A binary-response regression model based on support vector machines

Hien D Nguyen\(^1\) and Daniel V Fryer\(^2\)

July 24, 2020

\(^1\)Department of Mathematics and Statistics, La Trobe University, Bundoora, Victoria Australia. (Email: h.nguyen5@latrobe.edu.au) \(^2\)School of Mathematics and Physics, University of Queensland, St Lucia, Queensland Australia

Abstract

The soft-margin support vector machine (SVM) is a ubiquitous tool for prediction of binary-response data. However, the SVM is characterized entirely via a numerical optimization problem, rather than a probability model, and thus does not directly generate probabilistic inferential statements as outputs. We consider a probabilistic regression model for binary-response data that is based on the optimization problem that characterizes the SVM. Under weak regularity assumptions, we prove that the maximum likelihood estimate (MLE) of our model exists, and that it is consistent and asymptotically normal. We further assess the performance of our model via simulation studies, and demonstrate its use in real data applications regarding spam detection and well water access.

Key words: binary regression; support vector machines; maximum likelihood estimation; numerical optimization

1 Introduction

Let \( Y \in \{-1, 1\} \) be a binary response and let \( X \in \mathbb{X} \subseteq \mathbb{R}^d \) be some covariates. Furthermore, let \( Z_n = \{Z_i\}_{i=1}^n \), where \( Z_i = (X_i^\top, Y_i) \), be an independent and identically distributed (IID) random sample of \( n \in \mathbb{N} \) pairs of response and covariates.

A common problem that arises when considering binary response variables is to use the data \( Z_n \) to construct some discriminant function \( g : \mathbb{X} \to \{-1, 1\} \), such that the probability of misclassification: \( \Pr(g(X) \neq Y) \), is small (cf. Devroye et al. [1996] Ch. 1). In Cortes & Vapnik [1995], the authors proposed the so-called (linear soft-margin) support vector machine (SVM), whereupon the function \( g \) was proposed to take the form
\[ g(x) = \text{sign} \left( \alpha + x^\top \beta \right), \quad (1) \]

where \( \theta^T = (\alpha, \beta^T) \in T \subseteq \mathbb{R}^{d+1} \) are the parameters of \( g \), and \( \text{sign}(x) \) is equal to 1 if \( x \geq 0 \), and equal to 0, otherwise. Here, we will say that \( \alpha \) is the intercept term and \( \beta^T = (\beta_1, \ldots, \beta_d) \) is a vector of coefficients, where \( \beta_j \) is the coefficient of covariate \( j \in [d] = \{1, \ldots, d\} \). In order to estimate the parameters \( \theta \) from \( Z_n \), Cortes & Vapnik (1995) suggested an optimization process that is equivalent to solving the following problem (cf. Shawe-Taylor & Sun, 2011):

\[
\arg \min_{\theta \in T} \frac{1}{n} \sum_{i=1}^{n} l(Z_i; \theta) + \lambda \beta^\top \beta, \quad (2)
\]

where \( \lambda > 0 \) is a regularization constant for the size of \( \beta \), and

\[
l(Z_i; \theta) = \left[ 1 - Y_i x_i^\top \theta \right]_+ \]

is a loss function with \( x_i^\top = (1, x_i^\top) \) and \( [x]_+ = \max \{x, 0\} \). The SVM has become a ubiquitously successful tool for data analysts and applied researchers, and its virtues are well-exposed in volumes such as Abe (2005), Chen et al. (2004), Liang et al. (2016), and Murty & Raghava (2016).

Noting the form of the optimization problem, Polson & Scott (2011) proposed that one can consider the equivalent optimization routine (for \( \lambda = 0 \))

\[
\arg \max_{\theta \in T} -\sum_{i=1}^{n} l(Z_i; \theta) \quad (3)
\]

to be an approximation of the maximum likelihood estimation (MLE) of \( \theta \) under the probability model:

\[
\Pr(Y_i = y_i | X_i = x_i) = f(y_i | x_i; \theta) = \frac{\exp \left( -[1 - y_i x_i^\top \theta]_+ \right)}{\exp \left( -[1 - x_i^\top \theta]_+ \right) + \exp \left( -[1 + x_i^\top \theta]_+ \right)}, \quad (4)
\]

whereby the normalization term in the denominator of (4) is omitted. This approximation was also used by Fu et al. (2010), Mao et al. (2014), Lai et al. (2015), and Wenzel et al. (2017), where it is argued that it is more computationally feasible than MLE (since it is a concave optimization problem) and that the functional form more closely resembles the SVM problem from which it is derived.

In this paper we consider the MLE problem of computing

\[
\hat{\theta}_n = \arg \max_{\theta \in T} l_n(\theta), \quad (5)
\]

where

\[
l_n(\theta) = n^{-1} \sum_{i=1}^{n} \log f(Y_i | X_i; \theta), \quad (6)
\]
instead of (3). We prove that the log-likelihood function (6) is coercive, on average, conditional on the covariates \( \{X_i\}_{i=1}^n \), thus guaranteeing the existence of a global maximizer of the limiting function within the interior of some compact subset of \( T = \mathbb{R}^{d+1} \). This is sufficient for establishing consistency of the estimator. Furthermore, recent evidence suggests that one can compute the maximum of non-convex and non-differentiable functions, such as (6), using quasi-Newton methods such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS; [Fletcher, 1987]) algorithm (see, e.g., [Lewis & Overton, 2013] and [Keskar & Wächter, 2019]), especially with the aid of automatic differentiation (AD; see, e.g., [Bücker et al., 2006]). Regarding the maximum likelihood estimator (MLE), we further demonstrate that one can establish conditions under which consistency and asymptotic normality hold, and thus permit drawing of inference via model (4).

We assess the performance of our approach via a finite sample assessment of its asymptotic properties in simulation studies. Here, we also assess how well the model performs prediction of an unknown response \( Y \) given some observed covariate \( x \) in a similar manner to an SVM and logistic regression (see, e.g., [McLachlan, 1992, Ch. 8, and Hosmer et al., 2013]). We then apply our method to a pair of real-world data sets, regarding spam detection and well water access, and compare the inference drawn from model (4) to those drawn via logistic regression, as well as its ability to conduct prediction as compared to a SVM.

The paper proceeds as follows. In Section 2, we consider the existence, consistency and asymptotic normality of the MLE. In Section 3, we describe our computation strategy and conduct simulation studies. In Section 4, we present example applications. Finally, we present some concluding remarks in Section 5.

2 The maximum likelihood estimator

2.1 Existence

In order for maximum likelihood estimation (MLE) to make sense, we must demonstrate that the MLE exists in some useful sense. To that effect, we wish to show that conditional on \( \{X_i\}_{i=1}^n = \{x_i\}_{i=1}^n \) (for brevity, we shall write \( X_n = \{X_i\}_{i=1}^n \) and \( x_n = \{x_i\}_{i=1}^n \)), the expected value of the log-likelihood (6) has all of its global maxima, with respect to \( \theta \in T = \mathbb{R}^d \), in the interior \( \text{int}(S) \) of some compact set \( S \subset T \), for each \( n \). This can be achieved by showing that

\[-E[l_n(\theta) | X_n = x_n] \rightarrow \infty, \quad \text{if} \quad ||\theta|| \rightarrow \infty, \quad (7)\]

for each \( i \in [n] \) (cf. [Auslender & Teboulle, 2002, Sec. 3.1]), where \(||\cdot||\) denotes the Euclidean norm.

Let \( p_i = f(1|x_i; \theta_0) \in (0, 1) \) and let \( \hat{p}_i = 1 - p_i \). Here, \( \theta_0 \in T \) is the true value of \( \theta \), which arises from the data generating process of \( Z_n \). Then, we may
Figure 1: Visualization of the components of (8), for $p_i \in \{0.2, 0.5, 0.8\}$.

We firstly wish to show that (8) is coercive, with respect to $\hat{\theta}_i$. A visualization of $h$, $h_1$ and $h_2$ appears in Figure 1.

We may inspect $h_2$ at its limits and local extrema and observe that $|h_2(\hat{\theta}_i)| \leq 1 - \log(2)$, for all $\hat{\theta}_i$. Thus, $h_2$ is a bounded function. We observe that $h_1$ is coercive since $h_1(\hat{\theta}_i) \to \infty$ for $\hat{\theta}_i \to \pm\infty$. We thus also establish that $h = h_1 + h_2$ is coercive since $h_2$ is a function that is bounded from below (and in this case, also above).

Next, we appeal to Lemma 5.1 of Calatroni et al. (2019) (see also Clarke 2015, Lem. 2.7.1), which implies that if $h(\theta_i)$ is a proper, continuous, and coercive function, and if we have the null space condition:

$$\text{null}(\tilde{x}_i) = \{\theta \in \mathbb{R}^d : \tilde{x}_i^T \theta = 0\} = \{0\},$$

write

$$-\mathbb{E}[\log f(1|x_i; \theta)] = p_i \left[1 - \tilde{x}_i^T \theta\right]_+ + \bar{p}_i \left[1 + \tilde{x}_i^T \theta\right]_+ + \log \left[\exp \left(-\left[1 - \tilde{x}_i^T \theta\right]_+\right) + \exp \left(-\left[1 + \tilde{x}_i^T \theta\right]_+\right)\right],$$

for each $i \in [n]$.

Consider the substitution $\tilde{x}_i^T \theta = \tilde{\theta}_i$ and thus write

$$-\mathbb{E}[\log f(1|x_i; \theta)] = h(\tilde{\theta}_i) = h_1(\tilde{\theta}_i) + h_2(\tilde{\theta}_i),$$

where

$$h_1(\tilde{\theta}_i) = p_i \left[1 - \tilde{\theta}_i\right]_+ + \bar{p}_i \left[1 + \tilde{\theta}_i\right]_+,$$

and

$$h_2(\tilde{\theta}_i) = \log \left[\exp \left(-\left[1 - \tilde{\theta}_i\right]_+\right) + \exp \left(-\left[1 + \tilde{\theta}_i\right]_+\right)\right].$$
then (8) is continuous and coercive, with respect to $\theta$. Here $0$ denotes the zero vector, and proper is taken to mean that $h(\hat{\theta}_i) < \infty$ for at least one $\theta_i$ and $h(\hat{\theta}_i) > -\infty$ for all $\hat{\theta}_i \in \mathbb{R}$. Since $h$ is univariate, coercive, and bounded below, we automatically have the fact that $h$ is proper, and thus (8) is coercive as long as null$(\tilde{x}_i) = \{0\}$. Since the sum of coercive functions is coercive, we obtain the following result.

**Proposition 1.** If $\bigcap_{i \in [n]} \text{null}(\tilde{x}_i) = \{0\}$, then the expected conditional log-likelihood $E[l_n(\theta) | X_n = x_n]$ is coercive, and thus there exists some compact set $S \subset T = \mathbb{R}^{d+1}$, such that the set of global maxima

$$ \arg \max_{\theta \in T} E[l_n(\theta) | X_n = x_n] $$

is equal to

$$ \arg \max_{\theta \in \text{int}(S)} E[l_n(\theta) | X_n = x_n]. $$

From Proposition 1, we may conclude that (5) exists, so long we do not observe data $Z_n$, where all of the responses are equal to $-1$ or are all equal to $1$, and where we do not observe some pathological set of covariates $X_n$, where $\bigcap_{i \in [n]} \text{null}(X_i) \neq \{0\}$. One potential pathology is if all of the vectors of $X_n$ are linearly dependent. For $X_i$ arising from some continuous distribution, this event will occur with probability zero.

### 2.2 Consistency

We begin by establishing the consistency of the MLE over some arbitrarily large compact subset $S$ of $T$, as we are permitted to do via Proposition 1. Further assume that $X$ is a compact subset of $\mathbb{R}^d$. To this end, we firstly consider the limit of (6) conditional on $X_n = x_n$, for fixed $\theta \in T$. Using the independent but not identical law of large numbers of White (2001, Cor. 3.9), we have the fact that

$$ l_n(\theta) \xrightarrow{p} E[l_n(\theta) | X_n = x_n], $$

as $n \to \infty$, conditional on the existence of some constant $C < \infty$, such that $E \left[ \log^2 f(Y_i | x_i; \theta) \right] < C$, for all $i$, and fixed $\theta$. Here $\xrightarrow{p}$ denotes convergence in probability. This is easy to verify, since $Y_i \in \{-1, 1\}$ is a discrete random variable and $X$ is compact, so we may take

$$ C = \sup_{x \in X} \left[ \log^2 f(-1 | x; \theta) + \log^2 f(1 | x; \theta) \right]. $$

Next, we must make the convergence in probability uniform over some compact set $S \subseteq T$. That is, we require that

$$ \sup_{\theta \in S} |l_n(\theta) - E[l_n(\theta) | X_n = x_n]| \xrightarrow{p} 0. $$

5
We can verify this using the generic uniform law of large numbers of Newey (1991, Cor. 3.1). This can be established by verifying that $S$ is compact, that $|\log f(y|x; \theta) - \log f(y|x; \vartheta)| \leq L \|\theta - \vartheta\|$, for each fixed $(y, x) \in \{-1, 1\} \times X$, where $\theta, \vartheta \in S$ and $L < \infty$ is a constant.

As in (8), we consider the map $\tilde{\theta} = \tilde{x}^\top \theta$. Let $I \subset \mathbb{R}$ be a sufficiently large compact interval such that $\tilde{x}^\top \theta \in I$ for all values of $x \in X$ and $\theta \in S$. This is possible via the compactness of $S$ and $X$. Next we wish to establish the fact that $\tilde{h}(\theta) = -1 + y\tilde{x}$ is Lipschitz with respect to $\tilde{\theta} \in I$. This can be achieved by noting that $\tilde{h}$ is piecewise continuously differentiable, and by applying Scholtes (2012, Cor. 4.1.1). Using the affine map $\tilde{\theta} = \tilde{x}^\top \theta$, from $S$ to $I$, we establish (11) by the fact that Lipschitz compositions are Lipschitz. Thus, (10) is verified.

By the continuity of (6) and its uniform convergence in probability (10), we can now apply Nguyen & McLachlan (2016, Lem. 5), a non-smooth version of the extremum estimator consistency theorem of Amemiya (1985), in order to establish the following consistency result regarding (5).

**Proposition 2.** Let $S \subset T$ and $X \subset \mathbb{R}^d$ be compact, such that $S_n = \{ \theta \in \text{int}(S) : \theta$ is a local maximum of $l_n(\theta) \}$, and assume that $E[l_n(\theta) | X_n = x_n]$ attains a strict local maximum at $\theta_0 \in \text{int}(S)$. Then, for any $\epsilon > 0$, $\inf_{\theta \in S_n} \|\theta - \theta_0\| \overset{p}{\to} 0$.

This proposition is useful in the context of solving problem (5) since (6) is likely to have multiple local and global maxima, and similarly with the conditional expected log-likelihood $E[l_n(\theta) | X_n = x_n]$. The result ensures that if we follow the sequences of strict local maxima of (6), then we obtain sequences of consistent estimators for each of the local maxima $\theta_0$ of $E[l_n(\theta) | X_n = x_n]$. Of course, in any one run of an optimization algorithm, one tends to only find one local maximum. Thus, it is often advisable to run the optimization algorithm for computing (5) multiple times, with different initializations, in order to ensure that one has located the local maximum that corresponds to the $\theta_0$ one is seeking (cf. Amemiya, 1985, Sec. 4.1.1).

### 2.3 Asymptotic normality

We may now establish asymptotic normality via Amemiya (1985 Thm. 4.1.3). Here, we require that the local maximum of interest $\theta_0$ is such that the assumptions of Proposition 2 are satisfied, and also that there is an open and convex neighborhood around $\theta_0$, where $\log f(y_i | x_i; \theta)$ (for each $i \in [n]$) is three times
differentiable with respect to \( \theta \in S \), and all first, second, and third order partial derivatives are bounded (in order to apply Amemiya 1985, Thm. 4.1.4). Here, the boundedness can be established simply via the fact that \( Y_i \in \{-1, 1\} \) is a discrete random variable. We thus have the following result.

**Proposition 3.** Assume that \( S \subset T \) and \( X \subset \mathbb{R}^d \) are compact, and that \( \log f(y|x; \theta) \) is three times differentiable with respect to \( \theta \) in a open and convex neighborhood of \( \theta_0 \), for each \( i \in [n] \). Let \( \{\hat{\theta}_n\}_{n=1}^\infty \) be a sequence that is obtained by choosing one element of \( S_n \) (for each \( n \)), as defined in Proposition 2, such that \( \hat{\theta}_n \xrightarrow{p} \theta_0 \). Then, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) converges in law to a normal distribution with mean vector 0 and covariance matrix \( A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0) \), where \( A(\theta_0) \) is assumed to be non-singular, and

\[
A(\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \frac{\partial^2 f(Y_i|X_i; \theta)}{\partial \theta \partial \theta^\top} \bigg| X_i = x_i \right]
\]

and

\[
B(\theta) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \frac{\partial f(Y_i|X_i; \theta)}{\partial \theta} \bigg| X_i = x_i \right] \frac{\partial f(Y_i|X_i; \theta)}{\partial \theta^\top} \bigg| X_i = x_i \right].
\]

We note that the covariance form \( A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0) \) assumes that there may be misspecification between the model (4) and the data generating process of \( Z_n \) (cf. White, 1982). If there is no misspecification, then we may take \( A(\theta_0) = -B(\theta_0) \) and thus the covariance matrix reduces to \(-A^{-1}(\theta_0) = B(\theta_0) \). We may estimate \( A(\theta_0) \) and \( B(\theta_0) \) by

\[
\hat{A}_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i; \theta)}{\partial \theta \partial \theta^\top} \bigg| \hat{\theta}_n \tag{12}
\]

and

\[
\hat{B}_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i; \theta)}{\partial \theta} \bigg| \hat{\theta}_n \frac{\partial \log f(Y_i|X_i; \theta)}{\partial \theta^\top} \bigg| \hat{\theta}_n \tag{13}
\]

respectively, via the sample \( Z_n \) (see, e.g., Boos & Stefanski 2013, Thm. 7.3).

3 Implementation and numerical studies

3.1 Computational specifics

We compute the MLE (4) using the BFGS method as implemented in R (R Core Team 2020) via the `optim` function. Here, we use gradients that are computed via AD using the package autodiff (Li 2018). It is established in Lewis & Overton (2009), Lewis & Overton (2013), and Keskar & Wächter (2019), that the
BFGS performs well in non-differentiable and non-convex settings. In such situations, they prove that the line search steps are convergent under general conditions, although it is difficult to prove the global convergence of the algorithm, overall, except in the simple case of the Euclidean norm function. However, via comprehensive simulation studies, it is found that the BFGS method tends to be correct under standard settings.

In order to guarantee global convergence, Lewis & Overton (2009) and Lewis & Overton (2013) suggest that one should apply a gradient sampling method after the BFGS solution is found. It is established in Burke et al. (2005) that gradient sampling is globally convergent under standard settings, and we implemented the BFGS-then-gradient (also referred to by Lewis & Overton 2009 and Lewis & Overton 2013 as HANSO: hybrid algorithm for non-smooth optimization) sampling approach via the rHanso package of Mallik & Borchers (2013).

Via a battery of simulation settings, we found that the hanso function from rHanso produced exactly the same outcomes as BFGS method using optim in many cases, and in other cases was actually less optimal. Thus, since HANSO requires an gradient sampling step, which is significantly more computationally intensive, we opted to rely on the standard BFGS method only, for all of our computations. Code, for some of the computation in this section and the next, can be found online at: https://github.com/hiendn/svm_binary_regression.

3.2 Finite sample accuracy of the MLE

Although Proposition 2 implies that one can always arbitrarily accurately estimate $\theta_0$, that characterizes the data generating process (4), with the MLE (5) using a sufficiently large IID sample $Z_n$, it is unclear as to how large $n$ needs to be in order for Proposition 2 to apply. We thus assess the performance of the MLE when $n$ is a finite value. Here we choose $n \in \{100, 200, 500, 1000\}$.

For each $n$, we simulate $Z_n$ with each covariate $X_i$ arising from a multivariate normal distribution with mean 0 and covariance matrix $I$ (identity matrix), for dimensions $d \in \{1, 5, 10\}$. We then simulate each $Y_i$ using the model (4) with $\theta_0^T = (\alpha_0, \beta_0^T) = 1^T$ (the ones vector). An MLE $\hat{\theta}_n$ is then computed using the BFGS algorithm, as described above.

For each combination of $n$ and $d$, we repeat the simulation above $R = 100$ times. We then compute the mean squared error

$$\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \left\| \hat{\theta}_n^{(r)} - \theta_0 \right\|^2,$$

for each simulation combination and report it Table 1.

From Table 1 we observe, as expected, that for fixed $n$, an increase in the dimensionality $d$ increases the MSE, since there are more parameters to estimate and thus the complexity of the problem increases. Furthermore, for fixed $d$, we observe that as $n$ increases, the MSE decreases. This conforms with the conclusions from the consistency result of Proposition 2. We observe for the
Table 1: Mean squared errors from 100 replications of the MLE (5) for various combinations of dimension \(d\) and sample size \(n\). Here, \(a(b) = a \times 10^b\).

| \(n\)  | \(d = 1\) | \(d = 5\) | \(d = 10\) |
|-------|-----------|-----------|-----------|
| 100   | 1.80(–1)  | 2.75(–0)  | 2.30(+3)  |
| 200   | 7.27(–2)  | 3.78(–1)  | 3.48(+0)  |
| 500   | 2.68(–2)  | 1.57(–1)  | 3.30(–1)  |
| 1000  | 1.21(–2)  | 6.12(–2)  | 1.37(–1)  |
| 2000  | 6.82(–3)  | 2.85(–2)  | 8.76(–2)  |

larger values of \(n\) (500, 1000, and 2000), that the rate of decrease of the MSE is approximately linear in \(n\), which is as predicted by the asymptotic normality result of Proposition 3.

In Figure 2, we plot the conditional probability curves \(f(1|x, \hat{\theta}^r_n(n))\), with respect to \(x\), corresponding to each of the replications in the case of \(d = 1\), for \(n = 100\) and \(n = 1000\). We observe that there is a dramatic increase in accuracy of the estimation of the generative conditional probability curve, when \(n\) is increased from 100 to 1000.

### 3.3 Binary prediction accuracy

Here we assess the ability of model (4), fitted via MLE, to predict the value of \(Y\) given some observed covariate \(X\). This prediction is conducted in the same manner as when performing prediction using logistic regression. That is, we use the maximum a posteriori approach, whereupon we predict \(Y\) via the rule:

\[
\hat{y}(X) = \arg \max_{y \in \{-1, 1\}} f(y|X; \hat{\theta}_n) .
\]

(14)

In order to assess the performance of rule (14), we conduct the following simulation study. A sample \(Z_n\) of \(n \in \{100,1000\}\) pairs of responses and covariates are simulated from a two-component normal mixture model (see, e.g., McLachlan & Peel, 2000, Ch. 3), where each \(X_i\) is of dimension \(d \in \{2,5\}\), using the MixSim package of Melnykov et al. (2012). Here, the package allows for control of level of overlap between the mixture components via a parameter \(\bar{\omega}\) (larger implies greater overlap), which has default value 0.05. Here, we assess situations where \(\bar{\omega} \in \{0.05, 0.5\}\).

We then estimate the MLE \(\hat{\theta}_n\) using data \(Z_n\). An additional \(N = 1000\) pairs \(Z'_N = \{(X'_i, Y'_i)\}_{i=1}^N\) is generated from the same data generating process as \(Z_n\). Rule (14) is then applied to estimate each \(Y_i\) via \(\hat{y}(X'_i)\). The accuracy of the prediction is then recorded as

\[
\text{ACC} = N^{-1} \sum_{i=1}^N ||Y'_i = \hat{y}(X'_i)|| ,
\]
Figure 2: Curve of the estimated conditional probability function $f(1|x; \hat{\theta}_n^{(r)})$ for each replication from the accuracy simulation for the $d = 100$ scenario, where the MLE is computed using samples of sizes $n = 100$ and $n = 1000$. The dashed curve indicates the generative conditional probability function, where $\theta_0 = 1$. 
Table 2: Accuracies averaged over 100 replications (along with standard deviations, in italic) of predictions using Rule (14), logistic regression (LR), and SVM are provided for various combinations of dimension $d$, sample size $n$, and separation coefficient $\bar{\omega}$.

| $n$   | $d$ | $\bar{\omega} = 0.05$ | $\bar{\omega} = 0.5$ |
|-------|-----|------------------------|------------------------|
|       |     | (14) LR SVM            | (14) LR SVM            |
| 100   | 2   | 0.963 0.963 0.964      | 0.663 0.662 0.670      |
|       |     | 0.016 0.016 0.018      | 0.074 0.073 0.070      |
| 5     | 2   | 0.952 0.950 0.959      | 0.600 0.599 0.604      |
|       |     | 0.014 0.015 0.009      | 0.051 0.052 0.047      |
| 1000  | 2   | 0.967 0.968 0.967      | 0.691 0.689 0.702      |
|       |     | 0.014 0.014 0.015      | 0.059 0.059 0.049      |
| 5     | 2   | 0.966 0.966 0.966      | 0.614 0.614 0.620      |
|       |     | 0.007 0.007 0.007      | 0.049 0.049 0.047      |

where $[A] = 1$ if statement A is true and $[A] = 0$, otherwise. We also compute the accuracy of predicting the responses of $Z'_N$ via the covariates, using the logistic regression rule and SVM rule (1), where the respective models are estimated using the data $Z'_n$, only. Here, logistic regression and SVM are implemented using the glm function and the svm function (in the package e1071; Meyer et al., 2019) in R, respectively.

The experiment is repeated $R = 100$ times for each combination of $(d, n, \bar{\omega})$. The accuracies for each of the three assessed methods are averaged and a standard deviation is computed. These results are presented in Table 2. Example decision boundaries for each of the three prediction rules for the $(d, n, \bar{\omega}) = (2, 1000, 0.05)$ case are visualized in Figure 3.

We now discuss some observations regarding Table 2. Firstly, all three methods appear to perform equally well across each of the simulation scenarios. However, there is a tendency for SVM to perform better than Rule (14), which also has a tendency of performing equal or better than logistic regression. This ordering makes some sense as Rule (14) is a probabilistic version of the usual SVM rule (1), and it is also constructed in a manner similar to that of logistic regression. We note that both our model and logistic regression have the advantage over SVM in that they both generate posterior probabilities of $Y'$ given a fixed value of $X'$, whereas SVM does not since it is not probabilistic in construction. Thus, the posterior probabilities of can be calculated directly using our method and logistic regression, whereas SVM requires an approximate calculation via techniques such as those of Platt (1999) and Lin et al. (2007).

Next, we observe that our usual intuition regarding difficulty of prediction is met by these results. That is, as $n$ increases, accuracy improves, since more data is used to learn the prediction models. Further, greater dimensionality $d$ decreases accuracy for each fixed $n$ and $\bar{\omega}$, since the greater dimensionality increases the model complexity of the model. Lastly, increasing overlap drastically decreases prediction accuracy, since the heterogeneity of the data becomes
Figure 3: Scatter plot of the data $Z_N'$ for an instance of the $(d, n, \bar{\omega}) = (2, 1000, 0.05)$ simulation scenario. Here, circles and crosses indicate that $y_i$ equals $-1$ or $1$, respectively. The solid, dashed, and dotted lines represent the decision rule (14), and the logistic regression and SVM rules, respectively.
more difficult to recognize.

4 Applications

4.1 Wells data

We investigate the wells data set attributed to Gelman & Hill [2007], from the carData package [Fox et al., 2019]. The data are obtained from households in an area of Arahazar Upazila, Bangladesh, where people were exposed to unsafe levels of arsenic in their well water supply. The data consists of \( n = 3020 \) households, where the response of interest \( y_i \) indicates whether the household \( i \) switched from using their arsenic contaminated well to a safer one. Here, \( y_i = -1 \) indicates that the household did not switch, whereas \( y_i = 1 \) indicates a switch of water supply. In order to characterize the switching behavior, the level of arsenic contamination in the original well of the household in hundreds of micrograms per liter (arsen), the distance to the closest known safe well (dist), the education level of the head of the household in years (edu), and an indicator as to whether any members were associated with a community organization (assoc; 1 indicates an association) were also measured as covariates \( x_i \). A plot of the data appears in Figure 4.

To draw inference from these data, via MLE, we fit both a logistic regression model (using \texttt{glm}) and model (4). The log-likelihoods of the estimated logistic regression and model (4) were \(-1953.91\) and \(-1953.32\), respectively. This implies that (4) provided a slightly better fit to these data, but the closeness of the two log-likelihood outcomes indicates that the inference drawn from both models should be similar.

The estimated intercept term for logistic regression was \( \hat{\alpha}_n = -0.1567 (0.1006) \), and the coefficients for each of the covariates were estimated to be \( \hat{\beta}_{\text{arsen},n} = 0.4670 (0.0452) \), \( \hat{\beta}_{\text{dist},n} = -0.0090 (0.0010) \), \( \hat{\beta}_{\text{edu},n} = 0.0424 (0.0095) \), and \( \hat{\beta}_{\text{assoc},n} = -0.1243 (0.0771) \). Here, the bracketed terms are asymptotic mis-specification robust standard errors, as computed via the \texttt{sandwich} function, via the \texttt{sandwich} package [Zeileis, 2004]. Using Wald tests for the hypotheses \( H_0 : \beta_j = 0 \) versus \( H_1 : \beta_j \neq 0 \), we found that arsen, dist, and edu were all significant at at least the \( \alpha = 10^{-5} \) level, under asymptotic normality. We found that assoc was not significant at any \( \alpha < 0.1 \) level.

Moving onto model (4), we estimated the intercept term to be \( \hat{\alpha}_n = -0.0871 (0.0505) \), via MLE. The corresponding estimates for the coefficients of the covariates were \( \hat{\beta}_{\text{arsen},n} = 0.2407 (0.0230) \), \( \hat{\beta}_{\text{dist},n} = -0.0045 (0.0005) \), \( \hat{\beta}_{\text{edu},n} = 0.0210 (0.0048) \), and \( \hat{\beta}_{\text{assoc},n} = -0.0594 (0.0387) \). Here, the bracketed terms are asymptotic standard errors, computed using Proposition 3 and expressions (12) and (13). Wald tests for the hypotheses \( H_0 : \beta_j = 0 \) versus \( H_1 : \beta_j \neq 0 \) found that arsen and dist were significant at the \( \alpha = 10^{-5} \) level, edu was significant at the \( \alpha = 10^{-4} \) level, and assoc was not significant at any \( \alpha < 0.1 \) level.

As expected both logistic regression and model (4) provided very similar inference, as we notice that all corresponding coefficients are of the same sign.
Figure 4: The pairwise scatter plots of the covariates of the Wells data are plotted. Observations that correspond to a response of $y_i = -1$ are plotted as circles. Observations corresponding to $y_i = 1$ are plotted as crosses.
Furthermore, both models concluded that there were significant effects due to arsen, dist, and edu, but not due to assoc.

4.2 Spam data

We next investigate the spam7 data set from the DAAG package of [Maindonald & Braun, 2006]. These data contain \( n = 4601 \) observations regarding features of emails, where the response to be predicted is the indicator as to whether the email is spam: \( y_i \), which equals to \(-1\) if it is not spam, and 1 otherwise. The \( d = 7 \) covariates stored in \( x_i \) by which \( y_i \) may be conditionally dependent upon are the total length of words in capitals, number of occurrence of the dollar sign, number of occurrence of the bang symbol, number of occurrences of the word 'money', number of occurrences of the string '000', and number of occurrences of the word 'make'.

Upon fitting an SVM and model (4), we conduct prediction on the data from which the models were fitted and compute the prediction accuracies to be \( 0.8444 \) and \( 0.8476 \), respectively. This indicates that model (4) fits the data set slightly better than SVM. We next consider 5-fold cross-validated accuracies of the two models (cf. Arlot et al., 2010, regarding cross-validation methods). Using the same partitioning of the data, we compute the cross-validated accuracies to be \( 0.8479 (0.0095) \) and \( 0.8444 (0.0119) \), respectively, where standard deviations are reported in parentheses. We observe that both methods perform comparably in the prediction task, although model (4) using Rule (14) yielded slightly higher accuracy levels.

5 Concluding remarks

Remark 1. A powerful concept in SVM is that of reproducing kernel Hilbert space (RKHS) embedding. That is, instead of considering the linear map \( x^\top \beta \) in (1), one considers a map \( \eta : \mathbb{R}^d \rightarrow \mathbb{R} \), where \( \eta \) is in some RKHS \( \mathcal{H} \) (cf. Steinwart & Christmann, 2008). If \( \eta(x) = \gamma^\top \phi(x) \), for some finite dimensional vector \( \gamma \in \mathbb{R}^q (q \in \mathbb{N}) \) and map \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^q \), then the analogous application of model (4) is straightforward. That is, one simply replaces \( \beta \) with \( \gamma \) in \( \theta \), and one replaces \( x_i \) by \( \phi(x_i) \). This is true for example when one considers the RKHS corresponding to polynomial kernels of the form

\[
\kappa(x, x') = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}} = (x^\top x' + c)^u,
\]

where \( c \in \mathbb{R} \) and \( u \in \mathbb{N} \). Here, \( \langle \cdot, \cdot \rangle_{\mathcal{H}} \) denotes the inner product of the RKHS \( \mathcal{H} \). The functions \( \eta \) and \( \kappa \) are related via the so-called reproducing property: \( \eta(x) = \langle \eta, \kappa(\cdot, x) \rangle_{\mathcal{H}} \). In the case where \( \eta \) does not correspond to some finite dimensional mapping \( \phi \), the situation is more complicated and is beyond the scope of this article.

Remark 2. The coerciveness result of Proposition 1 is proved in terms of (7) in order to facilitate the consistency result of Proposition 2. However, we may
consider instead coerciveness of the negative log-likelihood function \(-l_n(\theta)\), without any probabilistic assumptions on the data \(Z_n\) (or assuming that \(Z_n = \{(x_i^T, y_i)\}_{i=1}^n\) with probability one).

Without loss of generality, we assume that \(y_1 = 1\) and \(y_2 = -1\), and we write

\[
-l_n(\theta) = -n^{-1} \log f(y_1|\theta) - \log f(y_2|x_2; \theta) - n^{-1} \sum_{i=3}^n \log f(y_i|x_i; \theta)
\]

\[
= n^{-1} \left[ \tilde{h}_1(\theta) + \tilde{h}_2(\theta) \right],
\]

where

\[
\tilde{h}_1(\theta) = \left[ 1 - \tilde{x}_i^T \theta \right]_+ + \left[ 1 + \tilde{x}_i^T \theta \right]_+,
\]

and

\[
\tilde{h}_2(\theta) = 2 \log \left[ \exp \left( - \left[ 1 - \tilde{x}_i^T \theta \right]_+ \right) + \exp \left( - \left[ 1 + \tilde{x}_i^T \theta \right]_+ \right) \right]
\]

\[
+ \sum_{i=3}^n \log f(y_i|x_i; \theta).
\]

It is easy to see that \(\tilde{h}_2\) is bounded from below, since \(\left[ 1 - y_i \tilde{x}_i^T \theta \right]_+\) is bounded from below by zero and due to the bounding of \(h_2\) (from Section 2.1). Thus, by the fact that the sum of a coercive function and a function that is bounded from below is coercive, we are only required to establish conditions under which \(\tilde{h}_1\) is coercive. Here, we use the fact that \(\tilde{h}_1\) is convex and Corollary 2.5.3 of Auslender & Teboulle (2002), which implies that \(\tilde{h}_1\) is coercive if the function \(\tilde{h}_{1,\infty}(\theta) > 0\) for all \(\theta \in \mathbb{T} \setminus \{0\}\), where

\[
\tilde{h}_{1,\infty}(\theta) = \lim_{t \downarrow 0} t \tilde{h}_1 \left( t^{-1} \theta \right)
\]

\[
= \frac{|\tilde{x}_1^T \theta| + |\tilde{x}_2^T \theta| + \tilde{x}_2^T \theta - \tilde{x}_1^T \theta}{2}.
\]

If we consider all the possible sign combinations of \(x_1^T \theta\) and \(x_2^T \theta\), we end up with the following conditions that ensure \(\tilde{h}_{1,\infty}(\theta) > 0\): (1) \(\tilde{x}_1^T \theta \geq 0\) and \(\tilde{x}_2^T \theta > 0\), (2) \(\tilde{x}_1^T \theta \leq 0\) and \(\tilde{x}_2^T \theta > 0\), (3) \(\tilde{x}_1^T \theta < 0\) and \(\tilde{x}_2^T \theta \leq 0\), and (4) \(\tilde{x}_1^T \theta < 0\) and \(\tilde{x}_2^T \theta > 0\). Thus a minimal set of assumptions for the coerciveness of \(-l_n(\theta)\) is that \(y_1 = 1\), \(y_2 = -1\), and that \(x_1\) and \(x_2\) are such that for any \(\theta \neq 0\), one of situations (1)–(4) is true. This is sufficient for guaranteeing the existence of the MLE \((\hat{\theta})\). One situation when these conditions are fulfilled is if \(x_1\) and \(x_2\) have no zero elements, and if \(x_2 = C x_1\), for some \(C > 0\).

The conditions above can also be used to establish the existence of a solution to problem (3) and to the \(\lambda = 0\) case of the SVM problem (2). To the best of our knowledge, this is the first such set of conditions for establishing coerciveness of the SVM problem (2) when \(\lambda = 0\).
Remark 3. We note that the BFGS approach that we used for optimization in Sections 3 and 4 is by no means the only methods that can be applied to solve the MLE problem (5). Recently, there has been rapid development in the research of algorithms that are provably convergent for broad classes of non-differentiable and non-convex optimization problems. For example, the piecewise differentiable approximation approach of [Griewank & Walther, 2019] is applicable, here, as well as various techniques presented in [Bagirov et al., 2020], such as bundle methods and model-based derivative free methods.

References

Abe, S. (2005). *Support Vector Machines for Pattern Classification*. London: Springer.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.

Auslender, A. & Teboulle, M. (2002). *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. New York: Springer.

Bagirov, A., Gaudioso, M., Karmitsa, N., Makela, M., & Taheri, S., Eds. (2020). *Numerical Nonsmooth Optimization: State of the Art Algorithms*. Cham: Springer.

Boos, D. D. & Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. New York: Springer.

Bücker, H. M., Corliss, G., Hovland, P., Naumann, U., & Norris, B. (2006). *Automatic Differentiation: Applications, Theory, and Implementations*. Cham: Springer Science.

Burke, J. V., Lewis, A. S., & Overton, M. L. (2005). A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15, 751–779.

Calatroni, L., Lanza, A., Pragliola, M., & Sgallari, F. (2019). A flexible space-variant anisotropic regularization for image restoration with automated parameter selection. *SIAM Journal on Imaging Sciences*, 12(2), 1001–1037.

Chen, N., Lu, W., Yang, J., & Li, G. (2004). *Support Vector Machine In Chemistry*. Singapore: World Scientific.

Ciak, R. (2015). Coercive functions from a topological viewpoint and properties of minimizing sets of convex functions appearing in image restoration. *arXiv preprint arXiv:1506.08615*. 17
Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer Science.

Fletcher, R. (1987). *Practical Methods of Optimization*. New York: Wiley.

Fox, J., Weisberg, S., & Price, B. (2019). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-3.

Fu, Z., Robles-Kelly, A., & Zhou, J. (2010). Mixing linear SVMs for nonlinear classification. *IEEE Transactions on Neural Networks*, 21(12), 1963–1975.

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Griewank, A. & Walther, A. (2019). Relaxing kink qualifications and proving convergence rates in piecewise smooth optimization. *SIAM Journal on Optimization*, 29, 262–289.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken: Wiley.

Keskar, N. & Wächter, A. (2019). A limited-memory quasi-Newton algorithm for bound-constrained non-smooth optimization. *Optimization Methods and Software*, 34(1), 150–171.

Lai, V., Nguyen, D., Nguyen, K., & Le, T. (2015). Mixture of support vector data descriptions. In *2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)* (pp. 135–140): IEEE.

Lewis, A. S. & Overton, M. L. (2009). Nonsmooth optimization via bfgs.

Lewis, A. S. & Overton, M. L. (2013). Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141(1-2), 135–163.

Li, C. (2018). *autodifffr: Automatic Differentiation in R through 'Julia'* . R package version 0.1.0.9000.

Liang, Y., Xu, Q.-S., Li, H.-D., & Cao, D.-S. (2016). *Support Vector Machines and Their Application in Chemistry and Biotechnology*. Boca Raton: CRC Press.

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on Platt’s probabilistic outputs for support vector machines. *Machine Learning*, 68(3), 267–276.

Maindonald, J. & Braun, J. (2006). *Data Analysis and Graphics using R: an Example-Based Approach*. Cambridge: Cambridge University Press.
Mallik, A. & Borchers, H. W. (2013). \textit{rHanso: An R Implementation of Hybrid Algorithm for Non-Smooth Optimization (HANSO)}. R package version 0.3-1/r107.

Mao, X., Wu, O., Hu, W., & O’Donovan, P. (2014). Nonlinear classification via linear SVMs and multi-task learning. In \textit{Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management} (pp. 1955–1958).

McLachlan, G. J. (1992). \textit{Discriminant Analysis and Statistical Pattern Recognition}. Hoboken: Wiley.

McLachlan, G. J. & Peel, D. (2000). \textit{Finite Mixture Models}. Hoboken: Wiley.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. \textit{Journal of Statistical Software}, 51(12).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). \textit{e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)}, TU Wien. R package version 1.7-3.

Murty, M. N. & Raghava, R. (2016). \textit{Support Vector Machines and Perceptrons: Learning, Optimization, Classification, and Application to Social Networks}. Cham: Springer.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. \textit{Econometrica}, 59, 1161–1167.

Nguyen, H. D. & McLachlan, G. J. (2016). Laplace mixture of linear experts. \textit{Computational Statistics & Data Analysis}, 93, 177–191.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. \textit{Advances in Large Margin Classifiers}, 10(3), 61–74.

Polson, N. G. & Scott, S. L. (2011). Data augmentation for support vector machines. \textit{Bayesian Analysis}, 6(1), 1–23.

R Core Team (2020). \textit{R: A Language and Environment for Statistical Computing}. R Foundation for Statistical Computing, Vienna.

Scholtes, S. (2012). \textit{Introduction to Piecewise Differentiable Equations}. New York: Springer.

Shawe-Taylor, J. & Sun, S. (2011). A review of optimization methodologies in support vector machines. \textit{Neurocomputing}, 74(17), 3609–3618.

Steinwart, I. & Christmann, A. (2008). \textit{Support Vector Machines}. New York: Springer.
Wenzel, F., Galy-Fajou, T., Deutsch, M., & Kloft, M. (2017). Bayesian non-linear support vector machines for big data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 307–322).: Springer.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

White, H. (2001). *Asymptotic Theory for Econometricians*. San Diego: Academic Press.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11, 1–17.