Noncrossing ordinal classification[*]

XINGYE QIAO[†]

Ordinal data are often seen in real applications. Regular multicategory classification methods are not designed for this data type and a more proper treatment is needed. We consider a framework of ordinal classification which pools the results from binary classifiers together. An inherent difficulty of this framework is that the class prediction can be ambiguous due to boundary crossing. To fix this issue, we propose a noncrossing ordinal classification method which materializes the framework by imposing noncrossing constraints. An asymptotic study of the proposed method is conducted. We show by simulated and data examples that the proposed method can improve the classification performance for ordinal data without the ambiguity caused by boundary crossings.

KEYWORDS AND PHRASES: Classification, Mixed integer programming, Multivariate analysis, Statistical computing, Support vector machine.

1. INTRODUCTION

Data with ordinal class labels are very common in reality and they are collected from many scientific areas and social practices, such as disease diagnosis and prognosis, national security threat detection, and quality control. For example, the development of tumor can be classified to Stage I, Stage II, Stage III, *etc.*; the U.S. homeland security advisory system has five categories, Green, Blue, Yellow, Orange and Red, ordered from the least to the most severe threats; the quality of a randomly sampled product can be categorized to excellent, good, fair and bad. The goal of ordinal classification is to classify a data point to one of these ordinal categories, $y \in \mathcal{Y}$, based on the covariates $\boldsymbol{x} \in \mathcal{S} \subset \mathbb{R}^d$. Here we consider the case $\mathcal{Y} = \{1, 2, \ldots, K\}$. The actual labels are of no importance, so long as the order can be recognized.

Note that ordinal data are a special case of the more general multicategory data. Ignoring the order information, one may classify ordinal data in the same way as one would do multicategory data, by applying a multicategory classification method. There is a large body of literature for the latter. This includes those which combine multiple binary classifiers, such as the One-Versus-One and One-Versus-Rest paradigms (see for example Duda et al., 2001), and those which estimate multiple classification boundaries simultaneously, such as Weston and Watkins (1999), Crammer and Singer (2002), Lee et al. (2004), and Huang et al. (2013). While using multicategory classification method for ordinal data sometimes works, such treatment can be suboptimal, because the classes are treated equally without their connections and relative superiority being considered. Moreover, a counterexample in Section 2 reveals that it is desirable to use an approach which fully utilizes the ordinal information available.

Suppose there are K classes in total. A simple but very useful strategy for ordinal classification is to sequentially conduct binary classifications between the combined meta-class $\mathcal{C}_k \equiv \{1, \ldots, k\}$ and meta-class $\overline{\mathcal{C}}_k \equiv \{k+1, \ldots, K\}$, for $1 \leq k \leq K-1$, and then pool the classification results from these $(K-1)$ steps to reach a final prediction (see Frank and Hall, 2001). In binary classification, usually a discriminant function f is estimated, and data point $\boldsymbol{x}$ is classified to the positive class if $f(\boldsymbol{x}) > 0$, or to the negative class otherwise. The classification boundary is defined by $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$. As there are $(K-1)$ binary classifiers in this strategy, there are $(K-1)$ classification boundaries. This approach assumes that each class is *sandwiched* by two adjacent classification boundaries.

An inherent difficulty of this approach is that since these boundaries are trained separately, it is possible that they may cross with each other. Consequently, how to make a final conclusion becomes ambiguous for some data points.

In this article, we propose a flexible margin-based classification method for ordinal data. The direction we pursue is to construct the $(K-1)$ boundaries simultaneously. Our method is equipped with extra noncrossing constraints to fix the crossing issue, hence is named **N**oncrossing **Ordi**nal **C**lassification (NORDIC). Similar noncrossing constraints were studied and used in the quantile regression context (for example, Bondell et al., 2010; Liu and Wu, 2011). Compared to the vanilla idea of training $(K-1)$ binary classifiers separately, simultaneous learning can borrow the strength from different classes, which leads to better classification accuracy and improved robustness to mislabeled data. Moreover, compared to many existing methods, our method is more flexible, since it does not assume that the boundaries are parallel.

Among the existing related work in classifying ordinal data, Herbrich et al. (2000) tried to find the classification boundaries by maximizing the margin in the space of pairs of data vectors; Frank and Hall (2001) was among the first to consider the idea of pooling binary classifiers; Shashua and Levin (2003) generalized the support vector formulation for ordinal regression and proposed to optimize multiple thresholds to define parallel separating hyperplanes; Chu and Keerthi (2005) improved the work of Shashua and Levin (2003) and guaranteed that the thresholds were properly ordered; Chu and Ghahramani (2005) used a probabilistic kernel approach based on Gaussian processes; Cardoso and da Costa (2007) replicated the data and cast the ordinal classification problem to a single binary classification problem. Many of these approaches, although ensuring noncrossing, have posed a fairly strong assumption that the $(K-1)$ boundaries are parallel to each other (either in the original sample space or in the kernel feature space), which may be lack of flexibility and be unrealistic in many cases.

The rest of the article is organized as follows. In Section 2, we compare the multicategory classification with the ordinal one, and review a simple framework for the ordinal classification. We introduce the main idea of the NORDIC method and the computation algorithm in Section 3. A more precise version of NORDIC, which makes use of a less popular optimization algorithm, is introduced in Section 4. The theoretical properties are studied in Section 5. Several simulated examples are used to compare NORDIC with other methods in Section 6. A real data example is studied in Section 7. Concluding remarks are made in Section 8.

2. ORDINAL CLASSIFICATION

In this section, we first demonstrate, using a real example that, in some cases, it is better not to ignore the ordinal information by treating ordinal data as regular multicategory data. We then introduce a framework of ordinal classification via binary classifiers. Lastly, we compare the principles of multicategory and ordinal classifications.

2.1 An example in U.S. presidential election

In a multicategory classifier with K classes, usually K discriminant functions $g_k(\boldsymbol{x})$, $k = 1, \ldots, K$, are estimated and the class prediction for $\boldsymbol{x}$ is $\operatorname{argmax}_{k \in \{1, \ldots, K\}} g_k(\boldsymbol{x})$. Let $\eta_k(\boldsymbol{x})$ denote the conditional probability for the kth class, $\eta_k(\boldsymbol{x}) = \mathbb{P}(Y = k | \boldsymbol{X} = \boldsymbol{x})$. In this case, any multicategory classifier would aim to mimic the Bayes classification rule, $\phi_{MC}^{Bayes}(\boldsymbol{x}) \equiv \operatorname{argmax}_{k \in \{1, \ldots, K\}} \eta_k(\boldsymbol{x})$, which has the smallest conditional classification risk, $\mathbb{P}(\phi(\boldsymbol{X}) \neq Y | \boldsymbol{X} = \boldsymbol{x})$, among all possible rules.

For the ordinal data, one can opt to ignore the ordinal information and classify them using a multicategory classifier. However, a counterexample suggests that this may not always be a wise strategy. Consider the presidential election in the United States. Any voter can be viewed as being from a red state (a state which is most conservative and predominantly vote for the Republican Party), a blue state

(a state which is least conservative and predominantly vote for the Democratic Party) and a purple state (also known as a swing state, where both parties receive strong support). In 2012, the states of North Carolina, Florida, Ohio, and Virginia were the swing states. There are many more blue and red states in the U.S. than swing states (and a much larger population in the former two types of states than that in the latter). Suppose each voter is associated with a covariate vector $\boldsymbol{x} \in \mathcal{S}$ and the color of her home state is the class label. The statistical task here is to classify her to one of the three types of states, $\mathcal{Y} = \{\text{red, purple, blue}\}$.

Recall that the Bayes rule in multicategory classification classifies $\boldsymbol{x}$ to the class with the greatest $\eta_k(\boldsymbol{x})$. It is more likely for a multicategory classifier to classify a voter to a blue state or a red state, since both tend to have larger $\eta_k(\boldsymbol{x})$. To see this, note that $\eta_k(\boldsymbol{x}) = \pi_k d_k(\boldsymbol{x}) / \{\sum_{\ell \in \mathcal{Y}} \pi_\ell d_\ell(\boldsymbol{x})\}$, where $d_k(\boldsymbol{x})$ is the density of the covariate $\boldsymbol{X}$ given that she is from the kth class and π_k is the unconditional class probability for the kth class. Clearly, both π_{red} and π_{blue} are *much* greater than π_{purple}, leading to that their $\eta_k(\boldsymbol{x})$'s tend to be larger as well. The bottom line is, it seems to be unfair that the chance that a voter from the purple state is correctly identified is compromised simply because there is a smaller population in purple states. Ironically, in a U.S. presidential election, the swing states are the most important battleground, because it is the swing states that break the even in a presidential campaign.

In this example, the imbalanced class prior probabilities appear to be the proximate cause that leads to the aforementioned issue. The underlying root cause, however, is that the ordinal data nature herein has been ignored. A classification method which makes use of the ordinal information is more appropriate in this case. We describe a simple strategy for this example here which leads to the more formal methodology in the next subsection: for a randomly selected voter, we first consider classifying her to a *blue* state, versus a *purple or red* state. If she is classified to the latter, then she tends to be relatively more conservative (than blue state voters). We then classify her to a *blue or purple* state, against a *red* state. If she is classified to the former, then she is relatively less conservative (than red state voters). The results of the two comparisons can lead to the final conclusion that she is classified to a purple state.

2.2 Ordinal classification via binary classifiers

In general, consider an ordinal classification problem with K classes. Furthermore, consider $(K-1)$ binary classifiers, where the kth classification boundary separates the combined set $\{i : y_i \in \mathcal{C}_k\}$ from the combined set $\{i : y_i \in \overline{\mathcal{C}}_k\}$ where $\mathcal{C}_k \equiv \{1, \ldots, k\}$ and $\overline{\mathcal{C}}_k \equiv \{k+1, \ldots, K-1\}$. For the kth binary classification, we code the former the negative class and the latter the positive class by constructing a dummy class label $y^{(k)} \equiv -1$ if $y \leq k$ and $+1$ if $y > k$. The kth binary classifier is associated with a discriminant

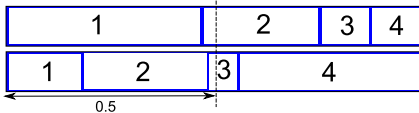|	BC I	BC II	BC III
Class 1	✗	✓	✓
Class 2	✓	✓	✓
Class 3	✓	✗	✓
Class 4	✓	✗	✗

Figure 1. In the top panel, the multicategory classification principle chooses Class 1 while the ordinal classification principle chooses Class 2. In the bottom panel, the multicategory classification principle chooses Class 4 while the ordinal classification principle chooses Class 3.

function $f_k(\boldsymbol{x})$ so that the classification rule is $\text{sign}\{f_k(\boldsymbol{x})\}$. Let $Z_k(\boldsymbol{x})$ denote the prediction set of observation $\boldsymbol{x}$ with respect to the kth subproblem, defined as $\mathcal{C}_k$ if $f_k(\boldsymbol{x}) < 0$, or $\overline{\mathcal{C}}_k$ otherwise. The final prediction for $\boldsymbol{x}$, aggregating all the results from the $(K-1)$ binary classifiers above, will be the intersection of $Z_k(\boldsymbol{x})$, $i.e.$, $\bigcap_{1 \leq k \leq K-1} Z_k(\boldsymbol{x})$.

In a four-class toy example, Table 1 tabulates the prediction of the three binary classifiers for some observation $\boldsymbol{x}$. The first binary classifier compares Class 1 and the metaclass $\{2,3,4\}$. The prediction is that the observation is from $\{2,3,4\}$. Similarly, the second binary classifier compares $\{1,2\}$ and $\{3,4\}$ and the prediction is $\{1,2\}$. Lastly, the third binary classifier classifies the observation $\boldsymbol{x}$ to $\{1,2,3\}$. Clearly, Class 2 is favored by all three binary classifiers and it is the final prediction for $\boldsymbol{x}$. This framework for reaching an ordinal classification prediction by pooling binary classifiers was first noted by Frank and Hall (2001).

2.3 Principle of ordinal classification

We are now ready to compare the principles of multicategory classification and ordinal classification. An illustration in Figure 1 can tellingly demonstrate the distinction between these principles. In a data set with $K = 4$, there are two example data points (shown in the top and the bottom rows respectively). For each data point, the length of each block denotes the conditional class probability $\eta_k(\boldsymbol{x})$. The sum of all four conditional probabilities is 1. The principle in multicategory classification chooses Class 1 in the top example and Class 4 in the bottom example, as they correspond

to the greatest $\eta_k(\boldsymbol{x})$ in both cases. In contrast, in ordinal classification, the desired prediction would be Class 2 and Class 3 respectively. For instance, for the top example, the data point is more likely from Class $\{1,2\}$ than from Class $\{3,4\}$, and more likely from Class $\{2,3,4\}$ than from Class $\{1\}$. Hence, Class 2 is the most plausible choice for this data point. Similarly, the data point in the bottom is most likely from Class 3. In particular, they both correspond to Class k such that $\sum_{\ell=1}^{k-1} \eta_\ell(\boldsymbol{x}) < 1/2$ and $\sum_{\ell=1}^{k} \eta_\ell(\boldsymbol{x}) > 1/2$ for each $\boldsymbol{x}$. In the illustration, a vertical line corresponding to 0.5 cuts the blocks for the desired predictions.

A useful notion here is that the principle of multicategory classification is to select the 'mode' of the class labels, based on $\eta_k(\boldsymbol{x})$, while that of the ordinal classification is to select the 'median'.

3. NONCROSSING ORDINAL CLASSIFICATION

Conducting ordinal classification via binary classifiers is very easy to implement as long as one has access to an efficient binary classifier. There are many options, such as Support Vector Machine (SVM; Cortes and Vapnik, 1995; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000), Distance Weighted Discrimination (DWD; Marron et al., 2007; Qiao et al., 2010), hybrids of the two (Qiao and Zhang, 2015b,a), ψ-learner (Shen et al., 2003), Large-Margin Unified Machines (Liu et al., 2011) and so on.

However, because the $(K-1)$ classification boundaries are trained separately, it is possible that they cross with each other. Figure 2 is an illustration which shows the possible crossing between classification boundaries. Here there are four classes (annotated as 1, 2, 3 and 4) and three estimated classification boundaries (I, II and III). The second and the third estimated boundaries cross with each other. Consequently, the red star point cannot be classified properly. In particular, it will be classified by classifier I to $\{2,3,4\}$, by classifier II to $\{1,2\}$ and by classifier I to $\{4\}$. The intersection of all three prediction sets is empty. Although one may argue that this point might be Class 2 or Class 4, no definite answer can be given, and there is an ambiguity as to how to classify this red star point.

Hence, it is desired that the estimated classification boundaries do not cross with each other. Let $f_k(\boldsymbol{x})$ be the discriminant function for the kth binary classification. Recall that its boundary is defined by $\{\boldsymbol{x} : f_k(\boldsymbol{x}) = 0\}$. For these boundaries to be noncrossing, mathematically, it is equivalent that for all $\boldsymbol{x} \in \mathcal{S}$ not on any boundary, where $\mathcal{S}$ is a subset of $\mathbb{R}^d$, there exists $k \in \{1, 2, \ldots, K-1\}$, such that $f_\ell(\boldsymbol{x}) > 0$ for all $\ell < k$ and $f_\ell(\boldsymbol{x}) < 0$ for all $\ell \geq k$. Let $S(\boldsymbol{x}, k) \equiv \text{sign}\{f_k(\boldsymbol{x})\}$. Then the condition above is the same as that $S(\boldsymbol{x}, k)$ is a monotonically decreasing function with respect to k for any fixed $\boldsymbol{x} \in \mathcal{S}$,

$$(1) \qquad S(\boldsymbol{x}, 1) \geq S(\boldsymbol{x}, 2) \geq \cdots \geq S(\boldsymbol{x}, K-1).$$
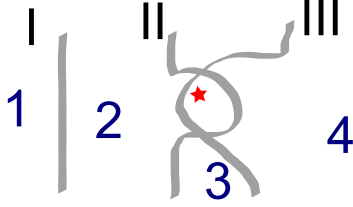
Figure 2. An illustration showing the possible crossing between estimated classification boundaries. Four classes of data (annotated as 1, 2, 3 and 4) with three estimated classification boundaries (I, II and III). Their true noncrossing boundaries are implied by their locations and are not shown. The second and the third estimated boundaries cross with each other. Consequently, the red star point cannot be classified properly.

3.1 Direct NORDIC

The noncrossing condition (1) can be fairly difficult to implement. We consider a sufficient condition first in this subsection. In this article, we use SVM as the basic binary classifier. For a Mercer kernel function $K(\cdot, \cdot)$, the Representer Theorem (Kimeldorf and Wahba, 1971) allows the kth classification function to be represented by $f_k(\boldsymbol{x}) = \sum_{j=1}^n \omega_{k,j} K(\boldsymbol{x}_j, \boldsymbol{x}) + b_k$.

Note that if we add the constraints that

$$\omega_{k,i} \geq \omega_{k+1,i} \text{ and } b_k \geq b_{k+1} \text{ for } k = 1, \ldots, K-2,$$

then as long as the kernel function is always nonnegative with $K(\cdot, \cdot) \geq 0$ (which is true for many kernel functions such as the Gaussian radial basis function kernel), we will have $f_k(\boldsymbol{x}) \geq f_{k+1}(\boldsymbol{x})$, and hence $S(\boldsymbol{x}, k) \geq S(\boldsymbol{x}, k+1)$ for any $\boldsymbol{x} \in \mathcal{S}$.

Hence we consider a direct approach to NORDIC, called NORDIC-0, by solving the following joint optimization problem with the extra noncrossing constraints (3)–(4):

$$(2) \quad \min_{\omega_{k,j}, b_k} \sum_{k=1}^{K-1} \left[\sum_{i=1}^n \left(1 - y_i^{(k)} f_k(\boldsymbol{x})\right)_+ + \frac{\lambda}{2} \boldsymbol{\omega}_k^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} \right],$$

where $f_k(\boldsymbol{x}) = \sum_{j=1}^n \omega_{k,j} K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b_k$, the coefficient vector for the kth function is $\boldsymbol{\omega}_{k,\cdot} \equiv (\omega_{k,1}, \ldots, \omega_{k,n})^T$, and $\boldsymbol{K}$ is an n by n matrix whose (i,j)th entry is $\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, subject to

$$(3) \quad b_k \geq b_{k+1}, \text{ for } k = 1, \ldots, K-2,$$

$$(4) \quad \omega_{k,i} \geq \omega_{k+1,i}, \text{ for } i = 1, \ldots n, \ k = 1, \ldots, K-2.$$

Here $\boldsymbol{\omega}_k^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot}$ is the regularization term for the kth discriminant function.

The term inside the square bracket of (2) is the objective function of kernel SVM corresponding to the kth classifier. We try to minimize the sum of these $(K-1)$ objective functions with the extra noncrossing constraints (3)–(4).

3.2 Indirect NORDIC

The constraints (3)–(4) for NORDIC-0 are sufficient conditions for noncrossing boundaries. However, such condition may be too strong. A weaker, but *almost* sufficient set of conditions would be inequality (3) along with the inequality that $\boldsymbol{K}\boldsymbol{\omega}_{k\cdot} \geq \boldsymbol{K}\boldsymbol{\omega}_{(k+1)\cdot}$, for $k = 1, \ldots, K-2$. Note that they ensure that $f_k(\boldsymbol{x}_i) \geq f_{k+1}(\boldsymbol{x}_i)$ for all the data $\boldsymbol{x}_i$ in the training data set. Thus when the training data is rich enough to cover the base of $\mathcal{S}$, then they are *almost* sufficient conditions for noncrossing. This approach is an indirect approach to noncrossing through the training data points, which is called NORDIC-1 in this article. A bonus of this set of constraints compared to (3)–(4) is that one does not need to take the inverse of $\boldsymbol{K}$ later in the implementation, which we will explain in the next subsection.

Let $\boldsymbol{y}_{k\cdot} = (y_1^{(k)}, \ldots y_n^{(k)})^T$ be the dummy class label vector of the n observations for the kth classifier, and $\boldsymbol{e} = (1, \ldots 1)^T$. For neatness, we let $\boldsymbol{Y}_{k\cdot}$ denote the diagonal matrix with $\boldsymbol{y}_{k\cdot}$ as its diagonal elements, *i.e.*, $\boldsymbol{Y}_{k\cdot} \equiv \mathrm{diag}(\boldsymbol{y}_{k\cdot})$. By replacing the Hinge loss $\{1 - y_i^{(k)} f_k(\boldsymbol{x}_i)\}_+$ in (2) by a slack variable $\xi_{k,i} \geq 0$, and incorporating the new constraints, we can write the optimization problem for NORDIC-1 as,

$$(5) \quad \min_{\boldsymbol{\omega}_{k\cdot}, b_k, \boldsymbol{\xi}_{k\cdot}} \sum_{k=1}^{K-1} \left(\frac{1}{2} \boldsymbol{\omega}_k^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + C \boldsymbol{e}^T \boldsymbol{\xi}_{k\cdot} \right),$$

subject to

$$(6) \quad \boldsymbol{e} - \boldsymbol{Y}_{k\cdot}(\boldsymbol{K}\boldsymbol{\omega}_{k\cdot} + b_k \boldsymbol{e}) \leq \boldsymbol{\xi}_{k\cdot}, \text{ for } k = 1, \ldots, K-1,$$

$$(7) \quad \boldsymbol{\xi}_{k\cdot} \geq \boldsymbol{0}, \text{ for } k = 1, \ldots, K-1,$$

$$(8) \quad b_k \geq b_{k+1}, \text{ for } k = 1, \ldots, K-2,$$

$$(9) \quad \boldsymbol{K}\boldsymbol{\omega}_{k\cdot} \geq \boldsymbol{K}\boldsymbol{\omega}_{(k+1)\cdot}, \text{ for } k = 1, \ldots, K-2.$$

3.3 Implementations of NORDIC

We start off by deriving the Wolfe duality of the optimization problem for NORDIC-1. The implementation of NORDIC-0 will come clearer later as a variant of that of NORDIC-1. We introduce nonnegative Lagrange multipliers $\boldsymbol{\alpha}_{k\cdot} = (\alpha_{k,1}, \ldots \alpha_{k,n})^T \in \mathbb{R}_+^n$, $\boldsymbol{\zeta}_{k\cdot} = (\zeta_{k,1}, \ldots \zeta_{k,n})^T \in \mathbb{R}_+^n$, $\gamma_k \in \mathbb{R}_+$ and $\boldsymbol{\varphi}_{k\cdot} = (\varphi_{k,1}, \ldots \varphi_{k,n})^T \in \mathbb{R}_+^n$ for the constraints (6), (7), (8) and (9) respectively. The Lagrangian for the primal problem (5)–(9) is,

$$\begin{aligned} \mathcal{L} = \sum_{k=1}^{K-1} \Bigg[& \left(\frac{1}{2} \boldsymbol{\omega}_k^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + C \boldsymbol{e}^T \boldsymbol{\xi}_{k\cdot} \right) \\ & + \boldsymbol{\alpha}_k^T \{ \boldsymbol{e} - \boldsymbol{Y}_{k\cdot}(\boldsymbol{K}\boldsymbol{\omega}_{k\cdot} + b_k \boldsymbol{e}) - \boldsymbol{\xi}_{k\cdot} \} \\ & - \boldsymbol{\zeta}_k^T \boldsymbol{\xi}_{k\cdot} - \gamma_k (b_k - b_{k+1}) \mathbb{1}_{\{k \neq K-1\}} \\ & - \boldsymbol{\varphi}_k^T (\boldsymbol{K}\boldsymbol{\omega}_{k\cdot} - \boldsymbol{K}\boldsymbol{\omega}_{(k+1)\cdot}) \mathbb{1}_{\{k \neq K-1\}} \Bigg]. \end{aligned}$$

It can be rearranged, so that in the square bracket, the subscripts for the primal variables are with the same index k,

as follows,

$$\mathcal{L} = \sum_{k=1}^{K-1} \left[\left(\frac{1}{2} \boldsymbol{\omega}_{k\cdot}^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + C \boldsymbol{e}^T \boldsymbol{\xi}_{k\cdot} \right) \right. \tag{10}$$
$$+ \boldsymbol{\alpha}_{k\cdot}^T \{ \boldsymbol{e} - \boldsymbol{Y}_{k\cdot} (\boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + b_k \boldsymbol{e}) - \boldsymbol{\xi}_{k\cdot} \}$$
$$- \boldsymbol{\zeta}_{k\cdot}^T \boldsymbol{\xi}_{k\cdot} - b_k \left(\gamma_k \mathbb{1}_{\{k \neq K-1\}} - \gamma_{k-1} \mathbb{1}_{\{k \neq 1\}} \right)$$
$$\left. - \boldsymbol{\omega}_{k\cdot}^T \boldsymbol{K} \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right) \right].$$

The Karush-Kuhn-Tucker (KKT) conditions for the primal problem require the following:

$$\boldsymbol{0} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\omega}_{k\cdot}} = \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} - \boldsymbol{K} \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot} \tag{11}$$
$$- \boldsymbol{K} \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right),$$

$$0 = \frac{\partial \mathcal{L}}{\partial b_k} = -\boldsymbol{y}_{k\cdot}^T \boldsymbol{\alpha}_{k\cdot} \tag{12}$$
$$- \left(\gamma_k \mathbb{1}_{\{k \neq K-1\}} - \gamma_{k-1} \mathbb{1}_{\{k \neq 1\}} \right),$$

$$\boldsymbol{0} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}_{k\cdot}} = C \boldsymbol{e} - \boldsymbol{\alpha}_{k\cdot} - \boldsymbol{\zeta}_{k\cdot}. \tag{13}$$

Once the KKT conditions (12) and (13) are inserted to (10), the items that are associated with b_k and $\boldsymbol{\xi}_{k\cdot}$ will be eliminated. Moreover, from (11), we have

$$\boldsymbol{K} \boldsymbol{\omega}_{k\cdot} = \boldsymbol{K} \left\{ \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot} + \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right) \right\},$$

which leads to

$$\boldsymbol{\omega}_{k\cdot} = \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot} + \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right)$$

when $\boldsymbol{K}$ is full rank. Let

$$\boldsymbol{R} = \left[\operatorname{diag} \{ \boldsymbol{Y}_{k\cdot} \}_{1 \leq k \leq K-1} \mid \mathbf{I}_{n(K-1)}^{(n)} \right.$$
$$\left. - \mathbf{I}_{n(K-1)}^{(-n)} \mid \boldsymbol{0}_{n(K-1) \times (K-2)} \right]$$

and $\boldsymbol{\theta} = (\boldsymbol{\alpha}; \boldsymbol{\varphi}; \boldsymbol{\gamma})$, where for a $m \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{A}^{(s)}$ denotes a $(m+s) \times n$ matrix whose upper m rows are $\boldsymbol{A}$ and the bottom s rows are all 0, and $\boldsymbol{A}^{(-s)}$ denotes a $(m+s) \times n$ matrix whose bottom m rows are $\boldsymbol{A}$ and the top s rows are all 0. Summarizing all these conditions, we can see that the optimality of the primal problem is given by the dual problem,

$$\max_{\boldsymbol{\theta}_k \equiv (\boldsymbol{\alpha}_{k\cdot}; \boldsymbol{\varphi}_{k\cdot}; \gamma_k)} - \frac{1}{2} \boldsymbol{\theta}^T \left\{ \boldsymbol{R}^T (\mathbf{I}_{K-1} \otimes \boldsymbol{K}) \boldsymbol{R} \right\} \boldsymbol{\theta} + \boldsymbol{e}^T \boldsymbol{\alpha},$$
$$\text{subject to} \quad -\boldsymbol{y}_{k\cdot}^T \boldsymbol{\alpha}_{k\cdot} - \left(\gamma_k \mathbb{1}_{\{k \neq K-1\}} - \gamma_{k-1} \mathbb{1}_{\{k \neq 1\}} \right) = 0,$$
$$\boldsymbol{0} \leq \boldsymbol{\alpha}_{k\cdot} \leq C \boldsymbol{e}, \ \boldsymbol{\varphi}_{k\cdot} \geq \boldsymbol{0}, \ \gamma_k \geq 0,$$

where $\otimes$ is the Kronecker product.

The dual problem above is nothing but a quadratic programming (QP) problem about $\boldsymbol{\alpha}_{k\cdot}, \boldsymbol{\varphi}_{k\cdot}, \gamma_k$ with equality and bound-inequality constraints, which can be solved by many third-party off-the-shelf QP subroutines. More efficient implementations, such as Platt's SMO (Platt, 1999), are possible, but is not explored here as it is beyond the scope of this paper.

The optimal primal variables $\boldsymbol{\omega}$ are calculated from the optimal dual variables using the relation $\boldsymbol{\omega}_{k\cdot} = \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot} + \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right)$. By the KKT complementary conditions, the bias term b_k for the kth classifier can be found from any $\boldsymbol{x}_i$ in the training data with $0 \leq \alpha_{k,i} \leq C$, due to the relations that $1 - y_i^{(k)} \{ \sum_{j=1}^n \omega_{kj} K(\boldsymbol{x}_j, \boldsymbol{x}_i) + b_k \} = 0$. Alternatively, one can fix the $\boldsymbol{\omega}$'s in the primal (5) as known and minimize (5)–(9) with respect to b_k and $\boldsymbol{\xi}_{k\cdot}$. This would lead to a linear programming problem.

The implementation for NORDIC-0 is similar, except that the Lagrangian is

$$\mathcal{L}_0 = \sum_{k=1}^{K-1} \left[\left(\frac{1}{2} \boldsymbol{\omega}_{k\cdot}^T \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + C \boldsymbol{e}^T \boldsymbol{\xi}_{k\cdot} \right) \right.$$
$$+ \boldsymbol{\alpha}_{k\cdot}^T \{ \boldsymbol{e} - \boldsymbol{Y}_{k\cdot} (\boldsymbol{K} \boldsymbol{\omega}_{k\cdot} + b_k \boldsymbol{e}) - \boldsymbol{\xi}_{k\cdot} \}$$
$$- \boldsymbol{\zeta}_{k\cdot}^T \boldsymbol{\xi}_{k\cdot} - \gamma_k (b_k - b_{k+1}) \mathbb{1}_{\{k \neq K-1\}}$$
$$\left. - \boldsymbol{\varphi}_{k\cdot}^T \underline{(\boldsymbol{\omega}_{k\cdot} - \boldsymbol{\omega}_{(k+1)\cdot})} \mathbb{1}_{\{k \neq K-1\}} \right].$$

The only difference of the Lagrangian of NORDIC-0 from that of NORDIC-1 is underlined. Consequently, the KKT conditions are almost the same, except that,

$$\boldsymbol{0} = \frac{\partial \mathcal{L}_0}{\partial \boldsymbol{\omega}_{k\cdot}} = \boldsymbol{K} \boldsymbol{\omega}_{k\cdot} - \boldsymbol{K} \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot}$$
$$- \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right).$$

This leads to $\boldsymbol{\omega}_{k\cdot}$ at the optimality being

$$\boldsymbol{\omega}_{k\cdot} = \boldsymbol{Y}_{k\cdot} \boldsymbol{\alpha}_{k\cdot} + \boldsymbol{K}^{-1} \left(\boldsymbol{\varphi}_{k\cdot} \mathbb{1}_{\{k \neq K-1\}} - \boldsymbol{\varphi}_{(k-1)\cdot} \mathbb{1}_{\{k \neq 1\}} \right),$$

assuming that $\boldsymbol{K}$ is invertible. The rest of the implementation is identical to that in NORDIC-0, except that we let

$$\boldsymbol{R} = \left[\operatorname{diag} \{ \boldsymbol{Y}_{k\cdot} \}_{1 \leq k \leq K-1} \mid \{ \mathbf{I}_{K-1} \otimes \boldsymbol{K}^{-1} \}^{(n)} \right.$$
$$\left. - \{ \mathbf{I}_{K-1} \otimes \boldsymbol{K}^{-1} \}^{(-n)} \mid \boldsymbol{0}_{n(K-1) \times (K-2)} \right]$$

and $\boldsymbol{\theta} = (\boldsymbol{\alpha}; \boldsymbol{\varphi}; \boldsymbol{\gamma})$.

4. EXACT NORDIC VIA INTEGER PROGRAMMING

Recall that the necessary and sufficient condition for noncrossing (1) is that the sign of $f_k(\boldsymbol{x})$, $S(\boldsymbol{x}, k)$, is a monotonically decreasing function with respect to k for any fixed $\boldsymbol{x} \in \mathcal{S}$, $S(\boldsymbol{x}, 1) \geq S(\boldsymbol{x}, 2) \geq \cdots \geq S(\boldsymbol{x}, K-1)$. The constraints for NORDIC-0 and NORDIC-1 that we have discussed in the last section is sufficient to ensure that $f_1(\boldsymbol{x}) \geq f_2(\boldsymbol{x}) \geq \cdots \geq f_{K-1}(\boldsymbol{x})$, which ultimately ensures noncrossing. However, they are not the weakest sufficient

conditions we can impose. As a matter of fact, the discriminant functions f_k themselves need not to be monotonically decreasing with respect to k in order for noncrossing. In this section, we explore an idea which aims for exact noncrossing by posing conditions on the sign of the discriminant functions.

For each $\boldsymbol{x} \in \mathcal{S}$, there are one out of two alternative situations with regard to the prediction result from a discriminant function f_k: either $f_k(\boldsymbol{x}) < 0$ or $f_k(\boldsymbol{x}) \geq 0$. According to the noncrossing condition (1), the former implies that $f_{k+1}(\boldsymbol{x}) < 0$ (recall that the sign is monotonically decreasing in k). Thus, the noncrossing condition (1) is logically equivalent to the condition that at least one of the following two constraints is satisfied,

$$\text{(i) } f_k(\boldsymbol{x}) \geq 0, \text{ and (ii) } f_{k+1}(\boldsymbol{x}) \leq 0;$$

i.e., (i) and (ii) cannot be both false. Specifically, if (i) is not true, *i.e.*, if $f_k(\boldsymbol{x}) < 0$, then (ii) is true. This leads to the noncrossing condition.

Such logical implication can be modeled by the following *Logical Constraints* which involve binary integer variables $z_{1k}, \ z_{2k} \in \{0, 1\}$,

$$-f_k(\boldsymbol{x}) - M_1 z_{1k} \leq 0,$$
$$f_{k+1}(\boldsymbol{x}) - M_2 z_{2k} \leq 0,$$
$$z_{1k} + z_{2k} \leq 1,$$

where M_1 and M_2 are two large numbers due to technicality. In particular, $z_{1k} + z_{2k} \leq 1$ implies that at least one between z_{1k} and z_{2k} has to be zero, hence (considering the first two constraints) either $-f_k(\boldsymbol{x}) \leq 0$ or $f_{k+1}(\boldsymbol{x}) \leq 0$, or both are true; this is the noncrossing condition discussed above. Note that if both z_{1k} and z_{2k} were 1, then the first two constraints became $-f_k(\boldsymbol{x}) \leq M_1$ and $f_{k+1}(\boldsymbol{x}) \leq M_2$, which would essentially impose no constraint on $f_k(\boldsymbol{x})$ and $f_{k+1}(\boldsymbol{x})$ so that the undesired case that $f_k(\boldsymbol{x}) < 0$ and $f_{k+1}(\boldsymbol{x}) > 0$ may occur. See Bradley et al. (1977) for an introduction to integer programming. We can use this technique to model the noncrossing constraints. In particular, we seek to

$$(14) \quad \min_{\omega_{k,j}, b_k} \sum_{k=1}^{K-1} \left[\sum_{i=1}^{n} \left(1 - y_i^{(k)} f_k(\boldsymbol{x}) \right)_+ + \lambda \|\boldsymbol{\omega}_{k\cdot}\|_1 \right],$$

subject to

$$(15) \quad -f_k(\boldsymbol{x}_i) - M_1 z_{1ik} \leq 0,$$
$$(16) \quad f_{k+1}(\boldsymbol{x}_i) - M_2 z_{2ik} \leq 0,$$
$$(17) \quad z_{1ik} + z_{2ik} \leq 1,$$
$$(18) \quad z_{1ik}, \ z_{2ik} \in \{0, 1\},$$

for $i = 1, 2, \ldots, n$ and $k = 1, \ldots, K - 2$. This method is referred to as NORDIC-2 in this article. Here the constrains (15)–(18) are almost sufficient and (exactly) necessary conditions to *noncrossing*. It is again not exactly sufficient be-

cause we impose the constraints to all the training data vectors, instead of all $\boldsymbol{x} \in \mathcal{S}$, similar to the case of NORDIC-1. However, again, if the data vectors in the training data are rich enough, noncrossing across the board can be expected. These conditions are weaker than those in NORDIC-0 and NORDIC-1 because they ensure the monotonicity of the sign of f_k, rather than the value of f_k itself.

Note that the objective function of NORDIC-2 is a little different from those of NORDIC-0 and NORDIC-1, especially in the use of the L_1 norm penalty. We choose not to use the more common L_2 penalty, which leads to a quadratic objective function in SVM, because it is rather difficult to solve a mixed integer programing problem with quadratic objective function. In fact, we are not aware of an efficient off-the-shelf computing freeware which solves such a problem. In order to show the usefulness of the new noncrossing constraints, which is the main point of this article, we choose to use the L_1 penalty for computational simplicity.

It is worth noting that so long as there is an efficient mixed integer programming package which is capable of dealing with quadratic objective functions, an extension will be very natural and readily available.

Indeed, integer programming can solve such nonstandard problem which traditional optimization methods such as QP or linear programming cannot. However, integer programming can been overlooked by statisticians for a long time (probably due to the high computational cost and few statistical problem that this method applies). To the author's best knowledge, this article is one of only a few works in the statistical literature which employs the integer programming technique. See Liu and Wu (2006) for another instance which uses mixed integer programming to solve a statistical problem.

5. THEORETICAL PROPERTIES

In this section, we study two aspects of the theoretical properties of NORDIC. The first subsection is about the Bayes rule and Fisher consistency of the loss function in ordinal classification. The second one pertains to the asymptotic normality of the NORDIC solution.

5.1 Bayes rule and Fisher consistency

For binary classification, a classifier with loss $V_1(yf(\boldsymbol{x})) : \mathbb{R} \mapsto \mathbb{R}_+$ is Fisher consistent if the minimizer of $\mathbb{E}[V_1(Yf(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]$ has the same sign as $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) - 1/2$. The latter is the Bayes rule for binary classification. Intuitively, Fisher consistency requires that the classifier yields the Bayes decision rule asymptotically. See Lin (2004) for Fisher consistency of binary large margin classifiers.

In multicategory classification, a classifier with loss function $V_2(y, \boldsymbol{f}(\boldsymbol{x})) : \mathbb{R} \times \mathbb{R}^K \mapsto \mathbb{R}_+$, where $\boldsymbol{f}(\boldsymbol{x}) : \mathcal{S} \mapsto \mathbb{R}^K$ is the K discriminant functions, is Fisher consistent if the minimizer of $\mathbb{E}[V_2(Y, f(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]$, $\boldsymbol{g}^*(\boldsymbol{x}) = (g_1^*(\boldsymbol{x}), \ldots, g_K^*(\boldsymbol{x}))^T$, satisfies that $\operatorname{argmax}_{k \in \{1, \ldots, K\}} g_k^*(\boldsymbol{x}) =$

argmax$_{k\in\{1,\dots,K\}}\eta_k(\boldsymbol{x})$. Here, argmax$_{k\in\{1,\dots,K\}}\eta_k(\boldsymbol{x})$ is the Bayes classification rule for multicategory classification. See, for example, Liu (2007) for some discussions on Fisher consistency for multicategory SVM classifiers.

Below we formally define the Bayes rule and Fisher consistency for ordinal classification. The Bayes rule for ordinal classification is $\phi_{OC}^{Bayes}(\boldsymbol{x}) = k$ where k is such that $\sum_{\ell=1}^{k-1}\eta_\ell(\boldsymbol{x}) < 1/2$ and $\sum_{\ell=1}^{k}\eta_\ell(\boldsymbol{x}) > 1/2$. This rule guarantees that each component binary classification in ordinal classification yields the Bayes rule in the binary sense.

DEFINITION 1. *(Generalized Fisher consistency for ordinal classification)* An ordinal classification method with loss function $V_3(\cdot,\cdot)$ is Generalized Fisher consistent if for any $\boldsymbol{x}$, the minimizer $\boldsymbol{f}^*(\boldsymbol{x}) = (f_1^*(\boldsymbol{x}),\dots,f_{K-1}^*(\boldsymbol{x}))^T$ of

$$\mathbb{E}\left[\sum_{k=1}^{K-1} V_3(Y^{(k)}, f_k(\boldsymbol{X}))\right]$$

satisfies that $\mathrm{sign}(f_k^*(\boldsymbol{x})) = \mathrm{sign}(1/2 - \sum_{\ell=1}^{k}\eta_\ell(\boldsymbol{x}))$ for $k = 1,\dots,K-1$. Here $Y^{(k)}$ is the dummy class label for Class Y in the kth binary classification subproblem.

Generalized Fisher consistency means that the $(K-1)$ discriminant functions $f_1^*(\boldsymbol{x}),\dots,f_{K-1}^*(\boldsymbol{x})$ jointly trained under the loss function V_3, is essentially the same as the Bayes rule $\phi_{OC}^{Bayes}(\boldsymbol{x})$, as $n\to\infty$. Note that $\phi_{OC}^{Bayes}(\boldsymbol{x})$ has the smallest risk with respect to the aggregated 0-1 loss for the $(K-1)$ binary subproblems. Hence it is also the one which has the smallest risk under the so-called distance loss, defined as $L(\phi,y) = |\phi - y|$ (see Qiao, 2015).

Because of the use of the Hinge loss function for SVM (which is Fisher consistent in the binary sense), our NORDIC method is Generalized Fisher consistent for ordinal classification. The proof is omitted.

5.2 Asymptotic normality of linear NORDIC

When the kernel function $K(\boldsymbol{x}_1,\boldsymbol{x}_2) = \boldsymbol{x}_1^T\boldsymbol{x}_2$, that is, the linear kernel, we can have the following linear NORDIC classifier, with the objective function,

$$(19)\quad \sum_{k=1}^{K-1}\left[\sum_{i=1}^{n}\left(1 - y_i^{(k)}\left(\boldsymbol{x}_i^T\boldsymbol{\omega}_{k\cdot} + b_k\right)\right)_+ + \frac{\lambda}{2}\boldsymbol{\omega}_{k\cdot}^T\boldsymbol{\omega}_{k\cdot}\right],$$

and one of the two following sets of constraints that correspond to NORDIC-1 and NORDIC-2 respectively,

$$\boldsymbol{x}_i^T\boldsymbol{\omega}_{k\cdot} + b_k \geq \boldsymbol{x}_i^T\boldsymbol{\omega}_{k+1,\cdot} + b_{k+1},$$

and

$$-\left(\boldsymbol{x}_i^T\boldsymbol{\omega}_{k\cdot} + b_k\right) - M_1 z_{1k} \leq 0,$$
$$\left(\boldsymbol{x}_i^T\boldsymbol{\omega}_{k+1,\cdot} + b_{k+1}\right) - M_2 z_{2k} \leq 0,$$
$$z_{1k} + z_{2k} \leq 1,$$
$$z_{1k},\ z_{2k} \in \{0,1\},$$

for $i = 1,2,\dots,n$ and $k = 1,\dots,K-2$.

Because linear kernel could be negative, the NORDIC-0 method cannot be directly extended to the linear kernel case. We can use the technique in Liu and Wu (2011) to create a new kernel that satisfies the nonnegativity assumption essential for NORDIC-0. In this subsection, we prove the asymptotic normality of linear NORDIC.

Koo et al. (2008) has provided a Bahadur representation of the linear SVM and proved its asymptotic normality under some conditions. In particular, they have shown that $(\tilde{\boldsymbol{\omega}},\tilde{b})^T - (\boldsymbol{\omega}^0,b^0)^T = O_p(n^{-1/2})$, where $(\tilde{\boldsymbol{\omega}},\tilde{b})$ are the solution to the SVM classifier and $(\boldsymbol{\omega}^0,b^0)$ are the minimizer of the expected loss function.

Theorem 1 below shows that the limiting distribution of the constrained NORDIC solution has the same limiting distribution to the unconstrained binary SVM classifiers. To prove this result, we need all the regularity conditions in Koo et al. (2008).

THEOREM 1. *For $1 \leq k \leq K-1$, let $(\hat{\boldsymbol{\omega}}_{k\cdot},\hat{b}_k)$ and $(\tilde{\boldsymbol{\omega}}_{k\cdot},\tilde{b}_k)$ be the constrained and unconstrained solutions, respectively, to the kth binary linear SVM problem in (19). Assume that the regularity conditions in Koo et al. (2008) are satisfied for k. Then for any $\boldsymbol{u}\in\mathbb{R}^{(d+1)(K-1)}$,*

$$\left|\mathbb{P}\left[n^{1/2}\left\{(\hat{\boldsymbol{\omega}}_{k\cdot},\hat{b}_k)^T - (\boldsymbol{\omega}_{k\cdot}^0, b_k^0)^T\right\} \leq \boldsymbol{u}\right]\right.$$
$$\left. - \mathbb{P}\left[n^{1/2}\left\{(\tilde{\boldsymbol{\omega}}_{k\cdot},\tilde{b}_k)^T - (\boldsymbol{\omega}_{k\cdot}^0, b_k^0)^T\right\} \leq \boldsymbol{u}\right]\right| \to 0,$$

so that the constrained solution has the same limiting distribution as the classical unconstrained solution.

Based on Theorem 1, inference for the constrained NORDIC can be obtained by applying the known asymptotic results for binary linear SVM, through the unconstrained NORDIC solutions. For example, we can show the asymptotic normality of the coefficients to the SVM components in linear NORDIC in the same way as those in Koo et al. (2008).

6. NUMERICAL RESULTS

We compare NORDIC-0, NORDIC-1, NORDIC-2, the vanilla ordinal classification method that uses $(K-1)$ separately trained (Frank and Hall, 2001) using binary SVM classifiers (BSVM), the data replication method by Cardoso and da Costa (2007) (DR) and the parallel discriminant hyperplane method by Chu and Keerthi (2005) (CK). We use our own experimental codes in the R environment to implement these methods. The Gaussian radial basis function kernel is used for all classifiers. The kernel parameter is tuned among the 10%, 50% and 90% quantiles of the pairwise distances between training vectors. The tuning parameters are tuned from a grid of possible values ranging from $2^{-4}, 2^{-3},\dots,2^4$.
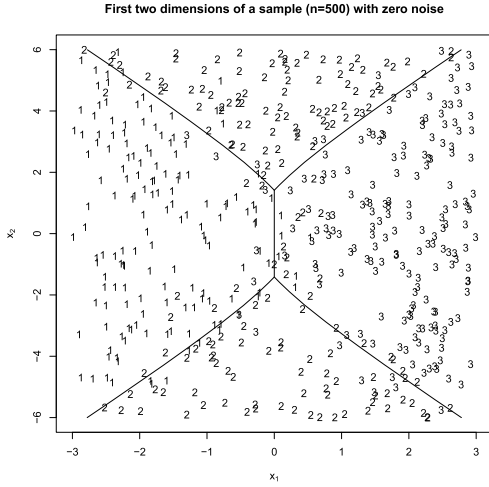
Figure 3. Nonlinear three-class examples: A scatter plot showing the first two dimensions of a realization with no additional error added. The Bayes rule is also shown.

6.1 Nonlinear three-class examples

We consider a data setting with three classes and d variables: $X_1, X_2, \ldots, X_d$, where

- $X_1 = \tilde{X}_1 + \sigma N(0,1)$ and $\tilde{X}_1 \sim \text{Uniform}(-3,3)$,
- $X_2 = \tilde{X}_2 + \sigma N(0,1)$ and $\tilde{X}_2 \sim \text{Uniform}(-6,6)$,
- and $X_3, \ldots, X_d \sim N(0,1)$.

Here, $\tilde{X}_1$ and $\tilde{X}_2$ truly determine the class labels (see below) but only their contaminated counterparts X_1 and X_2 are observed. In particular, let

$$f_1 = -2\tilde{X}_1 + 0.2\tilde{X}_1^2 - 0.1\tilde{X}_2^2 + 0.2,$$
$$f_2 = -0.4\tilde{X}_1^2 + 0.2\tilde{X}_2^2 - 0.4,$$
$$f_3 = 2\tilde{X}_1 + 0.2\tilde{X}_1^2 - 0.1\tilde{X}_2^2 + 0.2.$$

We assign each observation to class k with probability proportional to $\exp(f_k)$ for $k = 1, 2, 3$. We generate 100 data points in the training set, 100 in the tuning set and 10,000 in the test set. The standard deviation of the measurement error, σ, ranges from 0.5, 1 to 1.5. When $d = 5$ and $\sigma = 0$ (no perturbation), this is the same example as the nonlinear example in Zhang et al. (2008). However, we perturb the data and increase the dimension $(d = 10, 20, \ldots, 50)$ to make the problem more challenging.

Note that this example was initially designed by Zhang et al. (2008) as a regular multicategory classification, instead of an ordinal classification one. Figure 3 shows a sample realization of the data without perturbation at the first two dimensions. In a general sense, Class 2 can be viewed as in the middle of Class 1 and Class 3. We pretend that the class labels are of an ordinal nature and compare different ordinal classification methods.
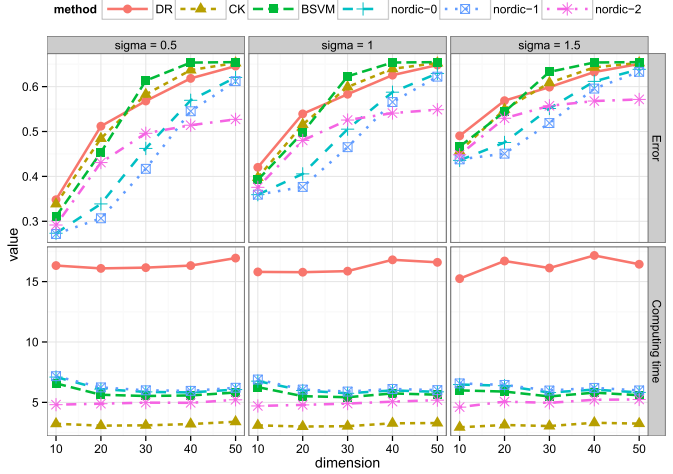
Figure 4. Nonlinear three-class examples: The top row shows the error rate for different methods (in different line types) with 3 noise levels (the left, middle and right panels) and 5 different dimensions (shown on the horizontal axis of each subfigure). The bottom row shows the computational time. In general, a NORDIC method is better than a non-NORDIC method for this example.

Figure 4 summarizes the results over 100 simulations. The NORDIC-0 and NORDIC-1 are the better classifiers in terms of classification performance when the dimensions are small. For higher dimensions, the NORDIC-2 method is better than the other methods. The DR method is the most computational costly and the CK method is the most efficient one. The reason that NORDIC works here is probably due to the perturbation that is added to this data set. A NORDIC method, with the help of the noncrossing constraints, can borrow strength from different classes and become more robust to perturbation.

6.2 Donut examples

We now consider a more challenging setting, which is tailered toward the ordinal data. We first generate data points from a 2-dimensional plate with radius 4 uniformly, and label them as from Class 1, except for those which are within a circle centered at $(1.9, 0)^T$ with radius 2, which are labeled as Class 2, and those which are within a circle centered at $(\sqrt{3} + 0.1, 0)^T$ with radius $\sqrt{3}$, which are labeled as Class 3. The observations for the additional $(d - 2)$ dimensions are all 0. We then perturb all the data points by adding independent d-dimensional Gaussian distributed random vector from $N_d(\mathbf{0}, \sigma\mathbf{I})$. We let $\sigma = 0.2$, 0.4 and 0.6 and let d range from 10 to 75. Figure 5 shows one realization of the data on the first two dimensions without the perturbation and the natural boundaries between the classes. This generalizes the classic donut examples in nonlinear classification.

Note that Class 2 is sandwiched by Class 1 and Class 3 from both outside and inside, and the high density region

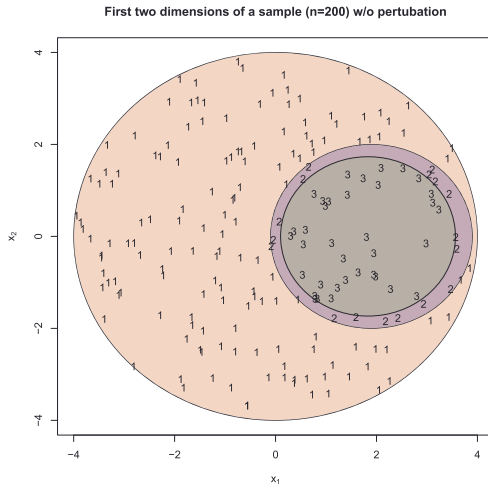First two dimensions of a sample (n=200) w/o pertubation

Figure 5. Donut examples: A scatter plot showing the first two dimensions of a realization, with no perturbation added. The natural boundaries between classes are also shown.
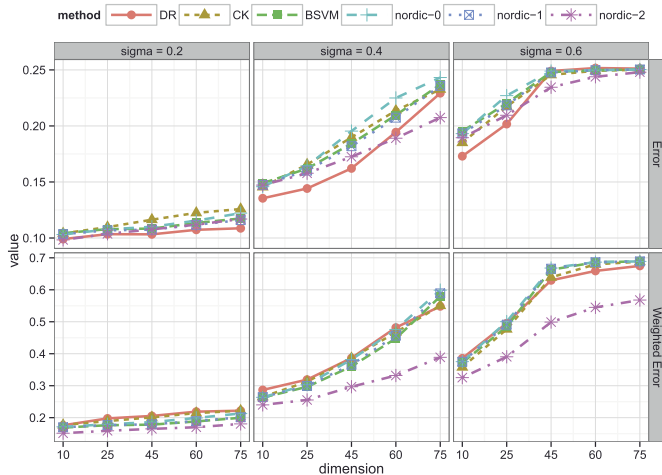

Figure 6. HD donut examples: The top row shows the error rate for different methods (in different line types) over 12 experiments with 3 noise levels (the left, middle and right panels) and 5 different dimensions (shown on the horizontal axis of each subfigure). The bottom row shows the weighted error rate. The NORDIC-2 is the best classifiers in terms of classification performance.

for Class 2 is very thin due to the construction. Hence, it is perceivable that a Class 2 observation is very difficult to be correctly classified. The noncrossing constraints here may be of some help because the boundary between Classes 1 and 2 may boost the estimation of the boundary between Classes 2 and 3, and vice versa.

The simulation results are reported in Figure 6. The first row shows the test error over 100 simulations. It appears that many times the DR method is the best. However, recall that in this data set the three classes are highly imbalanced

in terms of their sample size. On average, there are only 6.25% Class 2 points and 18.75% Class 3 points. A more reasonable measure to look into here is some weighted error rate that incorporates the different costs of misclassification. Here we report the weighted error with the configuration that:

- each misclassified point from Class 1 costs 1;
- each misclassified point from Class 2 to either Class 1 or Class 3 costs 2;
- each misclassified point from Class 3 to Class 2 costs 1, and from Class 3 to Class 1 costs 3.

Such assignment of the cost reflects the protection for Class 2, and the additional penalization for misclassifying across two boundaries (the cost for misclassifying from Class 3 to Class 1 is the sum of the costs for misclassifying from 3 to 2 and from 2 to 1.)

The second row of Figure 6 reports the weighted error rate. It is obvious that except for NORDIC-2, which is the best in this case, all other methods are more or less the same in terms of the weighted error. Interestingly, the NORDIC-0 and NORDIC-1 methods do not perform as well as their sibling NORDIC-2. They perform comparably to the other methods (they may have a very small advantage over CK and DR methods when the perturbation is small, for example, when $\sigma = 0.2$ and 0.4.) Recall that NORDIC-0 and NORDIC-1 imposes stronger constraints which aim for the monotonicity of the discriminant function $f_k(\boldsymbol{x})$ itself, as opposed to its sign. In contrast, the constraint from NORDIC-2 is much lighter, which may have left enough "degrees of freedom" to optimize the generalization performance.

One may argue that the choice of the costs in the weighted error may be arbitrary. In this case, it may be helpful to look into the confusion matrix to see the cause of the different performance. Figure 7 depicts the 3×3 confusion matrices for the case with contamination $\sigma = 0.4$ for different methods and different dimensions. For the (k, ℓ)th plot in the array, the reported value is the proportion of observations from Class ℓ that are classified to Class k ($\ell, k = 1, 2, 3$). Note that the aggregation of the three plots in the same column equals to 1. A good classifier is expected to have high rates in the diagonal plots and low rates in the off-diagonal plots. There are, on average, 7496.2 observations from Class 1, and almost all the methods classify them correctly. Class 2 (with only 626.5 observations) is clearly a very difficult class. Even our NORDIC-2 has a poor classification accuracy of 25%. That said, NORDIC-2 shows more advantages for higher dimensional cases. For Class 3, NORDIC-2 shows improved accuracy, especially with much fewer misclassifications into Class 1 (shown in the upper-left plot).

The computational time results are similar to what we have seen for the last example and are not reported here.

7. REAL APPLICATION

We use the scale balance data set from the UCI Machine Learning Repository (Lichman, 2013) to test the usefulness
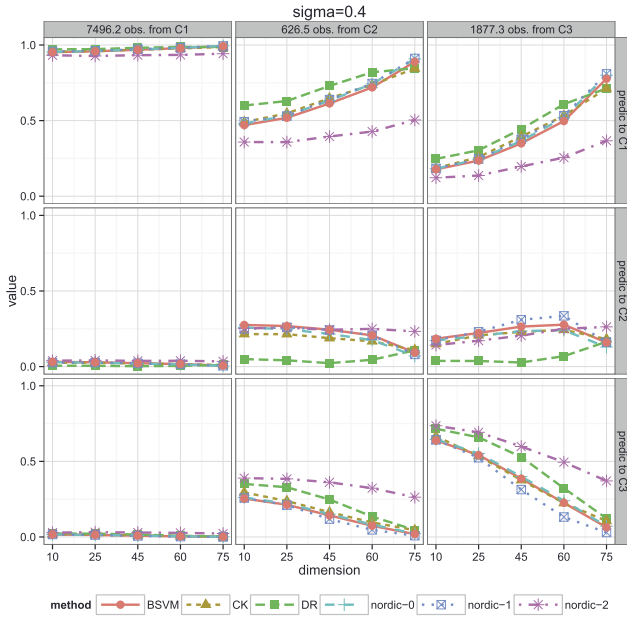
Figure 7. Confusion matrices for the examples with $\sigma = 0.4$ for different methods and dimensions.

Figure 8. Weighted error rate for the scale balance data set.

Figure 9. Confusion matrices for the scale balance data set.

of the NORDIC method. This data set, studied in Siegler (1976), was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The four attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced. There are 625 instances in the data, with 288 tip to the left (L), 49 balanced (B), and 288 tip to the right (R).

There is a clear order between the three classes (L, B and R,) and hence ordinal classification methods are appropriate. We randomly select n points from the data set for training, n for tuning, and the remaining $(625 - 2n)$ are for testing. The proportions of the three classes are preserved when the partitioning is conducted. The random experiment is repeated for 100 times. We consider four cases, where $n = 52$, 79, 125 and 208 respectively.

A naive coding of 1, 2 and 3 for these three classes followed by a regression method will prove to be suboptimal. In particular, in addition to the ordinal classification methods, we also compare with support vector regression (SVR Smola and Schölkopf, 2004, implemented by `svm()` in the R package `e1071`) with Gaussian radial basis function kernel. SVR is applied to the data with $\{1,2,3\}$ coding, and the predicted class is obtained by cut-off values 1.5 and 2.5 for the predicted numerical outcome.

Figure 8 shows the weighted error rates of different methods over 100 random splitting of the data set and 4 different sample sizes. Here, we let a misclassified point from Class 3 to Class 1, or from Class 1 to Class 3, to bear a cost
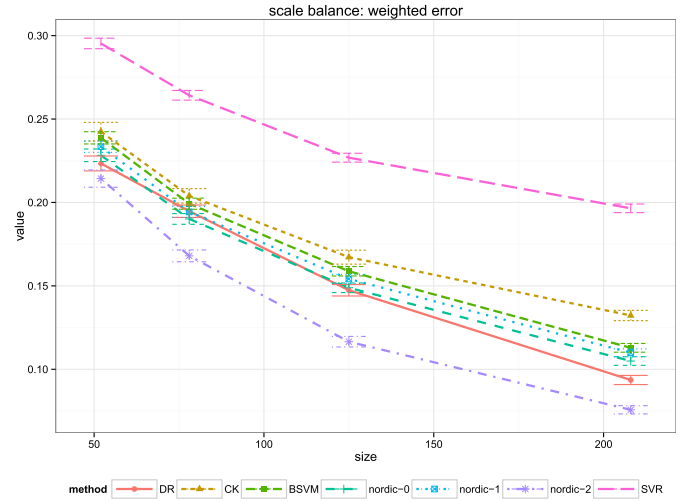
of 2; other types of misclassification cost only 1. All three NORDIC methods are among the best, with NORDIC-2 having a significant advantage. The other two NORDIC methods are comparable to the DR method especially for small sample cases. The SVR is the worst classifier in this experiment.

Figure 9 shows the confusion matrices. It can be seen that the poor performance of the SVR method is probably because it classifies much more instances to Class B, and this may be due to the arbitrary choice of the cut-off values 1.5 and 2.5. However, one may have no better way to choose the

cut-offs except through another layer of tuning parameter selection. On the other hand, NORDIC-2 stands out as the best classifier due to its best performance on Class B among the other methods (except for SVR.) Note that for Classes L and R, all methods (except for SVR) perform more or less the same.

8. CONCLUDING REMARKS

In this article, three versions of NORDIC classifiers are proposed to make use of the order information in classifying ordinal data. All three classifiers train $(K-1)$ binary SVM classifiers simultaneously with extra constrains to ensure noncrossing among classification boundaries. The NORDIC-0 and NORDIC-1 methods focus on a sufficient condition for noncrossing and are solved by QP. The NORDIC-2 method aims for the exact condition for noncrossing but has to be solved by the integer programming algorithm.

Let us turn our attention back to the formulation for NORDIC-0, (2)–(4). Without the additional constraints (3) and (4), the NORDIC-0 method is the combination of $(K-1)$ independently trained SVM classifiers (with the common tuning parameter). It is known that for a single binary SVM classifier, the discriminant function is given by $f(\boldsymbol{x}) = \sum_{j=1}^{n} \omega_j K(\boldsymbol{x}_j, \boldsymbol{x}) + b$. The coefficients $\omega_i = \alpha_i y_i$ is calculated by maximizing the following dual problem of SVM,

$$\mathcal{L}_{SVM} = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

(20)   subject to $\sum_i \alpha_i y_i = 0, \text{ and } 0 \leq \alpha_i \leq C.$

See, for example, Burges (1998) for a tutorial. The maximization problem above is the dual problem of SVM, while our NORDIC-0 method was based on the primal problem of SVM.

One may wonder if a dual-based NORDIC is possible. Indeed, a variant of NORDIC can be viewed as to maximize the sum of $(K-1)$ such objective functions as in (20), with extra noncrossing constraints that $\boldsymbol{\omega}_{k\cdot} \geq \boldsymbol{\omega}_{(k+1)\cdot}$, that is $\alpha_{k,i} y_i^{(k)} \geq \alpha_{k+1,i} y_i^{(k+1)}$. Note that the constraints are the same as in NORDIC-0 but the objective function is based on the dual objective function. However, one can show that this formulation ultimately reduces to the method proposed by Chu and Keerthi (2005), namely, all the $(K-1)$ classifiers share the same $\boldsymbol{\omega}$ vector. Hence, the CK method can be viewed as a special case in the NORDIC family. Note that in our NORDIC-0 proposal, we focus on the primal formulation. As a consequence, the resulting boundaries are not parallel to each other, leading to more flexibility.

The usefulness and efficiency of the proposed methods are supported by the comparison with the competitors. Promising results are obtained from simulated and data examples. Fisher consistency of the NORDIC method and asymptotic

normality of the linear NORDIC method further validate the proposed methods.

There is a natural connection between ordinal classification and ordered logistic regression. Both methods fully utilize the ordinal class information. Their difference can be viewed as analogous to the difference between binary SVM and (binary) logistic regression, or that between multicategory SVM and multinomial logistic regression. It is interesting to explore the benefit of using machine learning techniques including NORDIC, over the modeling approaches such as ordered logistic regression. See Lee and Wang (2015) for such a comparison in the binary case.

We have provided three distinct formulations. They may perform differently on different types of data sets, both in terms of the generalization error and the computational time; the derivation of these optimization problems may give insights into which kernels can more easily admit truly noncrossing boundaries. It is an interesting future research direction to identify specific kernels for which we can provide truly non-crossing boundaries.

APPENDIX

Proof to Theorem 1

Let $\hat{\boldsymbol{Z}}_n$ and $\tilde{\boldsymbol{Z}}_n$ denote $n^{1/2} \left\{ (\hat{\boldsymbol{\omega}}_{k\cdot}, \hat{b}_k)^T - (\boldsymbol{\omega}_{k\cdot}^0, b_k^0)^T \right\}$ and $n^{1/2} \left\{ (\tilde{\boldsymbol{\omega}}_{k\cdot}, \tilde{b}_k)^T - (\boldsymbol{\omega}_{k\cdot}^0, b_k^0)^T \right\}$, respectively. Then

$$\left| \mathbb{P} \left(\hat{\boldsymbol{Z}}_n \leq \boldsymbol{u} \right) - \mathbb{P} \left(\tilde{\boldsymbol{Z}}_n \leq \boldsymbol{u} \right) \right|$$
$$= \left| \mathbb{P} \left(\hat{\boldsymbol{Z}}_n \leq \boldsymbol{u} \mid \hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n \right) - \mathbb{P} \left(\tilde{\boldsymbol{Z}}_n \leq \boldsymbol{u} \mid \hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n \right) \right|$$
$$\times \mathbb{P} \left(\hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n \right)$$

Since the first term in the product is bounded by 2, it suffices to show that $\mathbb{P} \left(\hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n \right) \to 0$. The event, $\hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n$, is equivalent to the event that the unconstrained binary linear SVM classifiers have boundaries crossing from each other, that is,

$$n^{1/2} \left\{ \text{sign} \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{k\cdot} + \tilde{b}_k \right) - \text{sign} \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{(k+1)\cdot} + \tilde{b}_{k+1} \right) \right\} < 0$$

for some $\boldsymbol{x} \in \mathcal{S}$. This is only possible when $\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{k\cdot} + \tilde{b}_k < 0$ and $\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{(k+1)\cdot} + \tilde{b}_{k+1} > 0$. We consider their difference

$$n^{1/2} \left\{ \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{k\cdot} + \tilde{b}_k \right) - \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{(k+1)\cdot} + \tilde{b}_{k+1} \right) \right\}.$$

This difference can be written as

$$n^{1/2} \left\{ \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{k\cdot} + \tilde{b}_k \right) - \left(\boldsymbol{x}^T \boldsymbol{\omega}_{k\cdot}^0 + b_k^0 \right) \right\}$$
$$- n^{1/2} \left\{ \left(\boldsymbol{x}^T \tilde{\boldsymbol{\omega}}_{(k+1)\cdot} + \tilde{b}_k \right) - \left(\boldsymbol{x}^T \boldsymbol{\omega}_{(k+1)\cdot}^0 + b_{k+1}^0 \right) \right\}$$
$$+ n^{1/2} \left\{ \left(\boldsymbol{x}^T \boldsymbol{\omega}_{k\cdot}^0 + b_k^0 \right) - \left(\boldsymbol{x}^T \boldsymbol{\omega}_{(k+1)\cdot}^0 + b_{k+1}^0 \right) \right\}$$

Under the regularity conditions, and due to the results in Koo et al. (2008), the first two terms above are $O_p(1)$. Thus, $n^{1/2}\left\{\left(\boldsymbol{x}^T\boldsymbol{\omega}_{k\cdot}^0 + b_k^0\right) - \left(\boldsymbol{x}^T\boldsymbol{\omega}_{(k+1)\cdot}^0 + b_{k+1}^0\right)\right\} \leq -C < 0$. This contradicts the fact that $\text{sign}\left(\boldsymbol{x}^T\boldsymbol{\omega}_{k\cdot}^0 + b_k^0\right) \geq \text{sign}\left(\boldsymbol{x}^T\boldsymbol{\omega}_{(k+1)\cdot}^0 + b_{k+1}^0\right)$ due to the assumption that the conditional density for each class is positive. Thus $\mathbb{P}\left(\hat{\boldsymbol{Z}}_n \neq \tilde{\boldsymbol{Z}}_n\right) \to 0$ which completes the proof.

REFERENCES

BONDELL, H. D., REICH, B. J., and WANG, H. (2010), "Noncrossing quantile regression curve estimation," *Biometrika*, 97, 825–838. MR2746154

BRADLEY, S., HAX, A., and MAGNANTI, T. (1977), *Applied mathematical programming*, Addison Wesley.

BURGES, C. J. C. (1998), "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 2, 121–167.

CARDOSO, J. and DA COSTA, J. (2007), "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, 8, 1393–1429. MR2332436

CHU, W. and GHAHRAMANI, Z. (2005), "Gaussian processes for ordinal regression," *Journal of Machine Learning Research*, 6, 1019–1041. MR2249847

CHU, W. and KEERTHI, S. (2005), "New approaches to support vector ordinal regression," in *Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA: ACM, ICML '05, pp. 145–152.

CORTES, C. and VAPNIK, V. (1995), "Support-vector networks," *Machine Learning*, 20, 273–297.

CRAMMER, K. and SINGER, Y. (2002), "On the learnability and design of output codes for multiclass problems," *Machine Learning*, 47, 201–233, 10.1023/A:1013637720281.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000), *An introduction to support vector machines: and other kernel-based learning methods*, Cambridge University Press.

DUDA, R., HART, P., and STORK, D. (2001), *Pattern classification*, Wiley. MR1802993

FRANK, E. and HALL, M. (2001), "A simple approach to ordinal classification," in *Machine Learning: ECML 2001*, eds. De Raedt, L. and Flach, P., Springer, Berlin Heidelberg, vol. 2167 of *Lecture Notes in Computer Science*, pp. 145–156.

HERBRICH, R., GRAEPEL, T., and OBERMAYER, K. (2000), "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, eds. Bartlett, P. J., Schölkopf, B., Schuurmans, D., and Smola, A. J., the MIT Press, pp. 115–132. MR1820960

HUANG, H., LIU, Y., DU, Y., PEROU, C. M., HAYES, D. N., TODD, M. J., and MARRON, J. (2013), "Multiclass distance weighted discrimination," *Journal of Computational and Graphical Statistics*, 22, 953–969. MR3173751

KIMELDORF, G. and WAHBA, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95. MR0290013

KOO, J., LEE, Y., KIM, Y., and PARK, C. (2008), "A Bahadur representation of the linear support vector machine," *Journal of Machine Learning Research*, 9, 1343–1368. MR2426045

LEE, Y., LIN, Y., and WAHBA, G. (2004), "Multicategory support vector machines," *Journal of the American Statistical Association*, 99, 67–81. MR2054287

LEE, Y. and WANG, R. (2015), "Does modeling lead to more accurate classification?: A study of relative efficiency," *Journal of Multivariate Analysis*, 133, 232–250. MR3282028

LICHMAN, M. (2013), "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml.

LIN, Y. (2004), "A note on margin-based loss functions in classification," *Statistics & Probability Letters*, 68, 73–82. MR2064687

LIU, Y. (2007), "Fisher consistency of multicategory support vector machines," in *International Conference on Artificial Intelligence and Statistics*, pp. 291–298.

LIU, Y. and WU, Y. (2006), "Optimizing psi-learning via mixed integer programming," *Statistica Sinica*, 16, 441–457. MR2267244

LIU, Y. and WU, Y. (2011), "Simultaneous multiple non-crossing quantile regression estimation using kernel constraints," *Journal of Nonparametric Statistics*, 23, 415–437. MR2801302

LIU, Y., ZHANG, H., and WU, Y. (2011), "Hard or soft classification? Large-margin unified machines," *Journal of the American Statistical Association*, 106, 166–177. MR2816711

MARRON, J., TODD, M., and AHN, J. (2007), "Distance-weighted discrimination," *Journal of the American Statistical Association*, 102, 1267–1271. MR2412548

PLATT, J. C. (1999), "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*, eds. Schölkopf, B., Burges, C. J. C., and Smola, A. J., Cambridge, MA, USA: MIT Press, pp. 185–208.

QIAO, X. (2015), "Learning ordinal data," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 341–346. MR3385031

QIAO, X., ZHANG, H., LIU, Y., TODD, M., and MARRON, J. (2010), "Weighted distance weighted discrimination and its asymptotic properties," *Journal of the American Statistical Association*, 105, 401–414. MR2656058

QIAO, X. and ZHANG, L. (2015a), "Distance-weighted support vector machine," *Statistics and Its Interface*, 8, 331–345. MR3341331

QIAO, X. and ZHANG, L. (2015b), "Flexible high-dimensional classification machines and their asymptotic properties," *Journal of Machine Learning Research*, 16, 1547–1572.

SHASHUA, A. and LEVIN, A. (2003), "Ranking with large margin principle: two approaches," in *Advances in Neural Information Processing Systems 15*, eds. Becker, S., Thrun, S., and Obermayer, K., Cambridge, MA: the MIT Press, pp. 937–944.

SHEN, X., TSENG, G. C., ZHANG, X., and WONG, W. H. (2003), "On ψ-learning," *Journal of the American Statistical Association*, 98, 724–734. MR2011686

SIEGLER, R. S. (1976), "Three aspects of cognitive development," *Cognitive Psychology*, 8, 481–520.

SMOLA, A. J. and SCHÖLKOPF, B. (2004), "A tutorial on support vector regression," *Statistics and Computing*, 14, 199–222. MR2086398

VAPNIK, V. (1998), *Statistical learning theory*, Wiley. MR1641250

WESTON, J. and WATKINS, C. (1999), "Support vector machines for multi-class pattern recognition," in *European Symposium on Artificial Neural Networks*, pp. 219–224.

ZHANG, H., LIU, Y., WU, Y., and ZHU, J. (2008), "Variable selection for the multicategory SVM via adaptive sup-norm regularization," *Electronic Journal of Statistics*, 2, 149–167. MR2386091

Xingye Qiao
Department of Mathematical Sciences
Binghamton University
State University of New York
Binghamton, NY 13902-6000
USA
E-mail address: qiao@math.binghamton.edu