g(\alpha)}\lambda'(\alpha) \qquad \text{or} \qquad \zeta(\alpha) = \int \frac{g(\alpha) - 1}{g(\alpha)}\lambda'(\alpha)d\alpha. \tag{6}$$

The function ζ is convex if and only if $\frac{g(\alpha)-1}{g(\alpha)}\lambda'(\alpha)$ is an increasing function. Given g, such a function λ can be easily constructed (see Examples 1-2). Also,

since $\lambda' > 0$, ζ is a decreasing and convex function and

$$\zeta'(0) = \frac{g(0)-1}{g(0)}\lambda'(0) = -\lambda'(0),$$

so that φ is differentiable and continuous at $\alpha = 0$.

Without loss of generality, assume $\alpha_0 > 0$. With the chosen λ and φ given by (6), if $0 < \alpha < \alpha_b$,

$$C'(\alpha) = \lambda'(\alpha)\left\{g(\alpha_0)\frac{g(\alpha)-1}{g(\alpha)} + 1 - g(\alpha_0)\right\} = \frac{\lambda'(\alpha)}{g(\alpha)}\{g(\alpha) - g(\alpha_0)\},$$

which obviously equals 0 at $\alpha = \alpha_0$. Furthermore, since $\lambda'(\alpha) > 0$ and $g(\alpha) \neq g(\alpha_0)$ for any $\alpha \neq \alpha_0$, α_0 is the unique solution to $C'(\alpha) = 0$ for $0 < \alpha < \alpha_b$. If $-\alpha_b < \alpha \leq 0$, then

$$\begin{aligned} C'(\alpha) &= -g(\alpha_0)\lambda'(-\alpha) - \{1 - g(\alpha_0)\}\zeta'(-\alpha) \\ &= -\frac{\lambda'(-\alpha)}{g(-\alpha)}\left[g(-\alpha) - \{1 - g(\alpha_0)\}\right] \\ &= -\frac{\lambda'(-\alpha)}{g(-\alpha)}\left[g(\alpha_0) - g(\alpha)\right] < 0. \end{aligned}$$

Hence, α_0 is the unique minimizer of $C(\alpha)$ and φ in (5)-(6) is valid.

To summarize, under the parametric case, we need to find some increasing, convex, and differentiable function λ such that $\frac{g-1}{g}\lambda'$ is an increasing and integrable function. A valid surrogate φ can then be obtained by (5)-(6). For a given g, there may be many valid convex surrogates.

We now consider two examples.

Example 1 (Logistic model).

Under the logistic model, $g(\alpha) = \exp(\alpha)/\{1+\exp(\alpha)\}$, $\alpha \in (-\alpha_b, \alpha_b)$ with $\alpha_b = \infty$. We first consider $\lambda(\alpha) = \exp(\alpha)$. Then $\zeta'(\alpha) = \frac{g(\alpha)-1}{g(\alpha)}\lambda'(\alpha) = -1$ is a constant, and

$$\varphi(\alpha) = \begin{cases} -\alpha + 1 & \alpha \geq 0 \\ \exp(-\alpha) & \alpha < 0 \end{cases}$$

is a convex function. Without loss of generality, consider $\alpha_0 > 0$ and $\alpha > 0$. Then $\alpha = \alpha_0$ is a unique solution to

$$C'(\alpha) = -\frac{\exp(\alpha_0)}{1+\exp(\alpha_0)} + \frac{1}{1+\exp(\alpha_0)}\exp(\alpha) = 0.$$

Next, we consider $\lambda(\alpha) = \exp(\alpha/2)$. Then $\zeta'(\alpha) = -\exp(-\alpha/2)/2$, and $\varphi(\alpha) = \exp(-\alpha/2)$, which is the exponential loss function in Zhang (2004). Finally, we consider $\lambda(\alpha) = \exp(\alpha/3)$. Then $\zeta'(\alpha) = -\exp(-2\alpha/3)/3$, and

$$\varphi(\alpha) = \begin{cases} \frac{1}{2}\exp\left(-\frac{2\alpha}{3}\right) + \frac{1}{2} & \alpha \geq 0 \\ \exp\left(-\frac{\alpha}{3}\right) & \alpha < 0 \end{cases}$$

Example 2 (Truncated linear model). Consider

$$g(\alpha) = \begin{cases} 0 & \alpha < -\frac{1}{2b} \\ \frac{1}{2} + b\alpha & -\frac{1}{2b} \leq \alpha \leq \frac{1}{2b} \\ 1 & \alpha > \frac{1}{2b} \end{cases}$$

where b is a constant. In this case, $\alpha_b = 1/(2b)$. Let $\lambda(\alpha) = (b\alpha + 1/2)^2$. Then $\zeta'(\alpha) = 2b(b\alpha - 1/2)$ is increasing in α, and $\varphi(\alpha) = (1/2 - b\alpha)^2$ for $\alpha \leq 1/(2b)$. Since $\zeta'(1/(2b)) = 0$, $\zeta(\alpha) = 0$ for all $\alpha > 1/(2b)$. Hence,

$$\varphi(\alpha) = \begin{cases} \left(\frac{1}{2} - b\alpha\right)^2 & \alpha \leq \frac{1}{2b} \\ 0 & \alpha > \frac{1}{2b} \end{cases} \tag{7}$$

which is truncated quadratic loss in Bartlett et al. (2006).

2.2 The risk function and its minimizers

The risk function for the classification rule $T(X) = \text{sign}(\beta^T X)$ under the 0-1 loss is

$$\begin{aligned} R(\beta) &= E\{\ell(Y\beta^T X)\} = E[E\{\ell(Y\beta^T X)|X\}] \\ &= E\left[P(Y = 1, \beta^T X \leq 0|X) + P(Y = -1, \beta^T X \geq 0|X)\right] \\ &= \int_{\beta^T x \leq 0} g(\beta_0^T x) dF(x) + \int_{\beta^T x > 0} \{1 - g(\beta_0^T x)\} dF(x) \\ &= \int_{\beta^T x \leq 0} \{2g(\beta_0^T x) - 1\} dF(x) + \int \{1 - g(\beta_0^T x)\} dF(x), \end{aligned} \tag{8}$$

where F is the joint cumulative distribution function of X. Figure 1 shows a $3d$ plot of the function $R(\beta)$ for $g(\beta^T X) = \exp(\beta_1 x_1 + \beta_2 x_2)/\{1 + \exp(\beta_1 x_1 +$

$\beta_2 x_2)\}$, and (x_1, x_2) is bivariate normal with mean zero and identity matrix as the covariance matrix. From Figure 1, we can see that, when $\beta_1 = \beta_2$, the risk function $R(\beta)$ is minimized.

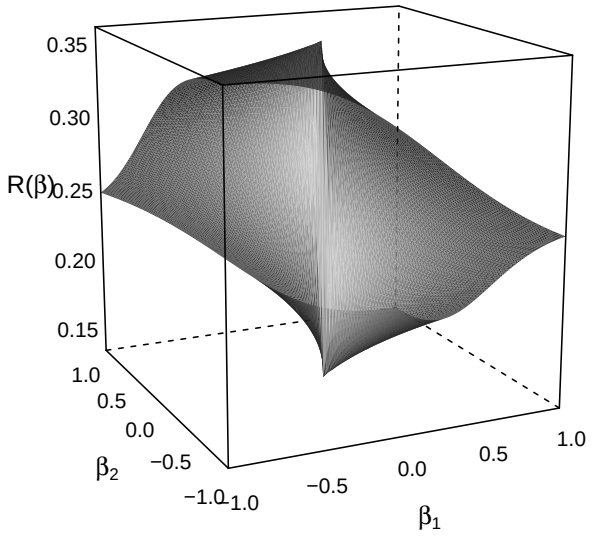

Figure 1: The plot of 0-1 risk $R(\beta)$

In general, it can be seen from (8) that if we choose $\beta = c\beta_0$ with a constant $c > 0$, then the first integral on the right side of (8) is 0 and $R(\beta)$ reaches its minimum $\int\{1 - g(\beta_0^T x)\}dF(x)$, i.e., $c\beta_0$ is a minimizer of $R(\beta)$ for any $c > 0$ (the minimizer is not unique). The following lemma provides more information. It shows that $R(\beta)$ is differentiable at $\beta = c\beta_0$ even though the 0-1 loss ℓ is not always differentiable. The proof is given in the Supplementary Material.

Lemma 1. *Assume that the non-constant part of X has a density f, and f and g are continuous. Then, $R(\beta)$ is differentiable at $\beta = c\beta_0$ for any constant $c > 0$ and $\partial R(\beta)/\partial\beta|_{\beta=c\beta_0} = 0$.*

We now turn to the φ-risk for convex surrogate φ, $R_\varphi(\beta) = E\{\varphi(Y\beta^T X)\}$. Following the discussion in Section 2.1, we define $\Psi(\beta_0)$ to be a set of valid

convex surrogates φ, i.e.,

$$\Psi(\beta_0) = \{\varphi : \text{convex, differentiable, for any } x \quad g(\beta_0^T x)\varphi'(\beta_0^T x) + \{1 - g(\beta_0^T x)\}\varphi'(-\beta_0^T x) = 0\}. \tag{9}$$

From Theorem 1, $\Psi(\beta_0)$ is not empty. Unlike the 0-1 risk, for any $\varphi \in \Psi(\beta_0)$, the minimizer of φ-risk is unique. The proof of the following result is given in Supplementary Material.

Lemma 2. *Let $\varphi \in \Psi(\beta_0)$ defined by (9) with finite $\int \sup_{\beta\in\mathcal{N}} |\varphi'(\beta^T x)|xdF(x)$ and $\int \sup_{\beta\in\mathcal{N}} |\varphi''(\beta^T x)|x^T xdF(x)$, where $\mathcal{N}$ is a neighborhood of β_0. Then, the unique minimizer of $R_\varphi(\beta)$ over β is at $\beta = \beta_0$.*

From Lemmas 1 and 2, we have the following conclusions.

1. By Lemma 1, the minimizers of the 0-1 risk $R(\beta)$ form a set $\{c\beta_0 : c > 0\}$.
2. By Lemma 2, the minimizer of $R_\varphi(\beta)$ is unique for each $\varphi \in \Psi$ under minor conditions.
3. For any positive constant c, $c\beta_0$ is actually a minimizer of the φ_c-risk with $\varphi_c(\alpha) = \varphi(c\alpha)$, which is a valid convex surrogate if φ is a valid convex surrogate.

These conclusions help us to simplify the problem in the semiparametric case by ignoring an unknown constant (slope).

3 Applications in Parametric Models

When g is known, model (1) is parametric and we can apply the maximum-likelihood estimation (MLE) to estimate the unknown β_0. The likelihood based on training data (Y_i, X_i), $i = 1, ..., n$, is

$$L(\beta) = \prod_{i=1}^{n} g\left(\beta^T X_i\right)^{\frac{1+Y_i}{2}} \left\{1 - g\left(\beta^T X_i\right)\right\}^{\frac{1-Y_i}{2}},$$

which is actually the joint probability mass function of $Y_1, ..., Y_n$ conditioned on X_i, $i = 1, ..., n$. The classical asymptotic theory shows that the MLE is asymptotically normal with mean β_0 and asymptotic covariance matrix

$$\frac{1}{n}\left[E\left\{XX^T\frac{g'^2(\beta_0^TX)}{g(\beta_0^TX)g(-\beta_0^TX)}\right\}\right]^{-1} \tag{10}$$

under the conditions on g in Theorem 1, where g' is the derivative of g, and the MLE is optimal in the sense that any other asymptotically normal estimator of β_0 has asymptotic covariance matrix no smaller than that in (10) in terms of matrix ordering.

On the other hand, the CSM estimator based on a convex surrogate φ is obtained by minimizing the empirical φ-risk defined in (3), i.e., minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i\beta^TX_i).$$

Let φ be given by (5) with $\zeta'(\alpha) = \frac{g(\alpha)-1}{g(\alpha)}\lambda'(\alpha)$. Then

$$\varphi(Y_i\beta^TX_i) = \begin{cases} \frac{1+Y_i}{2}\zeta(\beta^TX_i) + \frac{1-Y_i}{2}\lambda(\beta^TX_i) & \text{if } \beta^TX_i > 0 \\ \frac{1+Y_i}{2}\lambda(-\beta^TX_i) + \frac{1-Y_i}{2}\zeta(-\beta^TX_i) & \text{if } \beta^TX_i \le 0 \end{cases}$$

and

$$\begin{aligned} &\frac{\partial\varphi(Y_i\beta^TX_i)}{\partial\beta} \\ =&\begin{cases} \frac{1+Y_i}{2}X_i\frac{g(\beta^TX_i)-1}{g(\beta^Tx_i)}\lambda'(\beta^TX_i) + \frac{1-Y_i}{2}X_i\lambda'(\beta^TX_i) & \text{if } \beta^TX_i > 0 \\ -\frac{1+Y_i}{2}X_i\lambda'(-\beta^TX_i) - \frac{1-Y_i}{2}X_i\frac{g(-\beta^TX_i)-1}{g(-\beta^TX_i)}\lambda'(-\beta^TX_i) & \text{if } \beta^TX_i \le 0 \end{cases} \\ =&\frac{X_i\lambda'(|\beta^TX_i|)}{2g(|\beta^TX_i|)}\left[2g(\beta^TX_i) - 1 - Y_i\right]. \end{aligned}$$

Consequently, the CSM estimator is obtained by solving

$$H(\beta) = \frac{\partial}{\partial\beta}\left\{\sum_{i=1}^{n}\varphi(Y_i\beta^TX_i)\right\} = \sum_{i=1}^{n}\frac{X_i\lambda'(|\beta^TX_i|)}{2g(|\beta^TX_i|)}\left[2g(\beta^TX_i) - 1 - Y_i\right] = 0.$$

Since $H(\beta)$ is a sum of independent and identically distributed random vectors, the solution is asymptotically normal with mean β_0 and asymptotic covariance matrix

$$n^{-1}[E\{H'(\beta_0)\}]^{-1}\mathrm{Cov}\{H(\beta_0)\}[E\{H'(\beta_0)\}]^{-1}, \tag{11}$$

where $H'(\beta) = \partial H(\beta)/\partial\beta$,

$$E\{H'(\beta_0)\} = E\left\{XX^T \frac{\lambda'(|\beta_0^T X|)g'(\beta_0^T X)}{g(|\beta_0^T X|)}\right\}$$

and

$$\mathrm{Cov}\{H(\beta_0)\} = E\left\{XX^T \frac{\lambda'^2(|\beta_0^T X|)g(\beta_0^T X)g(-\beta_0^T X)}{g^2(|\beta_0^T X|)}\right\}.$$

Comparing (10) and (11), we conclude that, the matrices in (10) and (11) are the same if and only if the λ function satisfies $\lambda'(\alpha) = \frac{g'(\alpha)}{1-g(\alpha)}$, i.e., $\varphi(\alpha) = -\log\{g(\alpha)\}$, provided that $-\log\{g(\alpha)\}$ is convex.

Conclusion 1. When $-\log\{g\}$ is a convex function and g satisfies the conditions in Theorem 1, the CSM with $\varphi = -\log\{g\}$ is equivalent to the MLE. The CSM estimators with other φ's have asymptotic covariance matrices given by (11) and are asymptotically less efficient than the MLE, but may be computationally more efficient.

The fact that the CSM can be used as a substitute for MLE for computational saving is shown in the following examples.

Example 1 (continued). The most popular parametric model is the logistic model with $g(\alpha) = \exp(\alpha)/\{1+\exp(\alpha)\}$. If model (1) is viewed as a generalized linear model, then the logistic model corresponds to the canonical link (McCullagh and Nelder, 1989). It is well known that $-\log$ of likelihood is convex in this case. In fact, $-\log\{g(\alpha)\} = \log\{1+\exp(-\alpha)\}$ is convex. The CSM with $\varphi(\alpha) = \log\{1+\exp(-\alpha)\}$ is the same as the MLE.

Consider the CSM with the convex surrogate $\varphi(\alpha) = \exp(-\alpha/2)$ given in Example 1 in Section 2. From the previous discussion, this CSM is not equivalent to the MLE. We compare in a simulation the MLE and CSM with $\varphi(\alpha) = \exp(-\alpha/2)$ in a logistic model with a 5-dimensional $\beta_0 = (\beta_{01}, ..., \beta_{05})^T$ and $X = (1, \xi_2, ..., \xi_5)^T$, where $\xi_2, ..., \xi_5$ are independent standard normal random variables. The sample size is $n = 200$ and the simulation size is 1,000. Table 1 shows the means and root mean squared errors (rmse) of the estimated ratios β_{0j}/β_{01}, $j = 2, ..., 5$. The true values of the ratios and the time used to compute the estimators are also included in the table. All estimators are computed using the R function "optimize".

Table 1: Simulation mean, rmse, and time under the logistic model

simulation size $= 1,000$

		$n = 200$	
true ratio		CSM	MLE
$\beta_{02}/\beta_{01} = 0.5$	mean	0.524	0.521
	rmse	0.210	0.204
$\beta_{03}/\beta_{01} = 1$	mean	1.036	1.030
	rmse	0.256	0.248
$\beta_{04}/\beta_{01} = 0.5$	mean	0.529	0.526
	rmse	0.213	0.211
$\beta_{05}/\beta_{01} = 0$	mean	0.001	0.003
	rmse	0.196	0.188
	time in seconds	49.87	274.28

From Table 1, both the MLE and CSM estimator are almost unbiased and the CSM estimator has a larger but comparable rmse. However, the CSM is much faster.

Example 1A (probit model). Next, we consider probit model, i.e., $g(\alpha)$ in (1) is $\Phi(\alpha)$, the standard normal cumulative distribution function. We now show that $\varphi(\alpha) = -\log \Phi(\alpha)$ is convex so that the conclusions are the same as those in the logistic model, i.e., the CSM with $\varphi(\alpha) = -\log \Phi(\alpha)$ is the same as the MLE. Note that

$$\varphi'(\alpha) = -\frac{\Phi'(\alpha)}{\Phi(\alpha)} \qquad \text{and} \qquad \varphi''(\alpha) = -\frac{\Phi''(\alpha)\Phi(\alpha) - \{\Phi'(\alpha)\}^2}{\{\Phi(\alpha)\}^2}.$$

From the property of the normal distribution, $\Phi''(\alpha) = -\alpha\Phi'(\alpha)$. Then

$$\varphi''(\alpha) = \frac{\Phi'(\alpha)}{\{\Phi(\alpha)\}^2}\{\alpha\Phi(\alpha) + \Phi'(\alpha)\} = \frac{\Phi'(\alpha)}{\{\Phi(\alpha)\}^2}\int_{-\infty}^{\alpha} \Phi(t)dt > 0,$$

which proves that $\varphi(\alpha) = -\log \Phi(\alpha)$ is convex.

When $-\log$ of the likelihood is not convex, CSM estimators are different from the MLE. We consider the following example.

Example 2 (continued). Consider the truncated linear model

$$P(Y=1|X) = \begin{cases} 0 & \beta_0^T X \le -\frac{1}{2} \\ \frac{1}{2} + \beta_0^T X & -\frac{1}{2} < \beta_0^T X \le \frac{1}{2} \\ 1 & \beta_0^T X > \frac{1}{2} \end{cases}$$

with $\beta_0 = (\beta_{01}, ..., \beta_{05})^T = (0.25, 0.5, 0.25, 0.5, 0.5)^T$ and X the same as that in Example 1. In this case, $-\log$ of likelihood is not convex. As discussed in Section 2.1, we can use the truncated quadratic loss in (7) (Bartlett et al., 2006) as the convex surrogate. The CSM and MLE are different. The simulation results given in Table 2 show that the MLE is worse than the CSM when $n = 200$ in terms of rmse. This is because, although the MLE is asymptotically optimal, when $-\log\{g(\alpha)\}$ is not convex, the MLE requires a large n to appreciate the asymptotic effect. When $n = 200$, the MLE is still not stable and, hence, has a larger variability than the CSM estimator. To show this we run an additional simulation with $n = 500$. The results are included in Table 2 and show that the MLE is slightly better than the CSM. On the other hand, the computational gain in using the CSM is more in this case because the MLE is not a convex optimization.

Conclusion 2. When $-\log\{g(\alpha)\}$ is not convex, the CSM can be used as a substitute for the MLE for computational saving. Because of the complexity in computing the MLE, the CSM estimators may be better than the MLE with not very large n, although they are asymptotically less efficient than the MLE.

When $-\log\{g(\alpha)\}$ is not convex, there may not exist an optimal CSM estimator in terms of either estimation efficiency or computation complexity. Because the CSM is mainly for computational saving, it is not necessary to find a φ having the optimal computation efficiency. Using the asymptotic covariance matrix given by (11), we may perform some numerical studies to choose a surrogate φ having reasonable estimation efficiency. For example, in

Table 2: Simulation mean, rmse, and time under the truncated linear model

simulation size $= 1,000$

		$n = 200$		$n = 500$	
true ratio		CSM	MLE	CSM	MLE
$\beta_{02}/\beta_{01} = 2$	mean	2.075	2.095	2.025	2.018
	rmse	0.489	0.588	0.257	0.243
$\beta_{03}/\beta_{01} = 1$	mean	1.044	1.060	1.008	1.006
	rmse	0.305	0.355	0.159	0.142
$\beta_{04}/\beta_{01} = 2$	mean	2.086	2.110	2.022	2.016
	rmse	0.504	0.609	0.258	0.251
$\beta_{05}/\beta_{01} = 2$	mean	2.083	2.108	2.022	2.012
	rmse	0.510	0.613	0.259	0.251
	time in seconds	53.22	387.55	101.56	829.39

the three examples in this section, we have found some good surrogates.

4 Applications in Semiparametric Models

We now consider model (1) with unknown g and β_0, which is a semiparametric model. Although Theorem 1 shows the existence of a valid convex surrogate, when g is unknown, it is difficult to find a valid convex surrogate. We propose a method of constructing an approximate convex surrogate and establish its asymptotic properties in Section 4.1, and then provide some empirical results in Section 4.2.

4.1 Method and theory

Our idea is to first linearize g at 0 and consider the truncated quadratic loss in (7), which produces a valid convex surrogate φ when g is actually linear. After obtaining a φ, we apply kernel weighting to overcome the difficulty that g is not linear.

Consider the Taylor expansion of $g(\alpha)$ at $\alpha = 0$,

$$g(\alpha) \approx g(0) + g'(0)\alpha. \tag{12}$$

Although $g(0) = 1/2$, the derivative $g'(0)$ is unknown. From the discussion in the end of Section 2.2, if $\varphi(\alpha)$ is valid, so is $\varphi(c\alpha)$ for any $c > 0$, i.e., estimating $g'(0)\beta_0$ is sufficient for the purpose of classification. Thus, we do not need to estimate the unknown $g'(0)$. If g is actually linear, then $\approx$ in (12) becomes an equality, and based on Example 2, we can use the following valid convex surrogate

$$\tilde{\varphi}(\alpha) = \begin{cases} \left(\frac{1}{2} - \alpha\right)^2 & \alpha \le \frac{1}{2} \\ 0 & \alpha > \frac{1}{2} \end{cases} \tag{13}$$

However, (12) is an approximation and it only holds for α near 0 when g is not linear. Therefore, minimizing the $\tilde{\varphi}$-risk $\hat{R}_{\tilde{\varphi}}(\beta) = n^{-1}\sum_{i=1}^n \tilde{\varphi}(Y_i\beta^T X_i)$ may not lead to a satisfactory result since $Y_i\beta^T X_i$ may not be close to 0 for all i. To overcome this difficulty, we apply kernel weighting in conjuncture with the convex surrogate $\tilde{\varphi}$ in (13). Note that minimizing $\sum_{i=1}^n \tilde{\varphi}(Y_i\beta^T X_i)$ is the same as solving

$$\frac{1}{n}\sum_{i=1}^n \tilde{\varphi}'(Y_i\beta^T X_i)Y_i X_i = 0. \tag{14}$$

Let K be a symmetric probability density function (called a kernel) with support $[-1, 1]$ and

$$B_K = \int_{-1}^1 u^2 K(u)du < \infty \qquad \text{and} \qquad V_K = \int_{-1}^1 K^2(u)du < \infty,$$

and let $h > 0$ be a bandwidth and $K_h(t) = K(t/h)/h$. Then, we replace (14) by the following kernel weighted version,

$$\frac{1}{n}\sum_{i=1}^n \tilde{\varphi}'(Y_i\beta^T X_i)Y_i X_i K_h(\beta^T X_i) = 0. \tag{15}$$

By suitably choosing the bandwidth h, the solution of (15) is asymptotically valid.

Solving (15) is not a convex optimization since the weight $K_h(\beta^T X_i)$ involves β. We have to apply some algorithms such as Newton's method. This is the price we pay for using a semiparametric model.

The solution to (15) estimates $g'(0)\beta_0$, $g'(0) > 0$. We have the following result whose proof is given in Supplementary Material.

Theorem 2. *Let $\hat{\beta}$ the solution of (15). Assume the parameter space for β^0 is a compact set; model (1) holds with an unknown g that is third order differentiable with bounded third order derivative; the kernel K satisfies the previous stated conditions, $h \to 0$, $nh \to \infty$ and $nh^5 = O(1)$ as $n \to \infty$; and the non-constant part of X has a continuous density $f > 0$. Assume further that the matrix*

$$D = \frac{1}{|\beta_{0p}| g'(0)} \int \begin{pmatrix} 1 & z^T \\ z & zz^T \end{pmatrix} f(z)\, dx_2 \cdots dx_{p-1} \tag{16}$$

is positive definite, where z is the $(p-1)$-dimensional vector whose first $p-2$ components are $x_2, ..., x_{p-1}$ and the last component is $-(\beta_{01} + \beta_{02}x_2 + \cdots + \beta_{0(p-1)}x_{p-1})/\beta_{0p}$, and β_{0j} is the jth component of β_0. Then, as $n \to \infty$, we have the following conclusions.

(i) $\hat{\beta}$ converges in probability to $g'(0)\beta_0$.

(ii) If $nh^5 \to 0$, then $(nh)^{1/2}\{\hat{\beta} - g'(0)\beta_0\}$ converges in distribution to the p-dimensional multivariate normal distribution with mean 0 and covariance matrix $V_K D^{-1}/4$.

(iii) Based on the asymptotic mean squared error (MSE) of $\hat{\beta} - g'(0)\beta_0$, the optimal choice of h is $h \asymp n^{-1/5}$, where $a \asymp b$ means $a = O(b)$ and $b = O(a)$.

(iv) Let $R(\beta)$ be the 0-1 risk defined in (2). Assume that the density f is continuously differentiable. Then, $R(\hat{\beta}) = R(\beta_0) + O_p(n^{-4/5})$ when $h \asymp n^{-1/5}$.

In applications we need to choose a bandwidth h for a fixed sample size n. There is a rich literature on bandwidth selection in applying a kernel method. A popular method is to apply the cross-validation, which works by leaving out one data point at a time, and choosing the value of h that minimizes

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i \hat{\beta}_{(-i)}^T X_i),$$

where $\hat{\beta}_{(-i)}$ is the solution to (15) with the ith term in the summation deleted. Note that $\ell(Y_i \hat{\beta}_{(-i)}^T X_i)$ is the loss when we classify the ith subject in the training data set by using the CSM classification rule based on the data set with (Y_i, X_i) removed from the original data set. Thus, $\mathrm{CV}(h)$ quantifies the classification accuracy of the CSM based on h.

4.2 Simulation results

This section presents some results from two simulation studies under the semiparametric model (1) with unknown g and β_0, to illustrate the finite sample performance of the CSM and compare it with the five estimation-based methods described in Section 1, the SIR, SAVE, DR, EPFC, and LAD. These five are dimension-reduction methods, which can be applied by using the available LDR package in Matlab.

In the first simulation study we generate $X_1, ..., X_n$ from a multivariate normal distribution so that the linearity condition described in Section 1 is satisfied and all estimators are asymptotically normal with mean $c\beta_0$ for some $c > 0$. In the second simulation study $X_1, ..., X_n$ are generated from a distribution that does not satisfy the linearity condition. In both cases all components of X are non-constant, i.e., we know that the intercept is 0.

4.2.1 Results under a multivariate normal X

In the first simulation study, $P(Y = 1|X) = \exp(\beta_0^T X)/[1 + \exp(\beta_0^T X)]$ and $X \sim N_5(0, I_5)$, where $\beta_0 = (2, 1, 2, 1, 0)^T$ is 5-dimensional and I_5 is the identity

matrix of order 5. From the previous discussion we know that we can only estimate $c\beta_0$ and it is enough to estimate $c\beta_0$ for classification. Thus, we consider the estimation of the ratios β_{0j}/β_{01}, $j = 2, ..., 5$ with the true value $(0.5, 1, 0.5, 0)$, where β_{0j} is the jth component of β_0.

The results in Table 3 are based on $n = 100$ and the simulation size 1,000. We calculate the means and root mean squared errors (rmse) of estimators of β_{0j}/β_{01} based on the SIR, SAVE, DR, EPFC, LAD, and CSM. To evaluate the performance of the classification, in each simulation run we generate another sample $\{(\tilde{Y}_1, \tilde{X}_1), ..., (\tilde{Y}_m, \tilde{X}_m)\}$ of size $m = 50$ from the distribution of (Y, X), independent of $\{(Y_1, X_1), ..., (Y_n, X_n)\}$, and calculate the sample average of the loss, $\hat{R}(\hat{\beta}) = m^{-1} \sum_{i=1}^{m} \ell(\tilde{Y}_i \hat{\beta}^T \tilde{X}_i)$, for $\hat{\beta}$ obtained using each method under consideration, which can be treated as an estimate of the misclassification rate by using $\hat{\beta}$. The simulation mean and rmse of $\hat{R}(\hat{\beta})$ for each method is included in Table 3.

From this table, all the methods have good performances except that the SAVE has a larger rmse and slightly worse misspecification rate than the other methods.

4.2.2 Results under an asymmetric X

In the second simulation study, $P(Y = 1|X) = \Phi(\beta_0^T X)$ and each component of X has an asymmetric mixture distribution $\frac{1}{2}N(1, 2) + \frac{1}{2}N\left(-\frac{1}{2}, 1\right)$. The vector β_0 is 6-dimensional with the true ratios $(0.5, 1, 0.5, 0, 0)$. The simulation results are given in Table 4.

Under this setting, the linearity condition described in Section 1 does not hold and the asymptotic behaviors of the five methods, SIR, SAVE, DR, EPFC, and LAD, are unknown. On the other hand, Theorem 2 for the CSM does not require the linearity condition and, from Table 4, it is better than the other five methods in terms of rmse. The SAVE has the a substantially larger rmse than all other methods. The SAVE also has the worst misspecification rate. The CSM has the best misspecification rate although the SIR, DR,

Table 3: Simulation results for the first simulation study

simulation size $= 1,000$

		$n = 100$					
true value		SIR	SAVE	DR	EPFC	LAD	CSM
$\beta_{02}/\beta_{01} = 0.5$	mean	0.512	0.544	0.516	0.510	0.511	0.509
	rmse	0.199	0.408	0.221	0.208	0.202	0.198
$\beta_{03}/\beta_{01} = 1$	mean	1.028	1.003	1.027	1.030	1.024	1.030
	rmse	0.283	0.976	0.301	0.298	0.280	0.278
$\beta_{04}/\beta_{01} = 0.5$	mean	0.516	0.493	0.515	0.518	0.519	0.517
	rmse	0.213	0.702	0.218	0.220	0.217	0.202
$\beta_{05}/\beta_{01} = 0$	mean	0.002	-0.044	-0.000	0.002	0.001	0.003
	rmse	0.171	0.687	0.192	0.181	0.178	0.162
$R(\beta_0) = 0.157$	mean	0.168	0.182	0.169	0.169	0.168	0.168
	rmse	0.052	0.069	0.052	0.052	0.052	0.051

EPFC, and LAD have misspecification rate close to that of the CSM.

Supplementary Material

The supplementary material consists all theoretical proofs of Lemmas 1-2 and Theorem 2.

Acknowledgements

The authors would like to thank two referees and an associate editor for helpful comments and suggestions. The first and second authors' research was par-

Table 4: Simulation results for the second simulation study

simulation size $= 1,000$

		$n = 100$					
true value		SIR	SAVE	DR	EPFC	LAD	CSM
$\beta_{02}/\beta_{01} = 0.5$	mean	0.545	0.488	0.544	0.547	0.523	0.524
	rmse	0.181	0.800	0.185	0.186	0.169	0.141
$\beta_{02}/\beta_{01} = 1$	mean	1.022	1.056	1.028	1.020	1.025	1.020
	rmse	0.204	1.455	0.212	0.214	0.205	0.172
$\beta_{02}/\beta_{01} = 0.5$	mean	0.540	0.454	0.536	0.541	0.513	0.516
	rmse	0.172	1.297	0.173	0.177	0.164	0.139
$\beta_{02}/\beta_{01} = 0$	mean	0.007	0.001	0.004	0.009	0.006	0.020
	rmse	0.139	0.337	0.152	0.144	0.128	0.111
$\beta_{02}/\beta_{01} = 0$	mean	-0.004	-0.020	-0.008	-0.005	0.002	0.017
	rmse	0.141	0.563	0.156	0.144	0.129	0.112
$R(\beta_0) = 0.064$	mean	0.083	0.104	0.085	0.085	0.082	0.076
	rmse	0.043	0.079	0.043	0.043	0.041	0.039

tially supported by the Chinese 111 Project B14019 and the US National Science Foundation Grants DMS-1305474 and DMS-1612873. The last author's research was supported by the National Natural Science Foundation of China (11501208) and Postdoctoral Science Foundation of China (2014M560317).

References

Arora, S., Babai, L., Stern, J. and Sweedyk, Z. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences* **54**, 317-331.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classifi-

cation, and risk bounds. *Journal of the American Statistical Association* **101**, 138-156.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Collins, M., Schapire, R. E. and Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning* **48**, 253-285.

Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association* **104**, 197-208.

Cook, R. D. and Weisberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 328-332.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Lebanon, G. and Lafferty, J. (2002). Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems* **14**, 447-454.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997-1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-327.

Li, K. C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association* **87**, 1025-1039.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (Second edition)*. Chapman and Hall/CRC, London.

Nguyen, X. L., Wainwright, M. J. and Jordan, M. I. (2009). On surrogate loss functions and f-divergences. *The Annals of Statistics* **37**, 876-904.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics* **32**, 56-85.

Cui Xiong
School of Statistics, East China Normal University,
Shanghai 200241, China.
E-mail:cxiong531@163.com

Jun Shao
Department of Statistics, University of Wisconsin-Madison,
1300 University Ave., Madison, Wisconsin, 53706, U.S.A.
E-mail: shao@stat.wisc.edu

Lei Wang
LPMC and Institute of Statistics, Nankai University,
Tianjin 300071, China.
E-mail:leiwang.stat@gmail.com