Journal of Mathematics and Statistics

Original Research Paper

Cost-based reweighting for Principal LqSVM for sufficient dimension reduction

Andreas Artemiou[1]

[1] *School of Mathematics, Cardiff University, Cardiff, Wales, UK;*

Article history
Received:
Revised:
Accepted:

Corresponding Authors:
{Corresponding Author}
Email: {email}

Abstract: In this work we try to address the imbalance of the number of points which naturally occurs when slicing the response in Sufficient Dimension Reduction methods (SDR). Specifically, some recently proposed support vector machine based (SVM-based) methodology suffers a lot more due to the properties of the SVM algorithm. We target a recently proposed algorithm called Principal LqSVM and we propose the reweighting based on a different cost. We demonstrate that our reweighted proposal works better than the original algorithm.

Keywords: Support Vector Machines, Kernel methods, imbalance

Introduction

Sufficient Dimension Reduction (SDR) is a class of supervised linear and nonlinear feature extraction methods which are being developed mainly in a regression as well as in classification settings. In SDR we have a response variable Y (which we assume univariate without loss of generality) and a p-dimensional predictor vector X. Our objective is to reduce the dimension of X by finding d (where $d < p$) linear or nonlinear functions of X without losing information on the conditional distribution $Y|X$. In it's simpler form, we can express this using the linear independence model:

$$Y \perp\!\!\!\perp X|\beta^{\tau}X \tag{1}$$

and our effort is to estimate the $p \times d$ matrix β. It is obvious that if β is the identity matrix it satisfies the conditional independence model above but there is no dimension reduction achieved. The space spanned by the columns of β is called a Dimension Reduction Subspace (DRS). Since there are many β's that satisfy model (1) we focus on estimating the one which gives the minimum d. If such a space exists, we call it the Central Dimension Reduction Subspace (CDRS) or simply the Central Subspace (CS). CS does not always exist, but if it exists it is unique. The conditions of existence are relatively mild and we assume its existence throughout this paper. The interested reader is referred to Cook (1998) for more details on the existence of the subspace. Some methods under this model include Sliced Inverse Regression (SIR) by Li (1991), Sliced Average Variance Estimation (SAVE) by Cook and Weisberg (1991), Contour Regression (CR) by Li, Zha and Chiaromonte (2005), Directional Regression (DR) by Li and Wang (2007) and Sliced Inverse Mean Difference (SIMD) by Artemiou and Tian (2015). Most of the methods discussed here use inverse moments to perform feature extraction. More generally we express the nonlinear feature extraction using the model:

$$Y \perp\!\!\!\perp X|\phi(X)$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^d$ can be any linear or nonlinear function of the predictors. Well known works on this framework include Kernel SIR by Wu (2008) and Yeh et al (2009), Kernel regression by Fukumizu, Bach and Jordan (2009) and Principal Support Vector Machine (PSVM) by Li, Artemiou and Li (2011). The last method was the first among many methods that have been introduced the last few years and they fall into the category of SVM-based methodology. Among other methods that have been introduced we have the Artemiou and Dong (2016), Shin et al (2017), Shin and Artemiou (2017) and among others. Using the idea of slicing from classic SDR methodology like SIR and SAVE which estimate the CS using inverse-moment-based ideas within each slice, the SVM-based algorithms estimate the CS by deriving the optimal hyperplane that separates the points between slices. Li, Artemiou and Li (2011) proposed the use of the

Science Publications

"left vs right" (LVR) algorithm when the response is continuos and the "one vs another" (OVA) when the response is categorical. In both cases, the slices that are used to construct the separating hyperplane can be highly imbalanced, that is, one may contain more points than the other. In the classification setting, where the SVM were introduced by Cortes and Vapnik (1995), this has been a well known problem that has been addressed in a number of ideas. The interested reader is referred to He and Garcia (2009) for a selection of methods to tackle imbalance. In the dimension reduction framework, Artemiou and Shu (2014) used a cost based reweighted scheme to tackle imbalance on the PSVM algorithm proposed by Li, Artemiou and Li (2011). In this paper we will expand the work by Artemiou and Dong (2016) in using a cost-based reweighted scheme to tackle imbalance in Principal Lq SVM (PLqSVM). We give a brief review of the methodology in Section 2 and then we present our new method which we call Cost-based Reweighted Principal Lq Support Vector Machines (CRPLqSVM) in Section 3. In Section 4 we will give some theoretical results and we will present numerical Studies in section 5. We will close with a small discussion section.

Literature review

There is a long literature on Sufficient Dimension Reduction (SDR) as it goes back to the introduction of Sliced Inverse Regression (SIR) by Li (1991). In this section we will focus on the literature on some of the Support Vector Machine (SVM) based literature and we will also discuss reweighting approaches.

Before introducing the methods we will discuss some notation. First of all we assume that we have a univariate response variable Y with support Ω and a p dimensional predictor vector X. If we let A_1, A_2 be two disjont subsets of Ω we can define the binary version of the response variable to be:

$$\tilde{Y} = I(Y \in A_1) - I(Y \in A_2) \tag{2}$$

where $I(\cdot)$ denotes the indicator function. Also we use the equation $\psi^{\tau}X + t = 0$ to denote the hyperplane equation where $\psi \in \mathbb{R}^p$ is the normal vector and $t \in \mathbb{R}$ is the offset. Using now the discretized version of Y, that is $\tilde{Y}$, in the classification setting discussed by Cortes and Vapnik (1995) one can find the optimal hyperplane which separates the points according to their $\tilde{Y}$ value as the set $(\psi*, t) \in \mathbb{R}^p \times \mathbb{R}$ by minimizing the following objective function at the population level:

$$\psi^{\tau}\psi + \lambda E(1 - \tilde{Y}(\psi^{\tau}(X - E(X)) - t))^{+} \tag{3}$$

where λ is a fixed tuning parameter known as the cost (or misclassification penalty) and the $a^{+} = \max\{0, a\}$.

Principal Support Vector Machines (PSVM)

Li, Artemiou and Li (2011) introduced Principal Support Vector Machines (PSVM) which takes the classic SVM algorithm we discuss above and adapts it accordingly to allow to it to be used as a dimension reduction method in the SDR framework. First the authors suggest a slight modification to the objective function above and instead they propose the minimization of the objective function:

$$\psi^{\tau}\Sigma\psi + \lambda E(1 - \tilde{Y}(\psi^{\tau}(X - E(X)) - t))^{+} \tag{4}$$

where $\Sigma = \mathrm{var}(X)$. Although we are not going to deal with the nonlinear feature extraction algorithm here the inclusion of Σ in the objective function allows for a unified framework of linear and nonlinear feature extraction.

The algorithm for PSVM can be described in the following steps:

1. We first compute the sample mean $\bar{X}$ and the sample covariance matrix $\hat{\Sigma}$.

2. We find the dividing points q_r, for $r = 1, \ldots, H-1$ where H the number of slices and we define the $H-1$ response vectors $\tilde{Y}^r = (\tilde{Y}_1^r, \ldots, \tilde{Y}_n^r)^{\tau}$ where $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$.

3. We find $(\hat{\psi}_r, \hat{t}_r)$ to be the minimizer of the objective function:

$$\psi^{\tau}\hat{\Sigma}\psi + \lambda E_n(1 - \tilde{Y}^r(\psi trans(X - \bar{X}) - t))^{+}. \tag{5}$$

4. We construct the candidate matrix $\hat{V} = \sum_{i=1}^{H-1} \hat{\psi}^r(\hat{\psi}^r)^{\tau}$ and we do an eigenvalue decomposition to get the eigenvectors $u_1, \ldots, u_d$ corresponding to the largest d eigenvalues. One can find the subspace spanned by these eigenvectors as an estimate for the CS, $\mathscr{S}_{Y|X}$.

We note here that the description above fits the "left vs right" (LVR) approach. For the OVA approach we need to adjust the way we define the discretized version of the response vector $\tilde{Yb}$. The interested reader is referred to Li, Artemiou and Li (2011) for the details. Also we note that the objective function in (9) is solved by finding the dual problem using the Karush-Kuhn-Tucker (KKT) equations in a similar manner as the original SVM algorithm in Cortes and Vapnik (1995). The dual problem is then solved by using quadratic optimization.

Principal Lq Support Vector Machines (PLqSVM)

Artemiou and Dong (2016) identified, that there was a problem with the objective function of PSVM. Due to the fact that the second part, that is $a^{+} = \max\{0, a\}$ (also

known as the hinge loss in classification literature), is not strictly convex, under some conditions on the distribution of X we may have not have a unique solution for the offset t. Although this is not affecting the estimation of the CS as it only depends on the normal vector ψ (which is always unique), it created problems with the asymptotic theory of PSVM as the gradient function, the Hessian matrix and the influence function were dependent on the value of t. This meant that one couldn't use for example the asymptotic theory of PSVM to create sequential tests for dimension determination.

In an effort to avoid this, Artemiou and Dong (2016) proposed the use of LqSVM which raises the hinge loss to the power $q \geq 2$ creating a strictly convex function which can ensure the uniqueness of t. Therefore their objective function is:

$$\psi^{\tau}\Sigma\psi + \lambda E\left[(1-\tilde{Y}(\psi^{\tau}(X-E(X))-t))^{+}\right]^{q}. \quad (6)$$

The algorithm for estimation is essentially the same. Solving this objective function is slightly more challenging though, as it is not always possible to use quadratic programming optimization. In the case that $q = 2$ though, this is possible and therefore although Artemiou and Dong (2016) developed the theory for general q, they run simulations only for the case that $q = 2$.

Cost-based Reweighted Principal Support Vector Machine (CRPSVM)

Artemiou and Shu (2014) investigated the effect of imbalanced of slices in the dimension reduction framework. In the classification framework this is a well known issue and there are a number of algorithms that have been proposed to address this (see He and Garcia (2009)). To address this Artemiou and Shu (2014) proposed an algorithm that is based on using different costs (λ's) for each slice.

In the classification framework, let's assume there is a class (minority) that have much less observations than the other class (majority). Misclassifying one point from the minority class has a much bigger effect than misclassifying an observation from the majority class. One approach that was suggested to address this is to give the minority class a much bigger penalty than the majority class (see for example Veropoulos et al. (1999)). A similar approach for the dimension reduction framework was proposed by Artemiou and Shu (2014). Imbalance happens in the PSVM algorithm due to the construction of the cutoff points q_r, $r = 1,\ldots,H-1$. To ensure, like in previous algorithms that all slices have about the same number of points $q_r = ((100/H) \times r)^{\text{th}}$ percentile. Therefore if there are for example 100 observations and 10 slices, then q_1 is the 10^{th} percentile, which means on the first iteration of the algorithm one class will have 10 points and the other 90 points. Similarly, for q_2 we will have 20 points on one class and 80 on the other. This imbalance diminishes as we move to the middle of the dataset and then starts to increase again as we move to the higher percentiles.

When combining the two costs with the PSVM objective function we get the following objective function which we call Cost-based Reweighted Principal Support Vector Machines (CRPSVM):

$$\psi^{\tau}\Sigma\psi + E\left[\lambda_{\tilde{Y}}(1-\tilde{Y}(\psi^{\tau}(X-E(X))-t))^{+}\right]. \quad (7)$$

One question is how to choose the two values for the cost. One easy approach is to use the relationship $\lambda_{-1}/\lambda_1 = n_1/n_{-1}$ where n_j represents the number of observations with $\tilde{Y} = j$ and λ_j is the cost associated with the class that represents $\tilde{Y} = j$ for $j = 1,2$.

Further studies on the reweighting in the dimension reduction framework can be found in Smallman and Artemiou (2017) who used a number of algorithmic approaches to address imbalance and Artemiou (2019) who used it to address robustness at the presence of outliers.

Cost-based Reweighted Principal Lq Support Vector Machines (CRPLqSVM)

In this paper we will address the imbalance with the Principal Lq Support Vector Machine and we will propose the Cost-based Reweighted Principal Lq Support Vector Machines (CRPLqSVM) algorithm.

Population level results

We will use a similar approach as Artemiou and Shu (2014) used for PSVM, that is, we will have address imbalance by using different costs for each class. therefore the objective function for proposed algorithm becomes

$$\psi^{\tau}\Sigma\psi + E\left[\lambda_{\tilde{Y}}(1-\tilde{Y}(\psi^{\tau}(X-E(X))-t))^{+}\right]^{q}. \quad (8)$$

The following theorem proves that the normal vector of the hyperplane that forms part of the solution of the above objective function is part of the CS under the linearity condition. The linearity condition is very common in linear feature extraction in the SDR literature.

Theorem 1 *Assume that the $E(X|\beta^{\tau}X)$ is a linear function of $\beta^{\tau}X$ and that (ψ^*,t^*) is the solution that minimizes the objective function (8) among all possible $(\psi,t) \in \mathbb{R}^p \times \mathbb{R}$. Then $\psi^* \in \mathscr{S}_{Y|X}$.*

The proof of the theorem is similar to the proof that was used in Artemiou and Shu (2014) with the only difference that we have the q^{th} power of the hinge loss on the second

term of the objective function. As it is claimed in there their result holds for any convex function that is used. Since $\left[\lambda_{\tilde{Y}}(1-\tilde{Y}(\psi^{\tau}(X-E(X))-t))^{+}\right]^{q}$ is a convex function then the theorem holds and we omit the details here.

Estimation algortihm

The algorithm for CRPLqSVM can be described in the following steps:

1. We first compute the sample mean $\bar{X}$ and the sample covariance matrix $\hat{\Sigma}$.
2. We find the dividing points q_r, for $r = 1,\ldots,H-1$ where H the number of slices and we define the $H-1$ response vectors $\tilde{Y}^r = (\tilde{Y}_1^r,\ldots,\tilde{Y}_n^r)^{\tau}$ where $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$.
3. We find $(\hat{\psi}_r, \hat{t}_r)$ to be the minimizer of the objective function:
$$\psi^{\tau}\hat{\Sigma}\psi + \lambda^{*r}E_n[(1-\tilde{Y}^r(\psi trans(X-\bar{X})-t))^{+}]^{q} \quad (9)$$
where λ^{*r} is an n-dimensional vector with the i^{th} entry $(i = 1,\ldots,n)$ the value of lambda corresponding to the class indicated by $\tilde{Y}_i^r$.
4. We construct the candidate matrix $\hat{V} = \sum_{i=1}^{H-1}\hat{\psi}^r(\hat{\psi}^r)^{\tau}$ and we do an eigenvalue decomposition to get the eigenvectors $u_1,\ldots,u_d$ corresponding to the largest d eigenvalues. One can find the subspace spanned by these eigenvectors as an estimate for the CS, $\mathscr{S}_{Y|X}$.

As with PSVM this algorithm corresponds to the LVR approach. One can easily adjust the algorithm accordingly to fit the OVA approach.

There are two things we need to address in the estimation part. The first is what type of values one will use for the λ's. We decided to use a similar approach to Artemiou and Shu (2014) who used the inverse ratio of the number of observations in each class. Let n_j where $j = -1, 1$ denote the number of observations that have $\tilde{Y}_i^r = j$. Then we use

$$\frac{\lambda_{-1}}{\lambda_1} = \frac{n_1}{n_{-1}}$$

in each of the dividing point q_r, $r = 1,\ldots,H-1$ algorithm. The second is how to solve the optimization problem in the third step of the algorithm above. Here we discuss how this can be done in general.

First of all, we note that the general sample version can also be written as:

$$\psi^{\tau}\hat{\Sigma}\psi + \frac{1}{nq}\sum_{i=1}^{n}\lambda_{\tilde{Y}_i}[(1-\tilde{Y}_i(\psi^{\tau}(X_i-\bar{X})-t))^{+}]^{q}$$

where we omit the superscript r from $\tilde{Y}_i$ for simplicity as the solution is the same for any dividing point. Now let's define $Z_i = \hat{\Sigma}^{-1/2}(X_i-\bar{X})$ and $\zeta = \Sigma^{1/2}\psi$ which means the above optimization can be written as:

$$\zeta^{\tau}\zeta + \frac{1}{nq}\sum_{i=1}^{n}\lambda_{\tilde{Y}_i}[(1-\tilde{Y}_i(\zeta transZ-t))^{+}]^{q}. \quad (10)$$

The following Proposition then gives the solution. It has been proven in the

Proposition 1 *Let $\zeta^* \in mathbbR^p$ be the minimizer of the objective function (10).Then $\zeta^* = (1/2)Z^{\tau}(\alpha\odot\tilde{Y})$ where α is the solution to the following optimization problem:*

$$\begin{aligned}\max\ \ & \alpha^{\tau}1_n - \frac{1}{4}(\alpha\odot\tilde{Y})^{\tau}ZZ^{\tau}(\alpha\odot\tilde{Y}) \\ & + \frac{1-q}{q}(D_{\lambda^*}n^{-1})^{\frac{1}{1-q}}(\alpha^{\tau})^{\frac{q}{q-1}}1_n \\ \text{subject to}\ \ & \alpha \geq 0_n,\ (\alpha\odot\tilde{Y})^{\tau}1_n = 0_n\end{aligned}$$

where $1_n = (1,\ldots,1)^{\tau} \in \mathbb{R}^n$, $0_n = (0,\ldots,0)^{\tau} \in \mathbb{R}^n$ and D_{λ^*} is an $n\times n$ diagonal matrix that has the entries of vector λ^* in the diagonal.

The proof of this proposition is very similar to Proposition 1 in Artemiou and Dong (2016) and therefore we omit it. As discussed in Artemiou and Dong (2016) who had a similar issue, this is not an easy problem to solve. In the case that $q = 2$ this becomes a quadratic optimization problem as the one solved in other SVM approaches in the literature including the dimension reduction approaches we discussed in the previous section. Therefore, using $q = 2$ and the fact that $\tilde{Y}_i = \pm 1$ which makes $\alpha^{\tau}\alpha = (\alpha\odot\tilde{Y})^{\tau}(\alpha\odot\tilde{Y})$ then the optimization problem in the above proposition simplifies to :

$$\begin{aligned}\max\ \ & \alpha^{\tau}1_n - \frac{1}{4}(\alpha\odot\tilde{Y})^{\tau}(ZZ^{\tau}+2nD_{\lambda^*}^{-1})(\alpha\odot\tilde{Y}) \\ \text{subject to}\ \ & \alpha \geq 0_n,\ (\alpha\odot\tilde{Y})^{\tau}1_n = 0_n\end{aligned}$$

which is a quadratic optimization problem and can be easily solved. We need to remember that since we have standardized the data, we have to use $\psi^* = \Sigma^{-1/2}\zeta^*$ to obtain the solution of the unstandardized problem.

Numerical Studies

In this section we discuss numerical results of this experiment.

Table 1. Performance of PL2SVM and CRPL2SVM for different dimensions of the predictor and different number of slices

		$H=10$		$H=20$		$H=50$	
Models	p	PL2SVM	CRPL2SVM	PL2SVM	CRPL2SVM	PL2SVM	CRPL2SVM
I	10	0.16 (0.043)	0.13 (0.037)	0.15 (0.041)	0.11 (0.032)	0.16 (0.048)	0.10 (0.032)
	20	0.25 (0.051)	0.20 (0.042)	0.24 (0.048)	0.17 (0.032)	0.23 (0.055)	0.15 (0.036)
	30	0.35 (0.062)	0.28 (0.053)	0.30 (0.057)	0.21 (0.049)	0.32 (0.057)	0.21 (0.048)
II	10	0.70 (0.168)	0.66 (0.154)	0.72 (0.165)	0.69 (0.137)	0.72 (0.186)	0.70 (0.176)
	20	1.06 (0.135)	1.03 (0.137)	1.02 (0.168)	0.99 (0.160)	1.03 (0.161)	1.00 (0.164)
	30	1.23 (0.120)	1.20 (0.128)	1.21 (0.126)	1.19 (0.123)	1.20 (0.137)	1.18 (0.131)
III	10	1.12 (0.225)	1.14 (0.239)	1.07 (0.234)	1.03 (0.274)	1.08 (0.220)	1.05 (0.240)
	20	1.43 (0.201)	1.41 (0.214)	1.45 (0.185)	1.41 (0.213)	1.40 (0.216)	1.36 (0.225)
	30	1.62 (0.132)	1.58 (0.145)	1.59 (0.152)	1.54 (0.156)	1.61 (0.144)	1.55 (0.168)

Simulated data

We simulate data from the following models:

$$\begin{aligned}
&\text{Model I}: Y = X_1 + X_2 + \sigma\varepsilon,\\
&\text{Model II}: Y = X_1/\{0.5 + (X_2+1)^2\} + \sigma\varepsilon,\\
&\text{Model III}: Y = X_1(X_1 + X_2 + 1) + \sigma\varepsilon,
\end{aligned}$$

where $X \sim N_p(0_p, I_p)$, $\varepsilon \sim N(0,1)$, $p = 10, 20, 30$. We also use $n = 100$, $\sigma = 0.2$, the number of slices $H = 10, 20, 50$ and we define the cutoff points q_r for $r = 1, \ldots, H-1$ to be equally spaced percentiles. To compare between the algorithms we use the distance between the projection matrices on the estimated and the real subspace, that is $\|P_\beta - P_{\hat{\beta}}\|$ where $P_\beta = \beta(\beta^\intercal\beta)^{-1}\beta^\intercal$ and $P_{\hat{\beta}} = \hat{\beta}(\hat{\beta}^\intercal\hat{\beta})^{-1}\hat{\beta}^\intercal$ and $\|\cdot\|$ is the Frobenius norm.

We compare the the PL2SVM algorithm in Artemiou and Dong (2016) with our proposed CRPL2SVM in Table 1 for $p = 10, 20, 30$ and for $H = 10, 20, 50$. As one can see generally the reweighted algorithm performs better than the original algorithm for all combinations of p and H. We also see that in most cases the higher the number of slices the best the performance of the reweighted algorithm. This is due to the fact that the imbalance is more intense the higher the number of slices are.

Real Data Analysis

In this section we use a real dataset to show the advantage of the cost reweighted method. We use the dataset on Computer Hardware (Ein-Dor and Feldmesser (1987)) from UC Irvine machine learning repository (Dua and Graff (2019)). The objective is to create a regression model that estimates relative performance of the Central Processing Unit (CPU) of a computer using some of its characteristics, including cache memory size, cycle time, minimum and maximum input/output channels, and minimum and maximum main memory. Relative performance was calculated using observations from users of different machines in the market. The dataset consists of 209 models where performance is not available. We apply both algorithms, the PL2SVM and the cost reweighted one using 10 slices. Figure 1 shows the expected nonlinear relationship with the performance in the first direction of both methods. The two directions are very strongly correlated (correlation is 0.97), but it is clear that the cost reweighted one is slightly better as the points are closer to the curve than the PL2SVM one.

Discussion

The effect of imbalance of classes in the classification setting has been studied well over the years. With the use of classification methods in the SDR framework, there is a need to study and understand the effect of imbalance in this setting. the two settings are fundamentally different as in classification we are interested to find the optimal hyperplane and reduce the misclassification rate, while in the dimension reduction setting we are only interested for a hyperplane alignment which will estimate accurately the CS. Also we note that the imbalance in the SDR framework is artificial as it depends on the way we select the number of slices, with higher number of slices leading to a more imbalance between the slices. In this work, we investigate the effect of imbalance when the LqSVM is used in the SDR framework and we see that addressing the imbalance helps in estimating the CS more accurately.

As He and Garcia (2009) have suggested there is a huge literature on addressing imbalance and we are only proposing the use of a single method here. Although, it has shown positive results a more substantial study is needed to understand the effect of imbalance on the dimension reduction framework we are discussing in this work.

References

1. Artemiou, A. (2019). Using adaptively weighted large margin classifiers for robust sufficient dimension reduction. *Statistics*. (appeared online)

2. Artemiou, A. and Dong, Y. (2016). Sufficient dimension reduction via principal L*q* support vector machines. *Electronic Journal of Statistics*, **10**, 783-805.

3. Artemiou, A. and Shu, M. (2014). A cost based reweighed scheme of principal support vector machine. In *Topics in Nonparametric Statistics*, Springer Proceedings in Mathematics and Statistics, **74**, 1-12.

4. Artemiou, A. and Tian, L. (2015). Using Sliced Inverse Mean Difference for Sufficient Dimension Reduction. *Statistics and Probability Letters*, **106**, 184-190.

5. Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

6. Cook, R.D., Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, **86**, 316-342.

7. Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 1-25.

8. Dua, D. and Graff, C. (2019) *UCI machine learning repository*. Irvine (CA), University of California, School of Information and Computer Science.

9. Ein-Dor, P. and Feldmesser, J. (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Communications ACM*, **30**, (4), 308 - 317.

10. Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, **37**, 1871-1905.

11. He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284.

12. Lee, K. K., Gunn, S. R., Harris, C. J., Reed, P. A. S. (2001). Classification of imbalanced data with transparent kernels. In: *Proceedings of International Joint Conference on Neural Networks (IJCNN 2001)*, Washington, D.C., **4**, 2410-2415.

13. Li B., Artemiou, A. and Li, L. (2011). Principal support vector machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, **39**, 3182 - 3210.

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

15. Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580 - 1616.

16. Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316 - 342.

17. Shin, S. J. and Artemiou, A. (2017). Penalized Principal Logistic Regression for Sparse Sufficient Dimension Reduction. *Computational Statistics and Data Analysis*, **111**, 48 - 58.

18. Shin, S. J., Wu, Y., Zhang, H. H. and Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*. **104**, 67 - 81.

19. Smallman, L. and Artemiou, A. (2017). A Study on Imbalance Support Vector Machine Algorithms for Sufficient Dimension Reduction. *Communications in Statistics, Theory and Methods*, **46**, 2751 - 2763

20. Veropoulos, K., Campbell, C., Cristianini, N. (1999). Controlling the sensitivity of support vector machines. In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 1999)*, Workshop ML3, Stockholm, 55 - 60.

21. Wu, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, **17**, 590 - 610.

22. Yeh, Y. R., Huang, S. Y. and Lee, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1590 - 1603.

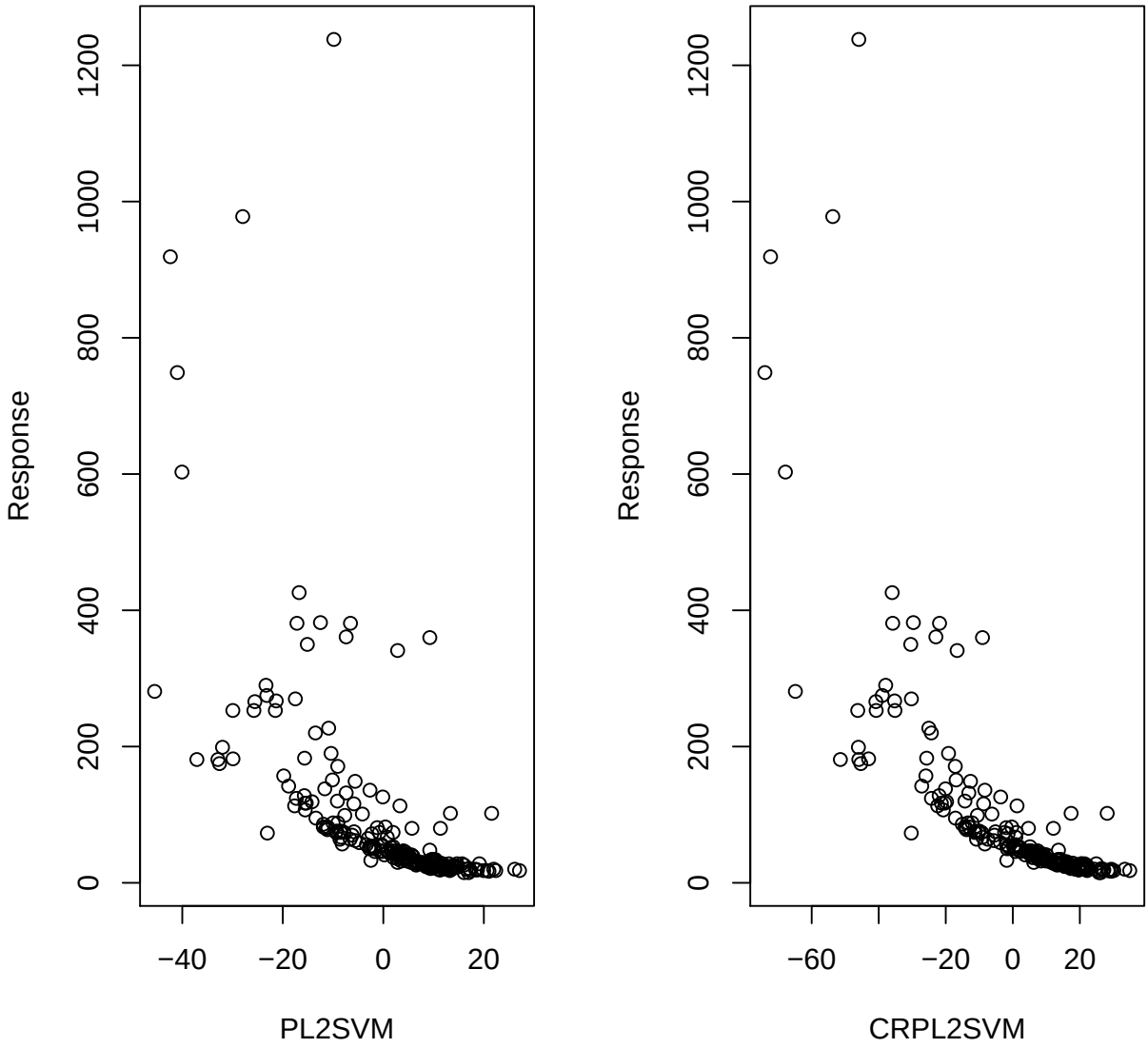

Fig. 1. Plots of the first directions for PL2SVM (left) and CRPL2SVM (right) plotted against the response variable. We can clearly see the quadratic nature of the relationship and the fact that the cost reweighted algorithm gives a stronger relationship.