# NEW STATISTICAL LEARNING METHODS FOR PERSONALIZED MEDICAL DECISION MAKING

Xuan Zhou

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2018

Approved by:

Donglin Zeng

Yuanjia Wang

Chunqin Deng

Gary G. Koch

Yufeng Liu

# ABSTRACT

Xuan Zhou: New Statistical Learning Methods for Personalized Medical
Decision Making
(Under the direction of Donglin Zeng and Yuanjia Wang)

This research focuses on developing new and computationally efficient statistical learning methods for multicategory classification and personalized medical decision making. Motivated by the challenge of multicategory classification problems, and the computational efficiency and theoretical properties of support vector machines (SVM), a novel learning algorithm is proposed. The method is then adapted to estimating multicategory individualized treatment rules by connecting with outcome weighted learning (Zhao et al., 2012). At last, an application to Electronic Health Record data is explored.

The proposed algorithm, forward-backward SVM (FB-SVM) is based on a sequential binary classification algorithm and relies on support vector machines for each binary classification and utilizes only feasible data in each step. Therefore, the method guarantees convergence and entails light computational burden. More importantly, we prove the theoretical property of Fisher consistency of the classification rule derived from the FB-SVM, which is not guaranteed by existing algorithms. We also obtain the risk bound for the predicted misclassification rate. We conduct extensive simulation and application studies, using popular benchmarking data and data from a newly completed real-world study, to demonstrate that the proposed method has superior performance, in terms of low misclassification rates and significantly improved computational speed when compared to existing methods.

Furthermore, we generalize the proposed FB-SVM with outcome weighted learning to estimate optimal individualized treatment rule (ITR) with multiple options of treatment, namely sequential outcome-weighted learning (SOM). Specifically, we solve a multicategory treatment selection problem via sequential weighted support vector machines. Theoretically, we show that the resulting ITR is Fisher consistent. We demonstrate the performance of proposed method with extensive simulations.

An application to a three-arm randomized trial of treating major depressive disorder shows that an individualized treatment strategy tailored to individual characteristics such as patients' expectancy of treatment efficacy and baseline depression severity reduces depressive symptoms more than non-personalized treatment strategies.

Finally, we discuss how the proposed SOM learning can be used to estimate optimal ITRs with safety concerns in high dimensional electronic health record data, which are collected from the Indiana Network for Patient Care database with patients' adverse reaction records who have taken statin medicine. We adopt sampling techniques, inverse probability weighting, propensity score adjustment, and variable clustering along with SOM learning in our analysis. Considering patients' electronic records of demographics and medical history, we are able to recommend the best statin drug which has the lowest risk to cause myopathy or rhabdomyolysis using SOM learning.

## ACKNOWLEDGEMENTS

First and foremost, I want to express my deepest gratitude to my advisors, Dr. Donglin Zeng and Dr. Yuanjia Wang, for their generous support, instructions and deep insights. It has been a great pleasure working with them the past few years. Their passion and profession in biostatistics are the leading sources in my growth throughout my dissertation research. I particularly appreciate the generous financial support they provided throughout the years of my graduate study. Further, I would like to thank the members of my committee, Drs. Chunqin Deng, Gary G. Koch, and Yufeng Liu, for their time and their comments which improved the quality of this work.

I would like to sincerely thank Dr. Koch for encouraging me to go on to PhD study after my graduation from master's program. He also provided generous financial support from Biometric Consulting Lab, and offered a fellowship at United Therapeutics. Dr. Koch is always available to discuss both my research and non-research issues. I also appreciate his time and insightful suggestions for my third project.

Many thanks to Dr. Haibo Zhou for being my academic advisor during the years of my master's study, and Dr. Jianwen Cai for introducing me to Dr. Zeng's project at Carolina Survey Research Laboratory. I'm also thankful to all the faculty and staff members for their various guidance and assistance throughout the past five years I spent in the department.

Next, sincere thanks to my family and friends for their support, encouragement, and company in the up and down days of my graduate life.

Finally, I would like to thank my husband, Chang, for his constant love, patience and support throughout the dissertation process.

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

Multicategory classification is a fundamental and challenging problem in machine learning and empirical application. While binary classification methods are well-developed, multicategory classification is a heavily debated research topic in the machine learning literature. In addition, a multicategory classification learning method can be well applied to estimate individualized treatment rules in trials with more than two treatment arms, which is also not sufficiently developed in existing literatures. Motivated by the need to create a computationally efficient method of consistent learning rules, we develop statistical learning methods for multicategory classification and estimating optimal personalized treatment rules.

In the first topic (see Chapter 2), we propose a new and computationally efficient learning algorithm, namely, forward-backward support vector machine (FB-SVM), to perform multicategory learning. The new method is based on a sequential binary classification algorithm: we first classify a target class by excluding the possibility of labeling as any other classes using a forward step of sequential SVM; we then exclude already classified classes and repeat the same procedure for the remaining classes in a backward step. The proposed algorithm relies on support vector machines for each binary classification and utilizes only feasible data in each step; therefore, the method guarantees convergence and entails light computational burden. More importantly, we prove the theoretical property of Fisher consistency of the classification rule derived from the FB-SVM, which is not guaranteed by existing algorithms. Furthermore, we obtain the risk bound for the predicted misclassification rate. We conduct extensive simulation and application studies, using popular benchmarking data and data from a newly completed real-world study, to demonstrate that the proposed method has superior performance, in terms of low misclassification rates and significantly improved computational speed when compared to existing methods.

Personalized medicine has received increasing interest among clinicians and statisticians. Particularly, one aspect of personalized medicine is to consider an individualized treatment strategy based on an individual's characteristics that leads to the greatest benefit in the patient population.

Recently, powerful machine learning methods have been proposed to estimate optimal individualized treatment rule (ITR) but are only restricted to the situation with only two treatments. When multiple treatment options are being considered, which is often the case in many studies, these methods have to handle complex treatment-treatment interactions by transforming multicategory treatment selection into multiple binary treatment selections. However, combining conclusions from multiple binary treatment selection is not straightforward and it may lead to inconsistent decision rules. There is a lack of literature on using multicategory learning to estimate optimal ITR. In the second topic (see Chapter 3), we fill this gap by proposing a novel and efficient method to generalize outcome weighted learning to multi-treatment settings. Theoretically, we show that the resulting ITR is Fisher consistent. We demonstrate the performance of the proposed method with extensive simulations. An application to a three-arm randomized trial of treating major depressive disorder shows that an individualized treatment strategy tailored to individual characteristics such as patients' expectancy of treatment efficacy and baseline depression severity reduces depressive symptoms more than non-personalized treatment strategies.

In the third topic (see Chapter 4), we discuss how the proposed learning algorithm can be used to estimate optimal ITRs with safety concerns in high dimensional data, which are collected from the Indiana Network for Patient Care database with patients' adverse reaction records who have taken statin medicine. In this work of observational data. we adopt sampling techniques, inverse probability weighting, propensity score adjustment, and variable clustering along with SOM learning to take care of potential bias and confounding, large computational cost, and sparsity. Tailored to patients' demographics and medical history, SOM learning is able to recommend the best statin drug which has the lowest risk to cause myopathy or rhabdomyolysis. Results of Q-learning and non-individualized treatment rules are also explored.

Chapters 2 to 4 each contains an introduction and relevant literature review, an explication of the problem, the proposed methods followed by theoretical properties, numerical examples and conclusions. Chapter 5 represents ideas for future research.

# CHAPTER 2: MULTICATEGORY CLASSIFICATION VIA FORWARD-BACKWARD SVM

## 2.1 Introduction

Multicategory classification is a fundamental and challenging problem in machine learning and empirical application. For instance, cancer is staged into more than two levels of categorization. An accurate classification of a patient's cancer stage is crucial for prognosis and the choice of an effective treatment. While binary classification methods are well-developed, multicategory classification is a heavily debated research topic in the machine learning literature.

The existing methods for multicategory classification can be divided into two general approaches. The first approach is to convert multicategory classification problems into sequential binary problems such as one-versus-all and one-versus-one methods (Dietterich and Bakiri, 1995; Kreßel, 1999; Allwein et al., 2001). These methods are relatively simple to implement because all work is accomplished using existing off-the-shelf techniques for binary classification problems, although the one-versus-one approach requires significantly more computational time than the one-versus-all approach (Hsu and Lin, 2002; Rifkin and Klautau, 2004). However, although easy to implement, these methods suffer from the risk of inconsistent results because an input sample may be assigned to multiple classes depending on the choice of binary classification. Thus, these methods do not satisfy the Fisher consistency property, and their misclassification error rates cannot achieve the optimal Bayesian error rate. Additionally, with the one-versus-all approach, the training set is imbalanced because there are many more subjects in the remaining classes than in the targeted class (Bishop, 2006).

The second approach to learning algorithms for multicategory classification is to form a single loss function for minimization so as to obtain a simultaneous multicategory objective function (Vapnik and Vapnik, 1998; Weston et al., 1999; Bredensteiner and Bennett, 1999; Crammer and Singer, 2002; Lee et al., 2004; Liu and Shen, 2006; Liu and Yuan, 2011). The key question is how to choose a sensible multicategory loss function. To better understand each loss function, we introduce a few notations. Let $\boldsymbol{f} = (f_1, f_2, ..., f_k)$ be a vector of decision functions, where each component represents one of $k$ classes. Suppose an input $\boldsymbol{x}$ belongs to class $y$, then the loss function used in Vapnik and Vapnik (1998), Weston et al. (1999) and Bredensteiner and Bennett (1999) is $\sum_{j \neq y}[1-(f_y(x)-f_j(x))]_+$, which pays a certain penalty when $f_y(x)-f_j(x)<2$. The authors in Lee et al. (2004) and Liu (2007) have shown that the loss may be Fisher inconsistent by providing a special case for 3-category classification problems. When all probabilities of being assigned to one of the three classes are less than 1/2, the loss is inconsistent if the second largest probability is greater or equal to 1/3. The loss function in Crammer and Singer (2002) and Liu and Shen (2006) has a similar formulation, $[1-\min_j(f_y(x)-f_j(x))]_+$. The difference is that this loss function pays a penalty only for the smallest difference between the targeted class and the other classes. This loss function is not convex and is Fisher consistent only when there is a dominating class, which is not guaranteed for multicategory classification (Zhang, 2004; Tewari and Bartlett, 2007; Liu, 2007). In Crammer and Singer (2002), a dual decomposition algorithm for the Lagrangian dual with kernel product was developed to solve the quadratic programming problem. Liu and Shen (2006) applied the difference convex decomposition to address the problem of non-convex minimization and provided the convergence rate. The loss function proposed by Lee et al. (2004) is $\sum_{j \neq y}[1+f_j(x)]_+$, which is a consistent generalization of hinge loss. The authors derived a Lagrangian dual but did not provide the decomposition algorithm. Despite the non-convexity of the loss function, another drawback of Lee, Lin, and Wahba's (Lee et al., 2004) is that the tuning criterion based on generalized cross-validation tends to over-estimate the tuning parameter, which may result in inaccurate classification. Liu and Yuan (2011) proposed reinforced hinge loss function $\gamma[(k-1)-f_y(x)]+(1-\gamma)\sum_{j \neq y}[1+f_j(x)]_+$, which is a convex combination of Lee, Lin, and Wahba's loss function and a direct generalized hinge loss. The authors stated that the loss function is consistent when $\gamma \leq 1/2$ or under certain circumstances when $\gamma > 1/2$. The algorithm contains both quadratic and linear programming problems with respect to each part of the loss function.

In this paper, we propose a novel algorithm – *forward-backward support vector machine* (FB-SVM) – to break a multicategory classification problem into sequential binary classifications in a way that takes advantage of computational efficiency and also allows consistent theoretical properties. For each binary step in our algorithm, we use a weighted support vector machine (SVM) for classification because the theoretical properties of binary SVMs are well established and flexible to allow for the incorporation of powerful kernel learning (Cristianini and Shawe-Taylor, 2000; Friedman et al., 2001;

Schölkopf and Smola, 2002; Lin, 2004). We first classify a target class by excluding the possibility of labeling as any other classes using a forward step of sequential SVMs; we then exclude already classified classes and repeat the same learning approach for the remaining classes using a backward step. By carefully choosing weights in each SVM step and constructing the final classification rules to combine the decision functions from all SVM steps, we show that the derived rule from FB-SVM is Fisher consistent, and we obtain the risk bound for the predicted misclassification rate. Moreover, we demonstrate through extensive simulation and application studies that FB-SVM has superior performance in terms of low misclassification rates and significantly improved computational speed, in comparison to existing methods.

The paper is structured as follows. Section 2 introduces the concept of FB-SVM using a 3-category classification problem, with the general algorithm for more than 3 categories provided at the end of the section. In Section 3, we provide the theoretical properties of FB-SVM, including Fisher consistency and the convergence rate of the predicted misclassification rate. Sections 4 and 5 present simulations and applications for illustration. Finally, a discussion and concluding remarks are provided in Section 6. Proofs of the theoretical results are presented in the Appendix.

## 2.2 Methodology

### 2.2.1 An illustration of FB-SVM for 3-category learning

We use a 3-category problem to introduce FB-SVM. Suppose we have a training dataset with $n$ pairs of predictors and outcomes $\{\boldsymbol{x}_i, y_i\}$, $i = 1, ..., n$, where $\boldsymbol{x}_i$ is the $d$-dimensional input, and $y_i$ is the corresponding class label for each sample, taking value 1, 2, or 3.

First, use class 3 as the target and obtain a decision function for this class. The idea is to produce some rules in order to eliminate any possibility that a sample will be classified with label 1 or 2. For this purpose, we first obtain a rule to classify label 1 versus $\{2, 3\}$. Because this is a classification of one versus two classes, subjects with label 1 must be weighted twice as much as subjects labeled 2 or 3. After we obtain such a rule, we next consider the subjects with labels 2 or 3 who are not classified with label 1. Using these data, we then perform another classification to discriminate between classes 2 and 3. Hence, any subject who is not classified with label 1 or label 2 will be classified with label 3. The choice of beginning with label 1 is arbitrary, and a mislabeling can potentially affect the second classification between classes 2 and 3. To remedy this, we can start from label 2 by first discriminating between classes 2 and $\{1, 3\}$ and then discriminating between classes $\{1, 3\}$ using another sequential SVM. Combining the above, we classify a subject with label 3 if either of the two sequential classification rules classifies this subject with label 3. Because this ordered classification is based on sequential SVMs, where each step depends on the rules obtained in the previous steps, we term this procedure a *forward SVM*.

Next, we aim to improve the rule for classifying a subject with labels 2 or 1. This can be carried out by eliminating subjects with label 3 and those who are already classified as label 3. Thus, the remaining data consist only of subjects with labels 1 or 2 who cannot be classified with label 3. We then perform a simple SVM to discriminate between classes 1 and 2 using the reduced data. Because of this backward elimination step, we term this procedure a *backward SVM*.

In summary, we collectively obtain a classification rule via a forward weighted SVM to construct a decision rule for class 3. We then eliminate infeasible data using a backward step and further construct a decision rule to distinguish the remaining two classes using a restricted sample. Generalization to any number of multiple categories is provided in the next section.

### 2.2.2 FB-SVM for general multicategory learning

Now, we extend the 3-category problem to the $k$-category, where $k \geq 3$. The training dataset consists of $n$ pairs of predictors and outcomes $\{\boldsymbol{x}_i, y_i\}$, $i = 1, ..., n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is feature variables for subject $i$, and $y_i \in \{1, 2, ..., k\}$ is the class label. As described previously, the main idea of FB-SVM is to perform a sequence of binary SVMs in order to exclude any possibility of classification into the class other than a pre-specified target class (forward SVM) and then perform backward elimination to exclude already determined classes and identify decision rules for the remaining classes (backward SVM). We describe the details of the method for the $k$-category below.

We first order the classes based on the descending order of prevalences. Specifically, we first learn the most prevalent class, then the second prevalent class, and then the rest. Without loss of generality, we assume that the order of the labels is $k, k-1, k-2, ..., 1$. To learn a classification rule for class label $k$, we consider all permutations of labels $\{1, 2, ..., k-1\}$. For the $j$th permutation, say $\{j_1, ..., j_{k-1}\}$, we conduct a sequence of $(k-1)$ binary classifications. In the first binary classification, we learn a decision rule by comparing label $j_1$ versus the remaining labels, i.e., $j_2, j_3, ..., j_{k-1}, j_k = k$. In the second binary classification, we make use of samples that are not in class $j_1$ and not classified with label $j_1$ to learn a binary classification decision rule by comparing label $j_2$ versus the remaining labels, i.e., $j_3, ...., j_{k-1}, j_k = k$. We then continue this binary classification procedure so that at the $l$th step, we use samples that are not in the previously considered classes and not classified with any of the previous labels to learn a binary decision rule by comparing label $j_l$ versus $j_{l+1}, ..., j_k = k$. In the final step, i.e., the $(k-1)$th step, the binary classification compares label $j_{k-1}$ vs $j_k = k$.

Because we are comparing one class versus the remaining classes at each step, the binary classification should be weighed to balance the classes in order to account for the fact that one class is compared to multiple classes. Specifically, at the $l$th step, we weight class $j_l$ by $(k-l)/(k-l+1)$ and weight the combined remaining classes by $1/(k-l+1)$. We will show that this weighting scheme ensures Fisher consistency of the proposed method.

Mathematically, we can express the above forward SVM algorithm as follows. Consider the $j$th permutation $\{j_1, ..., j_{k-1}\}$, and let $j_k = k$.

1. At step 1, define $z_{ij_1} = 1 - 2I\{y_i \neq j_1\}$, and estimate a decision rule $\text{sign}(\widehat{f}_{j1}(\boldsymbol{x}))$ using a weighted SVM so that $\widehat{f}_{j1}(\boldsymbol{x})$ minimizes the following empirical risk of a weighted hinge loss

$$\begin{aligned} V_{nj_1}(f) &= n^{-1}\sum_{i=1}^{n}\left\{\frac{k-1}{k}I(z_{ij_1}=1)[1-f(\boldsymbol{x}_i)]_+ + \frac{1}{k}I(z_{ij_1}=-1)[1+f(\boldsymbol{x}_i)]_+\right\} \\ &\quad +\lambda_{nj_1}\|f\|^2, \end{aligned}$$

where $x_+ = \max(x, 0)$ is the hinge loss, $\|\cdot\|$ denotes a semi-norm for $f$, and $\lambda_{nj_1}$ is a tuning parameter. In particular, if we consider a linear decision rule, i.e., $f(\boldsymbol{x}) = \beta^T\boldsymbol{x} + \beta_0$, $\|f\|$ is chosen as the Euclidean norm of $\beta$; if a nonlinear decision rule is desired, $f$ is chosen from a reproduced kernel Hilbert space, and $\|f\|$ is the corresponding norm in that space.

2. At step 2, we restrict the training data to those samples whose labels are not $j_1$ and are not classified with $j_1$ from the previous step. Thus, we estimate a decision rule $\text{sign}(\widehat{f}_{j_2}(\boldsymbol{x}))$ using a weighted SVM by minimizing

$$\begin{aligned} V_{nj_2}(f) &= n^{-1}\sum_{i=1}^{n} I\left(y_i \neq j_1, \widehat{f}_{j_1}(\boldsymbol{x}_i) < 0\right) \\ &\quad \times\left\{\frac{k-2}{k-1}I(z_{ij_2}=1)[1-f(\boldsymbol{x}_i)]_+ + \frac{1}{k-1}I(z_{ij_2}=-1)[1+f(\boldsymbol{x}_i)]_+\right\} \\ &\quad +\lambda_{nj_2}\|f\|^2, \end{aligned}$$

where $z_{ij_2} = 1 - 2I(y_i \neq j_2)$, and $\lambda_{nj_2}$ is a tuning parameter.

In general, at step $l$ $(l = 1, 2, ..., k-1)$, we obtain the rule $\text{sign}(\widehat{f}_{jl}(\boldsymbol{x}))$ by minimizing

$$\begin{aligned} V_{nj_l}(f) &= n^{-1} \sum_{i=1}^{n} I\left(y_i \neq j_1, ..., y_i \neq j_{l-1}, \widehat{f}_{j_1}(\boldsymbol{x}_i) < 0, ..., \widehat{f}_{j_{l-1}}(\boldsymbol{x}_i) < 0\right) \\ &\quad \times \left\{ \frac{k-l}{k-l+1} I(z_{ij_l} = 1)[1 - f(\boldsymbol{x}_i)]_+ + \frac{1}{k-l+1} I(z_{ij_l} = -1)[1 + f(\boldsymbol{x}_i)]_+ \right\} \\ &\quad + \lambda_{nj_l} \|f\|^2, \end{aligned}$$

where $z_{ij_l} = 1 - 2(y_i \neq j_l)$. Note that we use weight $(k-1)/(k-l+1)$ vs $1/(k-l+1)$ to account for the classification of one class versus multiple classes.

Based on this sequence of the forward SVMs, we conclude that a subject will be classified into class $k$, the pre-determined target class, if

$$\widehat{f}_{j1}(\boldsymbol{x}) < 0, \quad \widehat{f}_{j2}(\boldsymbol{x}) < 0, \quad ... \quad \widehat{f}_{j,k-1}(\boldsymbol{x}) < 0.$$

For notation simplification, we denote $\widehat{\mathcal{D}}_j^k(\boldsymbol{x}) = 1$ if the above conditions hold and let $\widehat{\mathcal{D}}_j^k(\boldsymbol{x}) = -1$ otherwise.

The choice of this sequential binary classification is based on the $j$th permutation $(j_1, ..., j_{k-1})$, and so it may not exhaust the correct classification with label $k$ due to this specific choice. Thus, we repeat the above forward SVM for any possible permutation to obtain $\widehat{\mathcal{D}}_j^k(\boldsymbol{x})$ for all $j = 1, ...., (k-1)!$. Consequently, our final classification rule to classify a subject with label $k$ if and only if $\widehat{\mathcal{D}}_j^k(\boldsymbol{x}) = 1$ for at least one permutation $j$. That is, if we define

$$\widehat{\mathcal{D}}^k(\boldsymbol{x}) = \max_{j=1}^{(k-1)!} \widehat{\mathcal{D}}_j^k(\boldsymbol{x}),$$

then we classify a subject with label $k$ if and only if $\widehat{\mathcal{D}}^k(\boldsymbol{x}) = 1$.

Next, we aim to construct a decision rule for class label $(k-1)$. We adopt a backward elimination procedure. Delete the samples whose class labels are not $k$ and are not classified with label $k$ in the previous step. In other words, we restrict the training dataset to samples with $y_i \neq k$ and $\mathcal{D}^k(\boldsymbol{x}_i) = -1$. Because the data consist of only $(k-1)$ class labels, we use the same forward SVM as before but now set class label $(k-1)$ as the final class in the sequential classification algorithm. By this procedure, we obtain a decision rule at each step of each permutation of $\{1, 2, .., k-2\}$, denoted by $\widehat{\mathcal{D}}_j^{(k-1)}(\boldsymbol{x})$, $j = 1, ..., (k-2)!$. Consequently, we will classify a subject with label $(k-1)$ if and only if $\widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = 1$ and $\widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1$, where

$$\widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = \max_{j=1}^{(k-2)!} \widehat{\mathcal{D}}_j^{(k-1)}(\boldsymbol{x}).$$

We continue this backward elimination and forward SVM to obtain $\widehat{\mathcal{D}}^{(k-2)}(\boldsymbol{x}), ..., \widehat{\mathcal{D}}^{1}(\boldsymbol{x})$ in turn. Our final classification rule from FB-SVM is

$$\widehat{\mathcal{D}}(\boldsymbol{x}) = \begin{cases} k & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = 1 \\ k-1 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(k-1)}(\boldsymbol{x}) = 1 \\ \vdots & \quad \vdots \\ 2 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(2)}(\boldsymbol{x}) = 1 \\ 1 & \widehat{\mathcal{D}}^{(k)}(\boldsymbol{x}) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(\boldsymbol{x}) = -1, \widehat{\mathcal{D}}^{(2)}(\boldsymbol{x}) = -1. \end{cases} \tag{2.1}$$

Our algorithm for $k$-category FB-SVM can be summarized as follows:

Backward loop with target class $s \in \{k, ..., 1\}$:

> Inner loop: for each permutation of the remaining classes except the previously classified classes and target class $s$ , we perform a sequence of forward SVMs with weights to learn $\widehat{\mathcal{D}}_j^{s}(\boldsymbol{x})$ $(j = 1, ..., (s-1)!)$.
>
> We collect all rules to obtain $\widehat{\mathcal{D}}^{s}(\boldsymbol{x}) = \max_{j=1}^{(s-1)!} \widehat{\mathcal{D}}_j^{s}(\boldsymbol{x})$.

After eliminating all samples with true labels of already classified classes or are classified with any of the previous labels, go to the backward loop step.

We note that FB-SVM requires a total of

$$\sum_{l=1}^{k}(l-1)\times(l-1)! = k! - 1$$

weighted binary SVM classifications. However, because of the sequential data elimination, the size of the input dataset keeps decreasing in a proportional fashion. Therefore, FB-SVM can be computationally efficient due to the fast implementation of SVM and reduced data sizes. In our numeric implementation, SVM at each step is implemented in MATLAB with package LIBSVM (Chang and Lin, 2011).

## 2.3 Theoretical Properties

In this section, we provide important theoretical properties for the proposed FB-SVM, including its Fisher consistency and convergence rate in terms of $n$. Zhang (2004); Tewari and Bartlett (2007); Liu (2007) have discussed Fisher consistency for several existing convex, margin-based multicategory loss functions. For multicategory hinge losses, geometric illustrations of Fisher consistency have been provided by Hill and Doucet (2007). First, we prove that the proposed FB-SVM is Fisher consistent. That is, when the sample size is infinity, the derived classification rule is the same as the Bayesian rule given as

$$\text{argmax}_{l=1}^{k} P_l(\boldsymbol{x}),$$

where $P_l(\boldsymbol{x}) = P(Y = l|\boldsymbol{X} = \boldsymbol{x})$. We introduce some notation for the classification rule from FB-SVM when $n = \infty$. Let $f_{jl}^*(\boldsymbol{x})$ be the counterpart of $\widehat{f}_{jl}(\boldsymbol{x})$ in the FB-SVM procedure when $n = \infty$ and the tuning parameters are zeros. Similarly, we let $\mathcal{D}_j^{*l}(\boldsymbol{x})$ and $\mathcal{D}^{*l}(\boldsymbol{x})$ be the corresponding limits of $\widehat{\mathcal{D}}_j^l(\boldsymbol{x})$ and $\widehat{\mathcal{D}}^l(\boldsymbol{x})$, respectively, at $n = \infty$. The final classification rule is $\mathcal{D}^*(\boldsymbol{x})$ given by

$$\mathcal{D}^*(\boldsymbol{X}) = \begin{cases} k & \mathcal{D}^{*(k)}(\boldsymbol{X}) = 1 \\ k-1 & \mathcal{D}^{*(k)}(\boldsymbol{X}) = -1, \mathcal{D}^{*(k-1)}(\boldsymbol{X}) = 1 \\ \vdots & \quad \vdots \\ 2 & \mathcal{D}^{*(k)}(\boldsymbol{X}) = -1, \ldots, \mathcal{D}^{*(3)}(\boldsymbol{X}) = -1, \mathcal{D}^{*(2)}(\boldsymbol{X}) = 1 \\ 1 & \mathcal{D}^{*(k)}(\boldsymbol{X}) = -1, \ldots, \ \mathcal{D}^{*(3)}(\boldsymbol{X}) = -1, \mathcal{D}^{*(2)}(\boldsymbol{X}) = -1. \end{cases} \tag{2.2}$$

**Theorem 2.1.** *FB-SVM classification rule $\mathcal{D}^*(X)$ is Fisher consistent. That is, $\mathcal{D}^*(\boldsymbol{x}) = l$ if and only if $P_l(\boldsymbol{x}) = \max_{h=1}^k P_h(\boldsymbol{x})$ for $l = 1, ..., k$.*

Next, we provide the convergence of the predicted misclassification rate from the Bayes error under some additional conditions for $P_l(\boldsymbol{x}), l = 1, ..., k$, assuming that the functional spaces for $f_{j_l}$ are from a reproducing kernal Hilbert space (RKHS) with Gaussian kernel and bandwidth $1/\sigma_n$. Specifically, we assume:

**Condition 2.1.** *There exists a positive constant $\delta_0$ such that $0 < \delta_0 < P_l(\boldsymbol{x}) < 1 - \delta_0 < 1$ for any $l = 1, ..., k$ and $\boldsymbol{x}$ in the domain of $\mathbf{X}$.*

**Condition 2.2.** *(Geometric noise conditions) There exist $q, \beta > 0$, and a constant $c$ such that for any permutation of $\{1, ..., k\}$, denoted by $\{j_1, ..., j_k\}$, it holds that for any $s > l$,*

$$P\left\{\left|\eta_l(\mathbf{X}) - \frac{1}{s-l+1}\right| < t\right\} \leq (ct)^q, \quad l = 1, ..., k,$$

*where $\eta_l(\boldsymbol{x}) = P(Y = j_l \Big| Y \in \{j_l, j_{l+1}, ..., j_s\}, \mathbf{X} = \boldsymbol{x})$, and moreover,*

$$E\left[I(\Delta(\mathbf{X}) < t)\left|\eta_l(\mathbf{X}) - \frac{1}{s-l+1}\right|\right] \leq ct^\beta,$$

*where $\Delta(\mathbf{X})$ denotes the distance from $\mathbf{X}$ to set $\{\boldsymbol{x} : \eta_l(\boldsymbol{x}) = 1/(s-l+1)\}$.*

**Condition 2.3.** *The distribution of $\mathbf{X}$ satisfies tail component condition $P(|\mathbf{X}| \geq r) \leq cr^{-\tau}$ for some $\tau \in (0, \infty]$.*

**Condition 2.4.** *There exists $\lambda_n$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$. Moreover, all tuning parameters $\lambda_{nj}$ satisfy $M^{-1}\lambda_n \leq \lambda_{nj} \leq M\lambda_n$. We further assume $\sigma_n \to \infty$.*

**Remark 2.1.** *In condition (2.2), the constants $q$ and $\beta$ are called noise exponent and marginal noise exponent, respectively. They are used to characterize the data distribution near the classification boundary when we classify category $j_l$ versus $\{j_{l+1}, ..., j_k\}$. In particular, when the boundary is fully separable, that is $|\eta_l - 1/(s-l+1)| > \delta_0$ for a constant $\delta_0$, these conditions hold for $q = \beta = \infty$. In condition (2.3), $\tau$ describes the decay of the distribution of $\mathbf{X}$. Obviously, when $\mathbf{X}$ is bounded, $\tau = \infty$. Condition (2.4) assumes the choice of tuning parameter and bandwidth in RKHS. We choose this simplification for convenience, although we can allow the tuning parameter and bandwidth to be different for each classification in the proposed method.*

Under conditions (2.1)-(2.4), we show that the following theorem holds.

**Theorem 2.2.** *Under conditions (C.1)-(C.4), for any $\epsilon_0 > 0$, $d/(d+\tau) < p \leq 2$, there exists a constant $C$ such that for any $\epsilon > 1$, with probability at least $1 - e^{-\epsilon}$,*

$$P(Y \neq \widehat{\mathcal{D}}(\mathbf{X})) \leq P(Y \neq \mathcal{D}^*(\mathbf{X})) + C\left\{\lambda_n^{\frac{\tau}{2+\tau}}\sigma_n^{-\frac{d\tau}{d+\tau}} + \sigma_n^{-\beta} + \epsilon\left(n\lambda_n^p\sigma_n^{\frac{1-p}{1+\epsilon_0 d}}\right)^{-\frac{q+1}{q+2-p}}\right\}^{q/(1+q)}.$$

**Remark 2.2.** *Suppose that $\mathbf{X}$ is bounded such that $\tau = \infty$ in condition (C.3). By choosing the optimal $\sigma_n$, we find that the convergence rate of the misclassification rate is a polynomial of order $n$, where the order is given by $-\beta q/[\beta(q+2) + d(q+1)] + \rho$ for any $\rho > 0$. If furthermore, the separating boundaries are all completely separable such that $\beta = q = \infty$, then the convergence rate is close to $n^{-1}$, which is the convergence rate of a change point model in parametric models.*

## 2.4 Simulation Studies

We conducted extensive simulation studies to compare FB-SVM with other learning methods. Specifically, the competing methods included WW in Weston et al. (1999), LLW in Lee et al. (2004), and one-versus-all (OVA). The RMSVM method in Liu and Yuan (2011) was computationally much more intensive even with linear kernels so was not included in the comparison. In the simulation studies, we considered four different scenarios: the first two are taken from Liu and Yuan (2011) with slight changes, which generates data retrospectively with feature variables $\mathbf{X}$ simulated given each class of $Y$. The third and fourth scenario are taken from Lee et al. (2004), which generates data prospectively with four categories, where feature variables $\mathbf{X}$ are simulated first and then the class label $Y$ is generated based on a multinomial distribution. There were 300 observations (100 for each category) for the first three scenarios and 400 for the last one. The details of the data generation are given as follows.

*Scenario 1.* We considered $Y$ as a 3-class category random variable with equal prevalence, i.e., $P(Y=1) = P(Y=2) = P(Y=3) = 1/3$. Informative feature variable $\mathbf{X}$ is 2-dimensional and for each $Y$ class, the distribution of $\mathbf{X}$ is given as:

$$\begin{aligned} &\mathbf{X}|Y=1 \sim N((\sqrt{3},-1)^T, 1.5^2 I_2), \mathbf{X}|Y=2 \sim N((-\sqrt{3},-1)^T, 1.5^2 I_2), \\ &\text{and } \mathbf{X}|Y=3 \sim N((0.2,1)^T, 1.5^2 I_2), \end{aligned}$$

where $I_2$ is a $2 \times 2$ identity matrix.

*Scenario 2.* We considered data generation similar to Scenario 1, except that the distribution of $\mathbf{X}$ given $Y$ is from the following:

$$\begin{aligned} &\mathbf{X}|Y=1 \sim N((2,0)^T, 1.5^2 I_2), \quad \mathbf{X}|Y=2 \sim N((-2,0)^T, 1.5^2 I_2), \\ &\text{and } \mathbf{X}|Y=3 \sim 0.5N((0,2)^T, 1.5^2 I_2) + 0.5N((0,-2)^T, 1.5^2 I_2). \end{aligned}$$

Note that the distribution of $\mathbf{X}$ in class 3 is from a mixture normal distribution.

*Scenario 3.* We considered that $X$ is one-dimensional and from $Unif(0,1)$. The class label $Y$ is generated from a multinomial distribution with probabilities

$$\begin{aligned} &P(Y=1|X) = 0.95\exp(-6X), P(Y=2|X) = 0.98\exp\{-5(X-1)^2\}, \\ &\text{and } P(Y=3|X) = 1 - P(Y=1|X) - P(Y=2|X). \end{aligned}$$

*Scenario 4. 4-Category* $X_1$ and $X_2$ are both one-dimensional and from $Unif(0,1)$. The class label $Y$ is generated from a multinomial distribution with probabilities

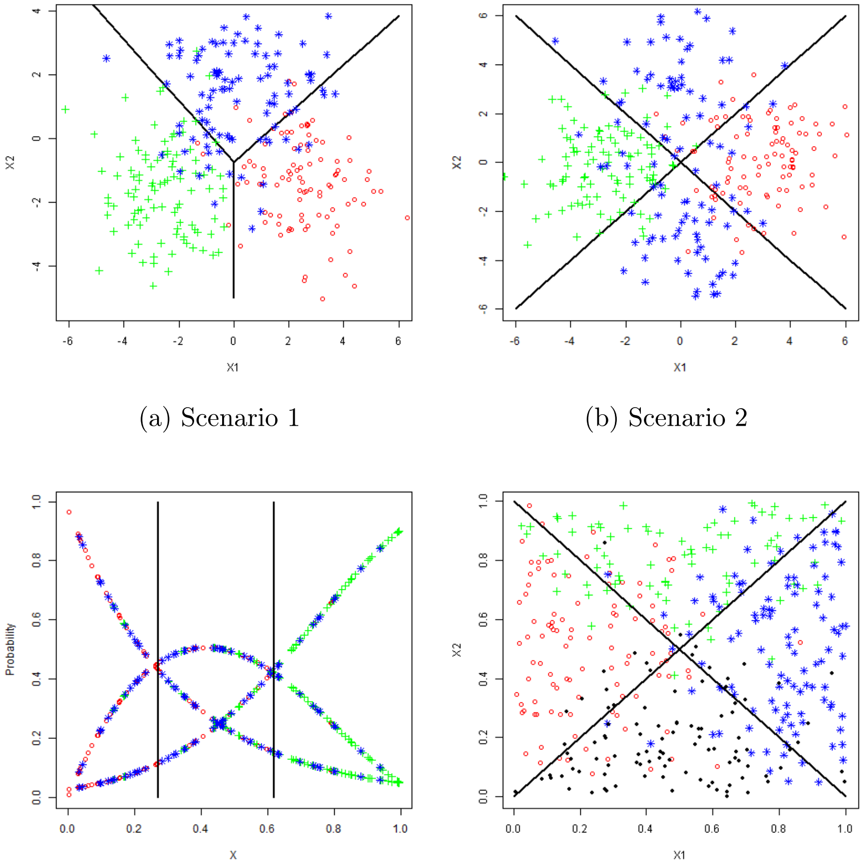
$$P(Y=1|\mathbf{X}) = c(\mathbf{X})\exp(-8(X_1^2 + (X_2-0.5)^2)),$$

$$P(Y=2|\mathbf{X}) = c(\mathbf{X})\exp(-8((X_1-0.5)^2+(X_2-1)^2)),$$

$$P(Y=3|\mathbf{X}) = c(\mathbf{X})\exp(-8((X_2-0.5)^2+(X_1-1)^2)),$$

$$\text{and } P(Y=4|\mathbf{X}) = c(\mathbf{X})\exp(-8(X_2^2+(X_1-0.5)^2)),$$

where $c(\mathbf{X})$ is normalizing function of $\mathbf{X}$ so that $\sum_{i=1}^{4} P(Y=i|\mathbf{X}) = 1$.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 2.1: Plots of each simulation setting with Bayes boundaries. Different classes are represented by 3 or 4 symbols and colors

Direct calculation yields that the optimal classification boundary in Scenario 1, 3 and 4 is linear but is nonlinear in Scenario 2, as shown in Figure 2.1. Additionally, for each setting, we also considered to include many noise variables unrelated to the class label in order to study the robustness of each method to noisy data in practice. Specifically, 50 noise variables normally distributed as $N(0, 0.5^2)$ and another 50 distributed as $N(0, 0.5^2) + 0.5X_1$ were added to Scenario 1 and 2; 20 noise variables normally distributed as $N(0, 0.1^2)$ and another 20 with $N(0, 0.1^2) + 0.5X_1$ were added to Scenario 3; and 20 noise variables normally distributed as $N(0, 0.1^2)$ and another 20 distributed as $N(0, 0.1^2) + 0.5X_1$ were added to Scenario 4. For each simulated dataset, we applied WW and LLW using package MSVMpack1.5 in MATLAB (Lauer and Guermeur, 2011). To apply our FB-SVM, we chose tuning parameter $\lambda$ among the set $\{2^{-5}, 2^{-4}, ..., 2^5\}$. We considered both linear kernel and Gaussian kernel in the estimation. The bandwidth of the latter kernel was set to $(10d)^{-1}$, where $d$ is the number of features, following the default setting for WW and LLW in MSVMpack1.5. Finally, to compare the prediction performance of all methods, we generated 5,000 observations independent for each class as testing data. All software was run on a Linux-based computing system with 2-core, 2.40 GHz Intel processors.

Table 2.1 summarizes the results based on 100 replicates in each simulation setting. Column “Test Error" is the average of misclassification error rates in the test dataset, and column “Std Dev" is the corresponding standard deviation from 100 replicates. Column “CPU (sec)" gives the average CPU running time for each replicate in terms of seconds. The corresponding results with noise variables are shown in Table 2.2. The corresponding Bayes error (i.e., the optimal misclassification rate) for these four scenarios is 0.2689, 0.2883, 0.2845 and 0.2798, respectively.

From Table 2.1, it is clear that the proposed method, FB-SVM, outperforms OVA by a large margin. When there are no noise variables, FB-SVM has similar misclassification rate to WW and LLW, but much faster to compute (up to more than 100 times improvement in speed). From Table 2.2 where noise variables are included, FB-SVM with Gaussian kernel continues to achieve misclassification close to Bayes error. Therefore, FB-SVM is quite robust to the presence of noise feature variables; however, some competing methods perform worse when noise variables are included. Although the LLW method has been shown to be Fisher consistent, it does not reveal a better classification accuracy than FB-SVM. WW has comparable misclassification rate compared to FB-SVM in this setting, but with a much slower computational speed. The computational time of our method can be as much as 100-fold faster than LLW and WW. The one-vs-all method has comparable computational cost as FB-SVM, but it has the lowest classification accuracy among all the methods.

We also included the results of multinomial regression. For scenarios with no noises, multinomial regression has comparable test error with FBSVM. However, it yields much higher test errors and

Table 2.1: Summary of Simulation Results from 100 Replicates (without Noise Variables)

| | Method | Linear kernel | | | Gaussian kernel | | |
|---|---|---|---|---|---|---|---|
| | | Test Err | Std Dev | CPU (sec) | Test Err | Std Dev | CPU (sec) |
| Sc 1 | FB-SVM | 27.3% | 0.3% | 0.6 | 28.5% | 0.7% | 0.9 |
| | OVA | 31.4% | 1.0% | 3 | 34.6% | 1.7% | 4 |
| | WW[1] | 27.2% | 0.3% | 16 | 27.5% | 0.4% | 11 |
| | LLW[2] | 30.5% | 3.6% | 58 | 27.2% | 0.3% | 17 |
| | MNR[3] | 27.1% | 0.2% | 0.1 | | | |
| Sc 2 | FB-SVM | 34.3% | 0.4% | 0.6 | 31.1% | 0.8% | 0.9 |
| | OVA | 43.6% | 0.8% | 3 | 37.8% | 1.8% | 2 |
| | WW | 34.3% | 0.5% | 26 | 29.8% | 0.5% | 11 |
| | LLW | 39.5% | 0.1% | 72 | 29.7% | 0.4% | 18 |
| | MNR | 40.3% | 1.1% | 0.1 | | | |
| Sc 3 | FB-SVM | 29.4% | 1.4% | 0.2 | 29.7% | 1.6% | 0.4 |
| | OVA | 41.7% | 8.7% | 0.8 | 33.9% | 6.3% | 1 |
| | WW | 32.0% | 1.5% | 14 | 32.6% | 0.5% | 14 |
| | LLW | 34.0% | 1.1% | 26 | 33.6% | 0.9% | 26 |
| | MNR | 28.6% | 0.4% | 0.1 | | | |
| Sc 4 | FB-SVM | 29.3% | 0.7% | 2 | 29.2% | 0.7% | 3 |
| | OVA | 40.5% | 1.7% | 2 | 39.6% | 4.3% | 1 |
| | WW | 28.3% | 0.4% | 28 | 28.5% | 0.6% | 32 |
| | LLW | 48.6% | 6.9% | 37 | 45.2% | 6.7% | 47 |
| | MNR | 28.1% | 0.3% | 0.1 | | | |

[1] Weston et al., 1999
[2] Lee et al., 2004
[3] Multinomial regression.

sometimes higher computational cost when the data contain noises. Although multinomial regression requires the least computational time for most of the time, the results relies heavily on the correctness of models.

## 2.5 Applications

### 2.5.1 Alzheimer's Disease Data

We applied the proposed method using data collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (`http://adni.loni.usc.edu/`; Weiner et al., 2010). The ADNI is a multi-center, ongoing, prospective study aimed at evaluating clinical, neuroimaging, genetic, and biochemical biomarkers for Alzheimer's disease (AD). An important research goal is to use baseline feature

Table 2.2: Summary of Simulation Results from 100 Replicates (with Noise Variables)

| | Method | Linear kernel | | | Gaussian kernel | | |
|---|---|---|---|---|---|---|---|
| | | Test Err | Std Dev | CPU (sec) | Test Err | Std Dev | CPU (sec) |
| Sc 1 | FB-SVM | 40.3% | 1.8% | 3 | 29.6% | 0.9% | 3 |
| | OVA | 44.3% | 1.7% | 1 | 33.9% | 0.8% | 2 |
| | WW[1] | 40.4% | 1.9% | 36 | 28.2% | 0.4% | 17 |
| | LLW[2] | 43.4% | 4.4% | 127 | 29.9% | 1.2% | 33 |
| | MNR[3] | 43.8% | 1.7% | 6 | | | |
| Sc 2 | FB-SVM | 46.3% | 1.3% | 3 | 33.8% | 0.4% | 3 |
| | OVA | 48.7% | 1.5% | 21 | 40.8% | 0.9% | 2 |
| | WW | 46.4% | 1.7% | 43 | 34.8% | 0.3% | 21 |
| | LLW | 49.6% | 4.4% | 129 | 38.0% | 1.5% | 41 |
| | MNR | 48.9% | 1.3% | 7 | | | |
| Sc 3 | FB-SVM | 32.4% | 1.2% | 1 | 32.7% | 1.0% | 1 |
| | OVA | 43.4% | 1.6% | 1 | 38.1% | 1.6% | 1 |
| | WW | 34.3% | 1.1% | 17 | 33.1% | 0.2% | 15 |
| | LLW | 36.5% | 1.5% | 32 | 35.6% | 3.1% | 29 |
| | MNR | 36.8% | 1.2% | 0.3 | | | |
| Sc 4 | FB-SVM | 34.3% | 7.7% | 11 | 36.6% | 1.7% | 13 |
| | OVA | 44.0% | 9.1% | 3 | 55.4% | 3.6% | 2 |
| | WW | 34.6% | 7.1% | 35 | 37.2% | 2.5% | 29 |
| | LLW | 48.1% | 16.4% | 60 | 42.5% | 2.2% | 43 |
| | MNR | 39.0% | 1.3% | 1 | | | |

[1] Weston et al., 1999
[2] Lee et al., 2004
[3] Multinomial regression.

variables to construct biomarker signatures that can distinguish cognitively normal (CN) healthy controls, subjects diagnosed as mild cognitive impairment (MCI) at baseline (according to ADNI protocol) but remained free of AD during the study (non-converters, MCI-nc), subjects diagnosed as MCI at the baseline but converted to AD during the study (converters, MCI-c), and patients diagnosed as AD at the baseline. Biomarkers that distinguish MCI-nc and MCI-c are especially useful for designing clinical trials that target MCI subjects with conversion to AD as the primary endpoint for the trial. Feature variables included in our analysis were personal health history, cerebrospinal fluid (CSF) biomarkers, neuropsychological assessments, and magnetic resonance imaging (MRI) measures. After eliminating missing values, the final analysis dataset contained 98 CN healthy controls and 74, 77, and 74 subjects in MCI-nc, MCI-c, and AD category, respectively (a total of 323 subjects). There were 32 feature variables, among which six features were derived from structural MRI scans (e.g., hippocampal volume). The goal of our analysis is to use all 32 feature variables to classify subjects into four categories of AD, MCI-nc, MCI-c, and CN. We used linear kernel with 10-fold cross validation to evaluate the performance of FB-SVM and compare it with LLW and WW. Tuning parameter $\lambda$ for each binary SVM was selected among the set $\{2^{-5}, 2^{-4}, ..., 2^{5}\}$. All continuous variables were standardized before analysis. Classification errors (overall and group-wise), standard deviation, and the CPU time in minutes are summarized in Table 2.3.

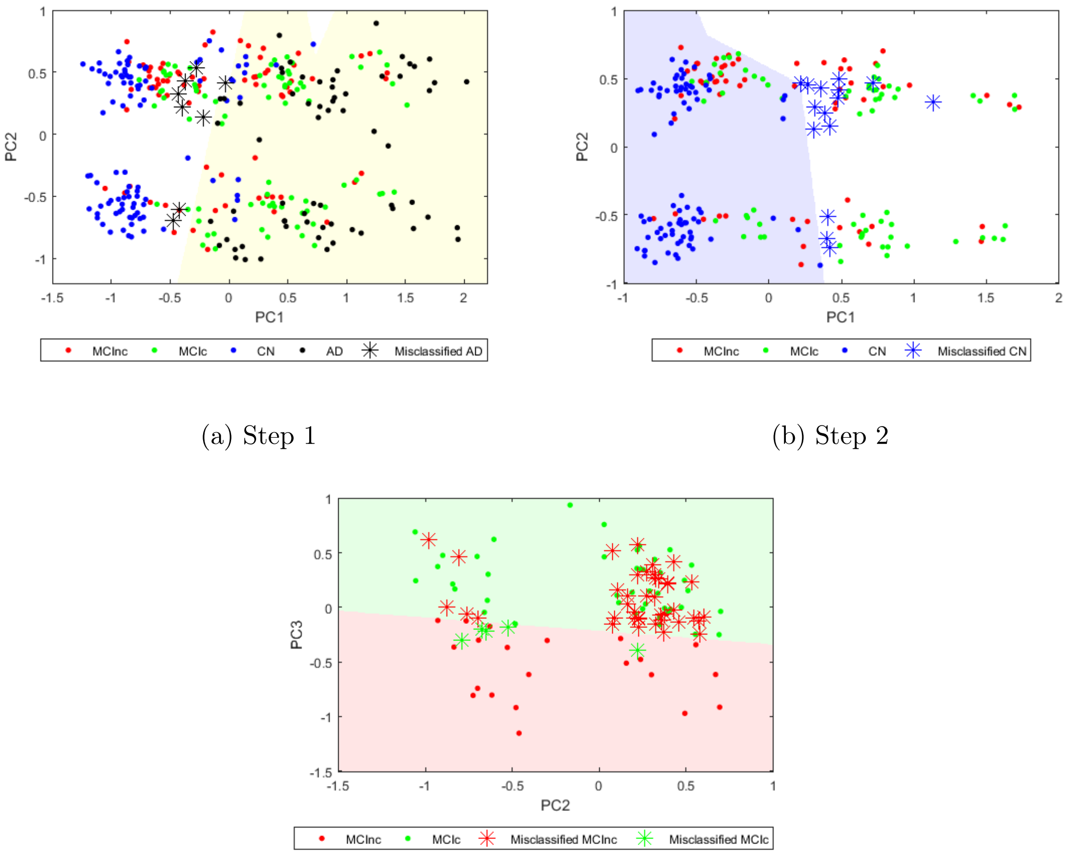
Table 2.3: Averages and standard deviation of the test errors based on 10-fold cross validation for the ADNI dataset.

| Method | Test Error | Std Dev | Misclassification Rate (%) | | | | CPU |
|---|---|---|---|---|---|---|---|
| | | | CN | MCI-nc | MCI-c | AD | (min) |
| FB-SVM | 28.8% | 6.8% | 7.1 | 20.4 | 16.8 | 13.3 | 0.8 |
| OVA | 31.2% | 8.6% | 5.9 | 25.7 | 20.1 | 10.8 | 0.3 |
| WW[1] | 29.1% | 5.8% | 6.5 | 22.0 | 20.1 | 9.6 | 6 |
| LLW[2] | 35.9% | 6.5% | 18.6 | 22.9 | 18.2 | 12.1 | 11 |

[1] Weston et al., 1999

[2] Lee et al., 2004

We can see that FB-SVM has the smallest overall test errors. Moreover, the proposed method is much more efficient and requires the least CPU time. The overall misclassification error for WW is close to FB-SVM, but for an important category, MCI-c, FB-SVM has a much better performance. To visualize the classification functions of FB-SVM, we projected the data onto the first two principal component (PC) directions of the feature variable space in Figure 2.2. In Figure 2.2a, our first step was to classify group AD from the other groups. In the second step, we kept only those subjects who had not been classified into AD in the first step and originally from the other 3 groups, and then classified group CN vs. MCI-nc and MCI-c. In step 3, we kept those subjects who had not been classified as AD or CN and had observed labels of MCI-nc or MCI-c in order to classify them into group MCI-nc or MCI-c. It can be seen from Figure 2.2 and Table 2.3 that the most difficult step is to classify group MCI subjects into converter and non-converter sub-groups. The misclassification rate for distinguishing MCI-nc and MCI-c is lower for FB-SVM than for WW and LLW. The accuracy of classifying MCI-nc versus MIC-c using multi-modal ADNI data as reported in the literature, is about 67% (Cui et al., 2011), which is lower than that obtained in our analysis. The percentage of the total variance explained by the first two PCs is 47% and 48% in steps 1 and 2, respectively; and is 26% explained by the second and third PC in step 3.



(a) Step 1

(b) Step 2

(c) Step 3

Figure 2.2: Principal conponent analysis (PCA) projection plots of FB-SVM classification steps with decision boundaries in ADNI data. Observed categories are represented by colored dots for correctly classified subjects and crosses for misclassified subjects from the target group. The background colors red, green, blue, and yellow mark the areas classified by FB-SVM as CN, MCI-nc, MCI-c, and AD, respectively. Step 1 classifies AD (black) vs. CN, MCI-nc, or MCI-c. Step 2 classifies CN (blue) vs. MCI-c or MCI-cn. Step 3 classifies MCI-nc (red) vs MCI-c.

### 2.5.2 Image Segmentation Data

We considered another application, the Image Segmentation dataset from the UCI Machine Learning website (`https://archive.ics.uci.edu/ml/datasets`). The problem to be solved was classification of image segmentations drawn from 7 outdoor images: 1: brickface, 2: sky, 3: foliage, 4: cement, 5: window, 6: path, and 7: grass. The images were hand-segmented to create a classification for every pixel. Each instance is a 3 × 3 region. We removed the third attribute, region-pixel-count, because the value is 9 for all instances. Thus, the final dataset contained 2310 instances and 18 continuous attributes. Again, 10-fold cross validation and linear rule were used to illustrate the performance of FB-SVM. Tuning parameter $\lambda$ for each binary SVM was selected among the set $\{2^{-5}, 2^{-4}, ..., 2^{5}\}$. Classification errors (overall and group-wise), standard deviation, and CPU time in hours are summarized in Table 2.4.

Table 2.4: Averages and standard deviation of the test errors based on 10-fold cross validation for the Image Segmentation dataset.

| Method | Test | Std | | | | Misclassification Rate (%) | | | | CPU |
|---|---|---|---|---|---|---|---|---|---|---|
| | Error | Dev | 1 | 2 | 3 | 4 | 5 | 6 | 7 | (hours) |
| FB-SVM | 9.2% | 2.0% | 0.3 | 0 | 4.1 | 4.5 | 5.8 | 3.5 | 0.3 | 18 |
| OVA | 14.0% | 1.8% | 10.8 | 0 | 2.9 | 6.3 | 5.8 | 7.7 | 0.2 | 0.02 |
| WW[1] | 13.7% | 2.1% | 0.3 | 5.2 | 5.4 | 8.3 | 6.9 | 1.0 | 0.4 | 110 |
| LLW[2] | 29.1% | 2.3% | 5.6 | 3.2 | 11.9 | 14.3 | 13.4 | 7.7 | 2.1 | 10 |

[1] Weston et al., 1999

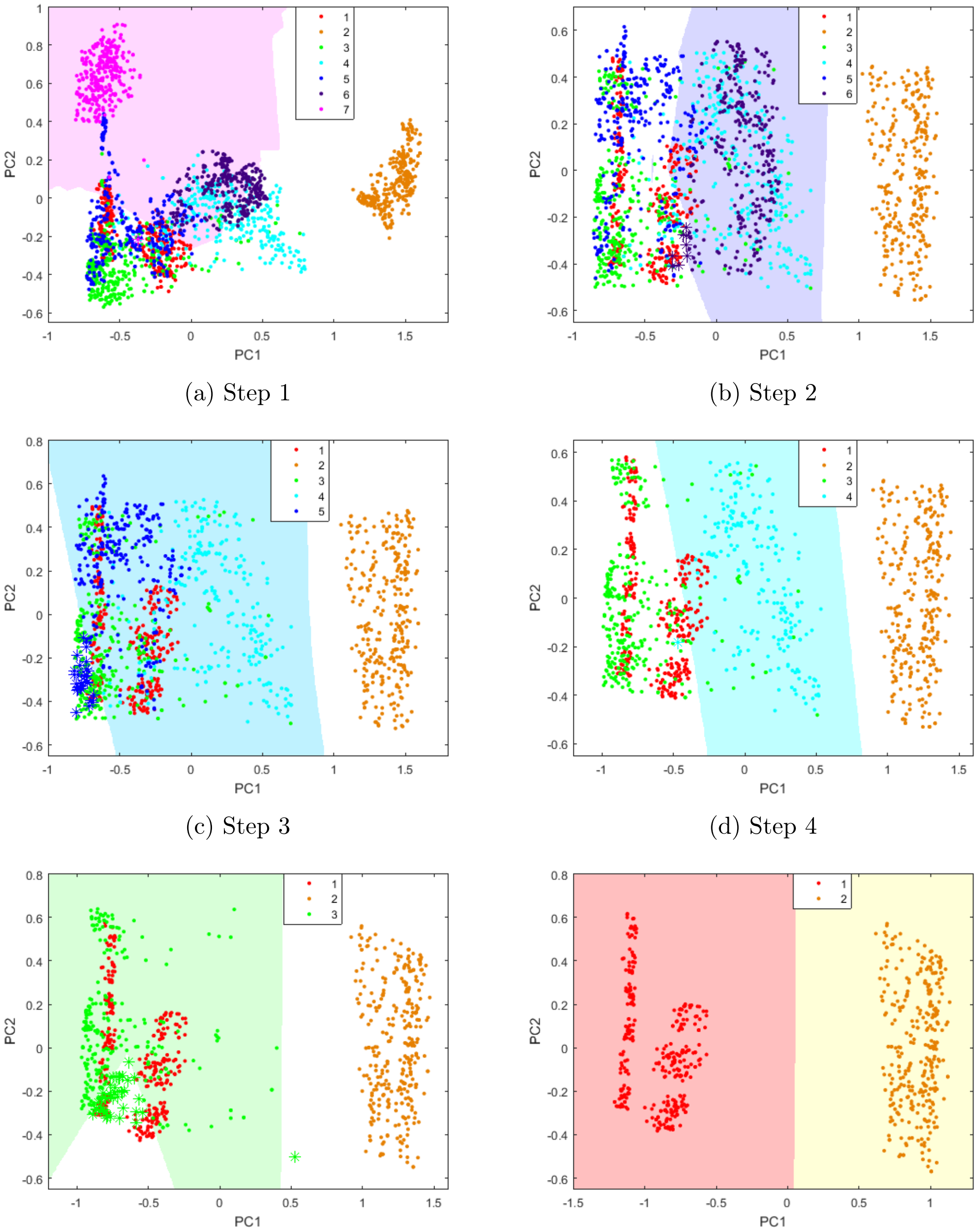
[2] Lee et al., 2004

From Table 2.4, we can see that FB-SVM has the smallest overall test error and competitive group-wise misclassification rates. OVA has the shortest running time but a worst test error. LLW requires a shorter computing time than FB-SVM, but has the lowest accuracy; this may be caused by its limitation on numeric convergence. Steps 1 through 6 in Figure 2.3 illustrate the FB-SVM classification function projected on the first two PC directions. As described in Section 2, our first step is to classify group 7 from the other 6 groups. In the second step, we only keep subjects not classified into group 7 and originally from the other 6 groups, and then classify group 6 versus groups 1 through 5. We complete the classification sequentially for all 7 groups in 6 steps. The classification functions in each step projected onto the first two PCs and overlaid on the observed class labels are depicted in Figure 2.3. The percentage of the total variance explained by the first two PCs is 74%, 79%, 83%, 77%, 80%, and 86% from steps 1 through step 6, respectively.

## 2.6 Conclusions

A number of algorithms are proposed in the literature as extensions of the SVM to solve the multicategory classification problems. However, most of the existing methods are neither Fisher consistent nor computationally efficient. In this paper, we propose a novel and computationally efficient learning algorithm to perform multicategory classification based on sequential binary SVMs, which is a convex optimization in each step and converges fast. We also establish the theoretical properties for the proposed FB-SVM, in terms of being Fisher consistent and the asymptotic convergence rate. In both simulation and application studies, FB-SVM outperforms alternatives, with a competitive running time, especially when compared with WW and LLW, where FB-SVM is seen to be as much as 100-fold faster in some settings.

Although FB-SVM is more computationally efficient when compared with other methods, the computational cost can be further reduced. One possible approach is to conduct parallel computing for all permutations in the algorithm in order to achieve a better efficiency, especially for applications with a large number of categories. Another extension is to explore a more efficient and effective methodology for a data-driven choice of tuning parameter $\lambda$.



(a) Step 1

(b) Step 2

(c) Step 3

(d) Step 4

(e) Step 5

(f) Step 6

Figure 2.3: PCA projection plots of FB-SVM classification steps with decision boundaries for Image Segmentation data. Observed categories are represented by colored dots for correctly classified subjects and crosses for misclassified subjects from the target group. The background colors red, yellow, green, cyan, blue, purple, and pink mark the areas classified by FB-SVM as groups 1 through 7, respectively.

# CHAPTER 3: SEQUENTIAL OUTCOME-WEIGHTED MULTI -CATEGORY LEARNING FOR ESTIMATING OPTIMAL ITR

## 3.1 Introduction

For many chronic diseases such as major depression and type 2 diabetes, treatment heterogeneity has been well documented where a treatment that is effective in the overall population may be highly ineffective in a subgroup of patients with specific characteristics (Trivedi Madhukar et al., 2008), or no longer beneficial after patients develop resistance (Lipska and Krumholz, 2014). On the other hand, in some cases a newly developed intervention may not be more efficacious compared to an existing active treatment in the overall population, but may reveal a large benefit in subgroups of patients (Carini et al., 2014). Henceforth, there has been a growing interest in understanding treatment heterogeneity and discovering individualized treatment rules tailored to patient-specific characteristics to maximize efficacy and achieve personalized medicine (Kosorok and Moodie, 2015). More specifically, tailored treatment strategy aims to recommend optimal treatment decision for an individual patient using information on a combination of his or her characteristics such as genomic features, medical treatment history and preference.

Recently, there has been a surge of statistical methods on estimating optimal treatment regimes involving a single decision point or multiple decision points using data collected from clinical trials or observational studies (Murphy, 2003; Robins, 2004; Moodie et al., 2007; Qian and Murphy, 2011; Zhao et al., 2011, 2012; Zhang et al., 2012, 2013). A class of popular methods is regression based Q-learning (Watkins, 1989; Murphy, 2005; Qian and Murphy, 2011), which relies on some postulated models to incorporate treatment-by-covariate interactions. Alternatively, Zhao et al. (2012), Zhang et al. (2012, 2013) proposed machine learning algorithms, for instance, outcome weighted learning (O-learning), to choose optimal treatment rules by directly optimizing the expected clinical outcome under the rule, referred as the value function, and draw connection with a classification problem. Most of these methods are designed to estimate optimal treatment rules for each patient between two treatment options. However, in many clinical applications it is common that more than two treatments are being compared. For example, in our motivating study, Research Evaluating the Value of Augmenting Medication with Psychotherapy (REVAMP) trial (Kocsis et al., 2009), non-responders or partial responders to a first-line antidepressant were randomized to three second-line treatment strategies.

When it comes to multiple-armed trials, the Q-learning approach, which relies heavily on the correctness of the postulated models, is more prone to model misspecification than for two-armed trials. For machine learning methods in Zhao et al. (2012), Zhang et al. (2012, 2013), multicategory treatment decision rules may be obtained via combining one-vs-one or one-vs-all comparisons. However, it is well documented in multicategory learning literature that the resulting classification rules from these methods are inconsistent so lead to sub-optimal performance (Dietterich and Bakiri, 1995; Kreßel, 1999; Allwein et al., 2001; Lee et al., 2004; Liu and Shen, 2006), due to possible conflicted decisions from pair-wise or one-vs-all comparisons. To the best of our knowledge, there has been no work on using multicategory learning to consistently estimate an optimal individualized treatment rules (ITRs) for multiple-armed trials.

In this paper, we propose a new approach to estimate optimal ITRs from multiple treatment options. Specifically, we transform the value maximization problem into a sequence of binary weighted classifications, which is referred as the sequential outcome-weighed multicategory (SOM) learning. At each step, we use a weighted binary support vector machine (SVM) to determine the optimal treatment for patients into one treatment category versus remaining treatment categories, where the weights are proportional to the outcome values and reflect the fact that a single treatment category is being compared to one or more treatment categories. We first estimate the optimal rule for a designated treatment option by excluding the possibility of declaring any other treatment as optimal via sequential SVMs; next, we exclude the treatments that have been already screened for optimality and repeat the same learning approach for the remaining treatment options. Theoretically, we show that the derived treatment rule is Fisher consistent. We demonstrate through extensive simulations that SOM learning has superior performance in comparison to Q-learning. Finally, an application of SOM learning to REVAMP shows that an ITR tailored to individual characteristics such as patients' expectancy of treatment efficacy and baseline depression severity reduces depressive symptoms more than a non-personalized treatment strategy.

The rest of the paper is structured as follows. Section 2 introduces the main idea and the mathematical framework for multicategory ITRs and formulate the problem for SOM learning. The detailed algorithm is then provided in the section. In Section 3, we provide theoretical justification for SOM learning. Sections 4 and 5 present extensive simulations and application to REVAMP to examine the performance of SOM. Finally, concluding remarks are provided in Section 6. The proof of theoretical results are presented in the Appendix.

## 3.2 Methodology

### 3.2.1 Optimal ITR with multiple treatments

Assume data are collected from a randomized trial with $n$ patients and $k$ different treatment options. For each patient $i$, we observe a $d$-dimensional vector of feature variables, denoted by $X_i \in \mathcal{X}$, a treatment assignment $A_i \in \mathcal{A} = \{1, 2, ..., k\}$, $i = 1, ..., n$, and the clinical outcome after treatment denoted by $R_i$ (also referred as the "reward"), with larger values of $R_i$ being more desirable. A multicategory ITR, denoted by $\mathcal{D}$, is a mapping from the space of feature variables, $\mathcal{X}$, to the domain of treatments, $\mathcal{A}$. An optimal ITR is a treatment assignment rule that maximizes the mean reward $E[R(\mathcal{D}(X))|X]$, where $R(a)$ is the potential outcome had treatment $a$ been given. According to Qian and Murphy (2011), for randomized trials and assuming consistency of the potential outcomes, the optimal ITR maximizes the following value function:

$$E\left[\frac{I\{A = \mathcal{D}(x)\}}{\pi_A(X)} R\right],$$

where $\pi_a(x) = pr(A = a|X = x)$ is the randomization probability for treatment $a$, $a = 1, ..., k,$, assumed to be bounded by a positive constant from below, so $\sum_{a=1}^{k} \pi_a(x) = 1$. The goal is to learn the optimal treatment rule using empirical observations $(R_i, A_i, X_i), i = 1, ..., n$.

Theoretically, it can be easily shown that the optimal ITR is

$$\mathcal{D}^*(x) = \text{argmax}_a E[R|A = a, X = x].$$

Therefore, one approach to estimate the optimal ITR is using a regression model to estimate the conditional means on the right-hand side. However, this approach heavily relies on the correctness of the postulated model, and model misspecification can lead to substantially non-optimal ITR even for a binary treatment situation (Zhao et al., 2012). Alternatively, Zhao et al. (2012) directly maximized the empirical version of the value function but replaced $I(A = \mathcal{D}(x))$ by $1 - \max(0, 1 - Af(x))$, where $f(x)$ is the decision function such that $\mathcal{D}(x) = \text{sign}(f(x))$. The latter corresponds to a weighted support vector machine where the weight for each observation is proportional to $R_i$. Because of this connection, the method is referred as outcome-weighted learning (abbreviated as O-learning). They demonstrated that O-learning outperformed the regression model based method in small sample. However, the proposed method can only be applied to estimate binary treatment decisions, and thus not directly applicable when more than two treatment options are of interest. Here, we aim to develop a robust method based on machine learning, which builds on O-learning for binary decision rules, to learn optimal multicategory treatment decision rules.

### 3.2.2 Main idea

The main idea of our method, namely, sequential outcome-weighted multicategory (SOM) learning, is to perform a sequence of binary treatment decision learning, where each step in the sequence determines whether the optimal treatment for a patient should be a candidate treatment category or the other treatments. To illustrate this idea, we order the candidate treatment categories based on the descending order of their prevalence. Without loss of generality, we assume that the order of the labels of treatment options are $k, k-1, ..., 1$.

We first aim to learn an optimal treatment rule that will determine whether a subject should be optimally treated with the $k$th option. Equivalently, we partition the domain of $X$ into $\mathcal{X}_k$ and $\mathcal{X}_k^c$ such that for subject with feature values $X \in \mathcal{X}_k$, the optimal treatment is the $k$th option; and for subject with $X \in \mathcal{X}_k^c$, the optimal treatment is not the $k$th option. To this end, we consider any ordered sequence of $\{1, ..., k-1\}$, denoted by $\{j_1, ..., j_{k-1}\}$, and let $j_k = k$. A sequential ITR learning is then conducted as follows.

In the first step, starting with $j_1$ versus $\{j_2, ..., j_k\}$, we determine whether a subject should be treated optimally with the $j_1$th option or some other choice. Since this is a binary decision problem, we can use existing methods for learning a binary treatment decision rule, for example, O-learning, with additional modifications as explained in later section. With this binary rule, for a future patient with feature variables $X$, if he or she is assigned to treatment $j_1 \neq k$, then clearly, $X \in \mathcal{X}_k^c$. Otherwise, we cannot determine whether $X$ should be in $\mathcal{X}_k$ or $\mathcal{X}_k^c$, since his/her optimal treatment can be one of $j_2, ..., j_k$.

In the second step of this sequential learning, we only consider patients whose optimal treatments are not determined as $j_1$ in the previous step. We then aim to learn a binary treatment rule to decide whether this subject should be optimally treated with $j_2$ or the remaining treatment choices, $\{j_3, ..., j_k\}$. Again, this is a binary treatment decision problem so we can perform estimation similar to the first step. With the second decision rule, we can check whether the patient should be treated with $j_2$ or some treatment among the remaining options. If $j_2 \neq k$ is selected, we conclude $X \in \mathcal{X}_k^c$; otherwise, we are still uncertain whether $X \in \mathcal{X}_k$.

Continue this process sequentially in the third step till the $k$th step when there is only treatment category $k$ in consideration. Consequently, for this given sequence, $\{j_1, ..., j_{k-1}\}$, the optimal treatment for this patient is $k$, i.e., $X \in \mathcal{X}_k$, if and only if at each step, the binary decision learning concludes that the patient should not be treated by $j_1, j_2, ..., j_{k-1}$ in turn. The choice of the ordered sequence $\{j_1, ..., j_{k-1}\}$ is arbitrary, so we propose to consider all possible permutations of $\{1, ..., k-1\}$. Then a patient with $X$ should be treated with treatment $k$ once he/she is determined to have the $k$th option as the optimal treatment in at least one permuted sequence.

The above procedure only provides a treatment rule that decides whether a subject has the optimal treatment as $k$ ($X \in \mathcal{X}_k$) or some other option ($X \in \mathcal{X}_k^c$). Thus, for a subject with features $X \in \mathcal{X}_k^c$, we need to determine which of the remaining $\{1, ..., k-1\}$ options is optimal. This can be carried out using the following procedure. We only consider patients whose optimal treatment is not $k$ based the previous procedure and whose actual treatments received are not $k$. For these patients, the optimal treatment options can only be one of $\{1, ..., k-1\}$, so the goal reduces to finding the optimal treatment decision within $(k-1)$ categories. Therefore, we can repeat the previous procedure but consider treatment $(k-1)$ as the target in treatment optimization. At the end, we obtain a treatment rule that determines whether a subject should be optimally treated with $(k-1)$. Finally, the same procedure can be carried out sequentially to decide which patients have the optimal treatment as $(k-2), \ldots, 1$, in turn.

Clearly, an advantage of SOM learning is that at every step of the sequential learning, one only needs to learn a binary decision rule, and thus many learning algorithms for binary decision are applicable. In particular, in our subsequent algorithm and implementation, we adopt the method from O-learning (Zhao et al., 2012) to use a weighted support vector machine (SVM) for this purpose. However, one significant difference is that due to the multicategory nature, weights in SOM learning should not only be proportional to the outcome $R$ as in O-learning, but should also reflect the imbalanced comparison between one treatment category and the combination of multiple treatment categories. The latter ensures that the derived optimal treatment rule is Fisher consistent, as will be shown in Section 3.

### 3.2.3 Method and algorithm

Mathematically, we can express SOM learning algorithm as follows. Start from the target decision for treatment $k$. Consider the $j$th permutation $\{j_1, ..., j_{k-1}\}$, and let $j_k = k$. The main algorithm is as follows:

1. At step 1, recall from the previous section that our goal is to learn a binary rule to decide whether a future patient should be treated by option $j_1$. Intuitively, we shall estimate the optimal decision function $f^*_{j_1}(X)$ such that the corresponding value for this decision, i.e.,

$$E\left[RI\{Z_{j_1} f_{j_1}(X) > 0\}/\pi_A(X)\right],$$

is maximized with $Z_{j_1} = 2I(A = j_1) - 1$. According to Liu et al. (2014), even if $R$ may be negative, this maximization is equivalent to minimize

$$\begin{aligned} & E \quad [|R|I\{Z_{j_1}\text{sign}(R) f_{j_1}(X) \leq 0\}/\pi_A(X)] \\ = & E \quad \left[\frac{|R|}{\pi_A(X)} I\{Z_{j_1}\text{sign}(R) = 1\}\{1 + f_{j_1}(X)\}\right] \\ + & E \left[\frac{|R|}{\pi_A(X)} I\{Z_{j_1}\text{sign}(R) = -1\}\{1 + f_{j_1}(X)\}\right]. \end{aligned}$$

Thus, using the empirical data, we may consider minimizing the empirical version the above expectation. However, there are two important issues to be considered. First, since solving the above problem is NP-hard, we can use a weighted SVM which essentially replaces the above 0-1 loss with a continuous and convex hinge-loss function. Second, since this learning is comparing one treatment category versus $(k-1)$ categories, it is necessary to weight observations with treatment $j_1$ by $(k-1)/k$ and the others by $1/k$ in order to balance the comparison.

Therefore, our algorithm is as follows following the above rationale. We define $\pi_{j_l}(x) = pr(A = j_l|X = x)$, where $l = 1, ..., k$, and let $Z_{ij_l} = 2I(A_i = j_l) - 1$. We then estimate the optimal decision rule as $\text{sign}(\widehat{f}_{j_1}(x))$, where $\widehat{f}_{j_1}(x)$ minimizes the following empirical risk of a weighted hinge loss:

$$V_{nj_1}(f) \quad = \quad n^{-1} \sum_{i=1}^{n} \left[\frac{|R_i|}{\pi_{j_1}(X_i)} I\{Z_{ij_1}\text{sign}(R_i) = 1\}\{1 - f(X_i)\}_+ \left\{\frac{k-1}{k} I(R_i > 0) + \frac{1}{k} I(R_i \leq 0)\right\}\right.$$

$$
+\frac{|R_i|}{\pi^*_{j_1}(X_i)} I\{Z_{ij_1}\text{sign}(R_i) = -1\}\{1 + f(X_i)\}_+ \left\{\frac{1}{k} I(R_i > 0) + \frac{k-1}{k} I(R_i \le 0)\right\}\Bigg]
$$
$$
+\lambda_{nj_1}\|f\|^2,
$$

where $\pi^*_{j_1}(X_i) = \sum_{l=2}^k I\{A_i = j_l\}\pi_{j_l}(X_i)$, $x_+ = \max(x, 0)$ is the hinge loss, $\|\cdot\|$ denotes a semi-norm for $f$ and $\lambda_{nj_1}$ is a tuning parameter. Particularly, consider a linear decision rule, i.e., $f(x) = \beta^T x + \beta_0$, $\|f\|$ is chosen as the Euclidean norm of $\beta$; if a nonlinear decision rule is desired, $f$ will be chosen from a reproducing kernel Hilbert space and $\|f\|$ is the corresponding norm in that space.

2. At step 2, recall from the previous section that this step aims to find the optimal decision to be either treatment $j_2$ or one of $\{j_3, ..., j_k\}$ among those patients whose optimal treatments are not determined as $j_1$ from step 1. Thus, restrict the data to those subjects who do not receive the $j_1$th treatment and whose optimal treatments are not $j_1$ as from the previous step. We then estimate a decision rule as $\text{sign}(\widehat{f}_{j_2}(x))$ using a weighted SVM by minimizing

$$
\begin{aligned}
V_{nj_2}(f) \quad = \quad & n^{-1}\sum_{i=1}^n I\left\{A_i \neq j_1, \widehat{f}_{j_1}(X_i) < 0\right\} \\
& \times \Bigg[\frac{|R_i|}{\pi_{j_2}(X_i)} I\{Z_{ij_2}\text{sign}(R_i) = 1\}\{1 - f(X_i)\}_+ \left\{\frac{k-2}{k-1} I(R_i > 0) + \frac{1}{k-1} I(R_i \le 0)\right\} \\
& + \frac{|R_i|}{\pi^*_{j_2}(X_i)} I(Z_{ij_2}\text{sign}(R_i) = -1)\{1 + f(X_i)\}_+ \left\{\frac{1}{k-1} I(R_i > 0) + \frac{k-2}{k-1} I(R_i \le 0)\right\}\Bigg] \\
& + \lambda_{nj_2}\|f\|^2,
\end{aligned}
$$

where $\pi^*_{j_2}(X_i) = \sum_{l=3}^k I\{A_i = j_l\}\pi_{j_l}(X_i)$, $Z_{ij_l}, \pi_{j_l}(X_i)$ are defined as same as in step 1, and $\lambda_{nj_2}$ is a tuning parameter. Note that in addition to weights based on the outcome values, we also weigh the observations from treatment $j_2$ by $(k-2)/(k-1)$ and the others by $1/(k-1)$ in order to account for the fact that the decision rule is based on comparing one category versus $(k-2)$ categories.

3. In turn, at step $h = 3, ..., k-1$, we obtain the rule as $\text{sign}(\widehat{f}_{jh}(x))$ by minimizing

$$
\begin{aligned}
V_{nj_h}(f) \quad = \quad & n^{-1}\sum_{i=1}^n I\left\{A_i \neq j_1, ..., A_i \neq j_{h-1}, \widehat{f}_{j_1}(X_i) < 0, ..., \widehat{f}_{j_{h-1}}(X_i) < 0\right\} \\
& \times \Bigg[\frac{|R_i|}{\pi_{j_h}(X_i)} I\{Z_{ij_h}\text{sign}(R_i) = 1\}\{1 - f(X_i)\}_+ \\
& \quad \times \left\{\frac{k-h}{k-h+1} I(R_i > 0) + \frac{1}{k-h+1} I(R_i \le 0)\right\} \\
& \quad + \frac{|R_i|}{\pi^*_{j_h}(X_i)} I\{Z_{ij_h}\text{sign}(R_i) = -1\}\{1 + f(X_i)\}_+ \\
& \quad \times \left\{\frac{1}{k-h+1} I(R_i > 0) + \frac{k-h}{k-h+1} I(R_i \le 0)\right\}\Bigg] + \lambda_{nj_h}\|f\|^2,
\end{aligned}
$$

where $\pi^*_{j_h}(X_i) = \sum_{l=h+1}^{k} I\{A_i = j_l\}\pi_{j_l}(X_i)$, $Z_{ij_l}, \pi_{j_l}(X_i)$ are defined as same as above steps. Again, we use weight $(k-h)/(k-h+1)$ for treatment $j_h$ versus $1/(k-h+1)$ for the others to balance comparison. At the end of this sequence, we conclude that if

$$\widehat{f}_{j_1}(x) < 0, \quad \widehat{f}_{j_2}(x) < 0, \quad ... \quad \widehat{f}_{j_{k-1}}(x) < 0,$$

the optimal treatment for a patient with features $x$ will be the $k$th option, the pre-determined target treatment category. For notational simplification, we denote $\widehat{\mathcal{D}}^k_{(j_1,\ldots,j_{k-1})}(x) = 1$ if the above conditions hold and let $\widehat{\mathcal{D}}^k_{(j_1,\ldots,j_{k-1})}(x) = -1$ otherwise.

The choice of this sequential decision rule is based on the permutation $(j_1, ..., j_{k-1})$, and thus may not exhaust the correct optimal treatment assignment for the $k$th option due to a specific choice. We thus repeat the above sequential learning for any possible $(k-1)!$ permutations to obtain $\widehat{\mathcal{D}}^k_{(j_1,\ldots,j_{k-1})}(x)$. Consequently, our final decision rule to assign a patient with treatment $k$ if and only if $\widehat{\mathcal{D}}^k_{(j_1,\ldots,j_{k-1})}(x) = 1$ for at least one permutation $(j_1, ..., j_{k-1})$. That is, if we define

$$\widehat{\mathcal{D}}^k(x) = \max_{(j_1,\ldots,j_{k-1}) \text{ is permutation of} \{1,..,k-1\}} \widehat{\mathcal{D}}^k_{(j_1,\ldots,j_{k-1})}(x),$$

then the optimal treatment for patient with $x$ is treatment $k$ if and only if $\widehat{\mathcal{D}}^k(x) = 1$.

4. To determine whether a patient should be optimally treated with the $(k-1)$th option, we adopt a backward elimination procedure. We exclude the patients who receive treatment option $k$ or whose optimal treatments are determined as $k$ in the previous step. In other words, we restrict the data to subjects with $A_i \neq k$ and $\widehat{\mathcal{D}}^k(x_i) = -1$. Because the data consist of only $(k-1)$ treatment options, we use the same SOM learning procedure as before but now set option $(k-1)$ as the target treatment, i.e., the last category in consideration in the above sequential learning algorithm. By this procedure, we obtain a decision rule at each step for each permutation of $\{1, .., k-2\}$, denoted by $\widehat{\mathcal{D}}^{(k-1)}_{(j_1,\ldots,j_{k-2})}(x)$ for permutation $(j_1, ..., j_{k-2})$. Let $\Pi_s$ denote all permutations of $(1, ..., s)$, and let

$$\widehat{\mathcal{D}}^{(k-1)}(x) = \max_{(j_1,\ldots,j_{k-2}) \in \Pi_{k-2}} \widehat{\mathcal{D}}^{(k-1)}_{(j_1,\ldots,j_{k-2})}(x).$$

Consequently, the optimal treatment for a patient with $x$ is $(k-1)$ if and only if $\widehat{\mathcal{D}}^{(k-1)}(x) = 1$ and $\widehat{\mathcal{D}}^{(k)}(x) = -1$.

5. Continue this backward elimination and sequential learning in turn for treatment $(k-2), ..., 1$ so as to obtain $\widehat{\mathcal{D}}^{(k-2)}(x), ..., \widehat{\mathcal{D}}^{1}(x)$. Our final estimated optimal ITR is

$$\widehat{\mathcal{D}}(x) = \begin{cases} k & \widehat{\mathcal{D}}^{(k)}(x) = 1 \\ k-1 & \widehat{\mathcal{D}}^{(k)}(x) = -1, \widehat{\mathcal{D}}^{(k-1)}(x) = 1 \\ \vdots & \quad \vdots \\ 2 & \widehat{\mathcal{D}}^{(k)}(x) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(x) = -1, \widehat{\mathcal{D}}^{(2)}(x) = 1 \\ 1 & \widehat{\mathcal{D}}^{(k)}(x) = -1, \ldots, \widehat{\mathcal{D}}^{(3)}(x) = -1, \widehat{\mathcal{D}}^{(2)}(x) = -1. \end{cases}$$

Our algorithm for $k$-category SOM learning can be summarized as follows:

Backward loop with target class $s \in \{k, ..., 1\}$:

Inner loop: for each permutation of the remaining treatment assignments except the previously classified ones and target treatment label $s$ , perform a sequence of weighted O-learning to learn $\widehat{\mathcal{D}}^{(j_1,\cdots,j_{s-1})}(x)$ for each permutation $(j_1, ..., j_s)$ of $\{1, .., s\}$.

Collect all rules to obtain $\widehat{\mathcal{D}}^{s}(x) = \max_{(j_1,\ldots,j_{s-1})\in\Pi_{s-1}} \widehat{\mathcal{D}}^{s}_{(j_1,\ldots,j_{s-1})}(x)$.

After eliminating all samples with actual treatment labels are previously considered treatment or whose optimal treatments are within any of the previous labels, go to the backward loop step.

We note that SOM learning requires a total of

$$\sum_{l=1}^{k} (l-1) \times (l-1)! = k! - 1$$

weighted binary SVM classifications. However, because of the sequential exclusion of subjects, the size of the input data decreases in a proportional fashion. Therefore, SOM learning can be computationally efficient due to the fast implementation of SVM and reduced data sizes in each step. In our numeric examples, SVM at each step is implemented in MATLAB with quadratic programming.

## 3.3 Theoretical Properties

In this section, we first establish the Fisher consistency of the optimal ITR estimated using SOM learning. Next, we obtain a risk bound for the estimated ITR and show how the bound can be improved for certain situations.

### 3.3.1 Fisher consistency

We provide the theoretical property of Fisher consistency for the proposed SOM learning. Specifically, when the sample size is infinity, we show that the derived ITR is the same as the true optimal ITR

$$\text{argmax}_{l=1}^{k} E(R|X=x, A=l).$$

Let $f^*_{j_l}(x)$ be the counterpart of $\widehat{f}_{j_l}(x)$ in the SOM learning procedure when $n=\infty$ and the tuning parameters vanishes. Let $\mathcal{D}^{*l}_{(j_1,\ldots,j_s)}(x)$ and $\mathcal{D}^{*l}(x)$ be the corresponding limits of $\widehat{\mathcal{D}}^{l}_{(j_1,\ldots,j_s)}(x)$ and $\widehat{\mathcal{D}}^{l}(x)$, respectively, when $n=\infty$. Then the limit of the ITR from SOM learning is

$$\mathcal{D}^*(x) = \begin{cases} k & \mathcal{D}^{*(k)}(x) = 1 \\ k-1 & \mathcal{D}^{*(k)}(x) = -1, \mathcal{D}^{*(k-1)}(x) = 1 \\ \vdots & \quad\vdots \\ 2 & \mathcal{D}^{*(k)}(x) = -1, \ldots, \mathcal{D}^{*(3)}(x) = -1, \mathcal{D}^{*(2)}(x) = 1 \\ 1 & \mathcal{D}^{*(k)}(x) = -1, \ldots, \mathcal{D}^{*(3)}(x) = -1, \mathcal{D}^{*(2)}(x) = -1. \end{cases}$$

The following result holds.

**Theorem 3.1.** *SOM learning rule $\mathcal{D}^*(X)$ is Fisher consistent. That is, $\mathcal{D}^*(x) = l$ if and only if $E(R|X=x, A=l) = \max_{h=1}^{k} E(R|X=x, A=h)$ for $l = 1, ..., k$.*

Theorem 1 provides a theoretical justification that the proposed SOM learning yields the true optimal ITR asymptotically. The proof of Theorem 1 is given in the appendix. The key result is to show that at each step of SOM learning, we compare the conditional mean $E[R|X, A=j_1]$ with the average value of $E[R|X, A=j_2]$, where $j_1$ is the treatment category in consideration at this step while $j_2$ is any treatment category among the remaining options.

### 3.3.2 Risk bounds

For any ITR $\mathcal{D}(x)$ associated with decision function $\mathcal{D}(x)$, define

$$\mathcal{R}(\mathcal{D}) = E\left[\frac{R}{\pi_A(x)} I\left\{A \neq \mathcal{D}(X)\right\}\right]$$

where $j = 1, ..., k$, $\pi_A(x) = \sum_{j=1}^{k} I(A = j)pr(A = j|x)$; and let $\mathcal{R}^* = \mathcal{R}(\mathcal{D}^*)$. Clearly, $\mathcal{R}(\mathcal{D})$ and $\mathcal{R}^*$ correspond to $E[R]$ subtracting the value for $\mathcal{D}$ and $\mathcal{D}^*$, respectively. In the section, we will derive the convergence rate of the estimated value function from the optimal value, which is equivalent to $\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}^*$, under some regularity conditions and assuming that the functional spaces for $f_{jl}$ in our SOM learning are from a reproducing kernel Hilbert space (RKHS) with Gaussian kernel and bandwidth $1/\sigma_n$.

For any $l$ and subset, $\mathcal{S}$, in $\{1, ..., k\}$ where $l \notin \mathcal{S}$, we define

$$\eta_{l,\mathcal{S}}(x) = \frac{E(R|X = x, A = l)}{|\mathcal{S}|^{-1} \sum_{h \in \mathcal{S}} E(R|X = x, A = h)},$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. That is, $\eta_{l,\mathcal{S}}(x)$ is the ratio between the mean outcome in treatment arm $l$ and the average mean outcome in treatment options from $\mathcal{S}$. We assume that the following conditions hold:

**Condition 3.1.** *(Geometric noise conditions) There exist $q, \beta > 0$, and a constant $c$ such that for any $l$ and set $\mathcal{S}$ with $l \notin \mathcal{S}$, it holds that*

$$r\left\{\Big|\eta_{l,\mathcal{S}}(X) - 1\Big| < t\right\} \leq (ct)^q,$$

*and moreover,*

$$E\left\{\exp\left(-\frac{\Delta(X)^2}{t}\right)\Big|\eta_{l,\mathcal{S}}(X) - 1\Big|\right\} \leq ct^{\beta},$$

*where $\Delta(X)$ denotes the distance from $X$ to the boundary defined as $\{x : \eta_{l,\mathcal{S}}(x) = 1\}$.*

**Condition 3.2.** *The distribution of $X$ satisfies tail component condition $pr(|X| \geq r) \leq cr^{-\tau}$ for some $\tau \in (0, \infty]$ and $E(|R||A = a, X = x)$ is uniformly bounded away from zero and infinity.*

**Condition 3.3.** *There exists $\lambda_n$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$. Moreover, all tuning parameters $\lambda_{nj}$'s in SOM satisfy $M^{-1}\lambda_n \leq \lambda_{nj} \leq M\lambda_n$ for a positive constant $M$. We further assume $\sigma_n \to \infty$.*

**Remark 3.1.** *In condition (3.1), the constants $q$ and $\beta$ are called noise exponent and marginal noise exponent, respectively. They are used to characterize the data distribution near the decision boundary at each step of SOM where we compare treatment $j_l$ versus any subset of $\{j_{l+1}, ..., j_k\}$. In particular, when the boundary is fully separable, that is, $|\eta_{l,\mathcal{S}} - 1| > \delta_0$ for a constant $\delta_0$, these conditions hold for $q = \beta = \infty$. In condition (3.2), $\tau$ describes the decay of the distribution of $X$. Obviously, when $X$ is bounded, $\tau = \infty$. Condition (3.3) assumes the choice of tuning parameter and bandwidth in RKHS. We choose this simplification for convenience, although we can allow the tuning parameter and bandwidth to be different for each treatment decision in the proposed method. Under conditions (3.1)-(3.3), the following theorem holds.*

**Theorem 3.2.** *Under conditions (3.1)-(3.3), for any $\epsilon_0 > 0$, $d/(d+\tau) < p \leq 2$, there exists a constant $C$ such that for any $\epsilon > 1$ and $\sigma_n = \lambda_n^{-q/(2\beta(1+q))}$, with probability at least $1 - e^{-\epsilon}$,*

$$\mathcal{R}(\widehat{\mathcal{D}}) \leq \quad \mathcal{R}^* + C\left\{\lambda_n^{-\frac{2}{2+p}+\frac{(2-p)(1+\epsilon_0)}{(2+p)(1+q)}} n^{-\frac{2}{2+p}} + \frac{\epsilon}{n\lambda_n} + \lambda_n^{\frac{q}{1+q}}\right\}^{\frac{q}{1+q}}.$$

**Remark 3.2.** *Suppose that $X$ is bounded such that $\tau = \infty$ in condition (3.2). By choosing the optimal $\lambda_n$ for the last two term in the right-hand side, i.e., $\lambda_n = n^{-(1+q)/(1+2q)}$, we find that the convergence rate is a polynomial order of $n$, where the order is $q/(1+2q)$. If furthermore, the separating boundaries are all completely separable such that $q = \infty$, then the convergence rate is close to the square-root rate.*

## 3.4 Simulation Studies

We conduct extensive simulation studies from four different settings to examine the small-sample performance of SOM learning. In the first three settings, 20 feature variables are simulated from a multivariate normal distribution, where the first 10 variables $X_1, X_2, ..., X_{10}$ have a pairwise correlation of 0.8, the remaining 10 variables are uncorrelated, and the marginal distribution for each variable is $N(0, 1)$. We generate 3-category random treatment assignments with equal probability, i.e. $pr(A = 1|X) = pr(A = 2|X) = pr(A = 3|X) = 1/3$. The reward functions for the first three settings are generated as follows:

*Setting 1.* $R = X_4 + (X_1 + X_2)I(A = 2) + (-X_1 + X_3)I(A = 3) + 0.5 \times N(0, 1)$

*Setting 2.* $R = X_4 + (X_2^2 - X_1^2)I(A = 2) + X_3^3 I(A = 3) + 0.5 \times N(0, 1)$

*Setting 3.* $R = (X_1 - 0.2) \times \{I(A = 1) - I(X_1 > 0.3)\}^2 + (X_2 + 0.3) \times \{I(A = 2) - I(X_2 > -0.5)\}^2 + (X_3 + 0.5) \times \{I(A = 3) - I(X_3 > 0)\}^2 + 0.5 \times N(0, 1)$.

In the last setting (*Setting 4*), we imitate a situation where the entire population consists of a finite number of latent subgroups for which the optimal treatment rule is the same within each subgroup. Specifically, we consider 10 latent groups and the true optimal treatment category of each group is, in turn, $A^* = 3, 3, 1, 2, 2, 1, 2, 3, 3, 1$. To generate data mimicking a three-arm randomized trial, for each subject, a 3-category treatment $A$, is randomly assigned with equal probability. The reward outcome is generated as $R = 4 \times I(A = A^*) - 1 + 0.5 \times N(0, 1)$. Furthermore, we imitate a common real world scenario where the treatment mechanism may not be known and thus the latent subgroup labels are not observed: instead of directly using group labels as observed data, we generate feature variables that are informative of the latent group membership as observed data. We simulate 30 feature variables from a multivariate normal distribution, where the first 10 variables $X_1, X_2, ..., X_{10}$ have a pairwise correlation of 0.8, the remaining 20 variables are uncorrelated, and the variance for each variable is 1. Moreover, $X_1, X_2, ..., X_{10}$ have mean values of $\mu_l$ for the latent group $l$, which are generated from $N(0, 5)$, while the means of $X_{11}, ..., X_{30}$ are all 0s. Therefore, only $X_1, X_2, ..., X_{10}$ are informative of the group labels due to different $\mu_l$. The observed data for each subject consist of the treatment assignment $A$, the feature variables $X_1, ..., X_{30}$, and the reward outcome $R$.

For each simulated data, we apply SOM learning to estimate the optimal ITR. At each step, we fit a weighted SVM with a linear kernel by solving the corresponding dual problem via quadratic programming. The tuning parameter is chosen using cross-validation. Furthermore, we compare SOM learning with regression-based Q-learning, one-vs-all (OVA) and one-vs-one (OVO) based on the value function (reward) of the estimated optimal treatment rules. Q-learning is obtained by fitting a linear model, regressing $R$ on $X$, $A$ and their interactions, in which $A$ is replaced by dummy variables created for each category of $A$. For OVA and OVO, to make sure the rewards are positive, we use the absolute value of rewards as weights, and for each binary step, new labels are calculated by multiplying the original labels with signs of reward. This approach is extracted from our SOM learning. Then the estimated value functions are obtained from weighted binary SVM models. Because both OVO and OVA algorithms break multi-treatment problems down to binary ones, the main drawback is that if a subject is assigned with more than one treatment through different binary comparisons, the final assignment will be the one with the smallest label value. For each setting, we compare the four methods with different sample sizes: $n$ =300, 600, and 900.

Figure 3.1 to 3.4 present the results of the optimal treatment mis-allocation rates and the estimated value functions from 100 replicates and difference sample sizes, which are computed in an independently generated test data of size 3 million. Furthermore, Table 3.1 to 3.4 summarize the average of the marginal mis-allocation rates of each category.

In the first setting, we observe that Q-learning gains higher values and lower mis-allocation rates of the optimal ITR compared to SOM under all sample sizes because the regression model used in Q-learning is correctly specified. The estimated values of SOM learning become closer to those of Q-learning as the sample size increases. In the latter three non-linear settings, the regression model in Q-learning is misspecified, so it performs worse under all sample sizes. Instead, SOM learning outperforms all comparators including OVA and OVO in all the simulation settings. For SOM learning, we also used Gaussian kernel in our method and found negligible difference from using linear kernel. However, since computation using the former is much more intensive, we recommend to use linear kernel in practice.

Table 3.1: Category Mis-allocation Rates (%) of Setting 1

| Cat | $n$=300 | | | | $n$=600 | | | | $n$=900 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO |
| 1 | 19.6 | 10.9 | 41.3 | 23.3 | 14.6 | 7.4 | 40.1 | 17.8 | 12.7 | 6.0 | 39.6 | 15.3 |
| 2 | 12.6 | 5.4 | 19.3 | 14.5 | 10.5 | 3.6 | 16.8 | 11.6 | 9.8 | 2.9 | 16.0 | 10.5 |
| 3 | 23.2 | 10.7 | 31.7 | 25.3 | 16.5 | 7.2 | 30.6 | 18.9 | 14.4 | 5.9 | 29.4 | 16.0 |

Table 3.2: Category Mis-allocation Rates (%) of Setting 2

| Cat | $n$=300 | | | | $n$=600 | | | | $n$=900 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO |
| 1 | 23.8 | 27.7 | 41.4 | 31.2 | 20.9 | 27.1 | 42.1 | 31.0 | 19.0 | 26.7 | 40.4 | 30.3 |
| 2 | 34.4 | 38.8 | 39.3 | 39.8 | 30.7 | 38.8 | 38.9 | 39.4 | 29.0 | 38.5 | 38.9 | 39.0 |
| 3 | 23.1 | 19.0 | 22.4 | 21.1 | 21.3 | 17.0 | 20.7 | 19.5 | 20.3 | 16.3 | 18.7 | 18.3 |

## 3.5 Application to REVAMP Study

We evaluate the performance of various methods when applied to the motivating REVAMP trial (Kocsis et al., 2009). The study aimed to evaluate the efficacy of adjunctive psychotherapy in the

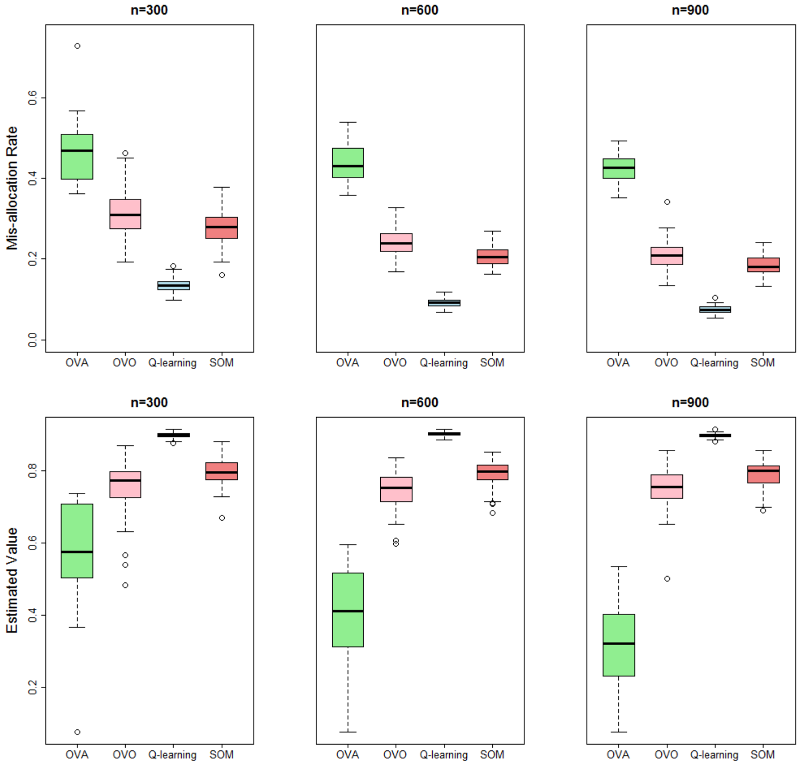


Figure 3.1: Box plots of the optimal treatment mis-allocation rates and estimated value functions of SOM, Q-learning, OVA and OVO for setting 1 with sample size of 300, 600 and 900. The optimal value is 0.9245.

treatment of patients with chronic depression who have failed to fully respond to initial treatment with an antidepressant medication. Among 808 participants in phase I, 491 were nonresponders or partial responders and entered phase II. At phase II, these 491 participants were then randomized to receive (1) continued pharmacotherapy and augmentation with brief supportive psychotherapy (MEDS+BSP), (2) continued pharmacotherapy and augmentation with cognitive behavioral analysis system of psychotherapy (MEDS+CBASP), or (3) continued pharmacotherapy (MEDS) alone, and were followed for 12 weeks. The primary outcome was the Hamilton Scale for Depression (HAM-D) scores at the end of 12-week follow-up. There were 17 baseline feature variables including participants' demographics, patient's expectation of treatment efficacy, social adjustment scale, mood and anxiety symptoms, and depression experience, as well as phase I depressive symptom measures such as rate of change in HAM-D score over phase I, HAM-D score at the end of phase I, rate of change of Quick Inventory of Depression Symptoms (QIDS) scores during phase I, and QIDS at the end of phase I.

After excluding participants with missing data, the final analysis consists of 318 participants,

Table 3.3: Category Mis-allocation Rates (%) of Setting 3

| Cat | $n$=300 | | | | $n$=600 | | | | $n$=900 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO |
| 1 | 28.1 | 31.0 | 35.1 | 35.8 | 25.4 | 27.7 | 34.6 | 31.7 | 23.7 | 25.7 | 36.6 | 29.9 |
| 2 | 33.0 | 39.1 | 36.7 | 40.0 | 31.5 | 38.2 | 36.9 | 38.6 | 30.3 | 37.5 | 36.2 | 37.6 |
| 3 | 38.5 | 38.8 | 37.2 | 40.7 | 36.6 | 36.6 | 37.0 | 38.7 | 36.0 | 36.1 | 36.9 | 39.5 |

Table 3.4: Category Mis-allocation Rates (%) of Setting 4

| Cat | $n$=300 | | | | $n$=600 | | | | $n$=900 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO | SOM | Qlearn | OVA | OVO |
| 1 | 14.7 | 35.7 | 43.3 | 30.6 | 13.1 | 33.0 | 43.6 | 30.2 | 13.2 | 31.2 | 44.3 | 28.6 |
| 2 | 27.5 | 37.5 | 30.9 | 32.0 | 25.7 | 34.8 | 29.4 | 31.2 | 25.1 | 33.4 | 29.0 | 30.8 |
| 3 | 33.9 | 37.1 | 33.4 | 32.0 | 32.6 | 34.8 | 32.8 | 31.8 | 31.9 | 34.0 | 32.3 | 31.4 |

among whom 134, 123, and 61 were assigned in MEDS+BSP, MEDS+CBASP, and MEDS, respectively. The mean HAM-D at the end of phase II study in each treatment arm under the designed non-personalized, random treatment assignment is summarized in Table 3.5. MEDS+CBASP group has the lowest post-treatment HAM-D score, but there is no statistically significant differences in the changes of HAM-D scores during phase II detected among the 3 treatment groups (Kocsis et al., 2009).

Table 3.5: Mean (standard error) of the HAM-D under randomized treatment assignment and value function of ITR from 2-fold cross-validation procedure with 500 repetitions

| Treatment arm† | MEDS+BSP | MEDS+CBASP | MEDS | |
|---|---|---|---|---|
| Value in test sample* | 13.51 (0.05) | 10.75 (0.05) | 12.66 (0.13) | |
| Method for ITR | SOM learning | Q-learning | OVA | OVO |
| Value in test sample* | 8.91 (2.91) | 16.16 (2.64) | 12.18 (1.67) | 10.85 (2.24) |

†: Treatment arm under the designed non-personalized, random assignment rule. *: Value function is the average HAM-D score at end of phase II for patients following an estimated optimal treatment (a smaller HAM-D score indicates a better outcome).

Our analysis goal is to use 17 feature variables to estimate the optimal ITR among three different options so that the value function (average HAM-D scores) under the ITR can be as low as possible. All feature variables are standardized before the analyses. We apply SOM learning and compare with Q-learning that uses $(1, X, A, XA)$ in the regression model, where $X$ represents feature variables and $A$ is the randomized treatment assignments, as well as OVA and OVO. The expected HAM-D for an ITR is calculated from 2-fold cross-validation of the data with 500 replicates: at each replicate, we

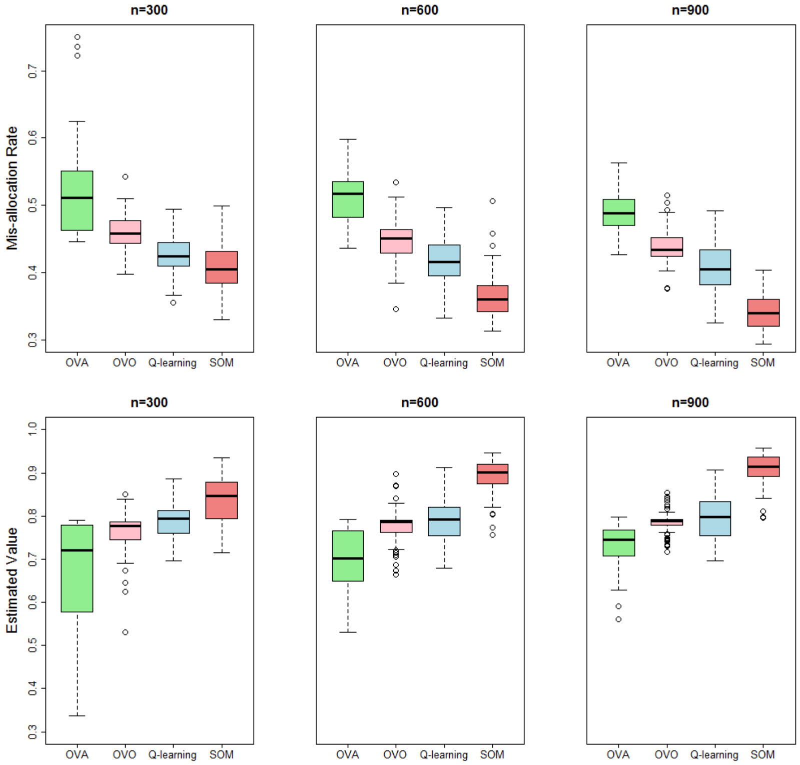


Figure 3.2: Box plots of the optimal treatment mis-allocation rates and estimated value functions of SOM, Q-learning, OVA and OVO for setting 1 with sample size of 300, 600 and 900. The optimal value of setting 2 is 1.0585.

randomly split the data into a training sample and a testing sample; we then apply SOM learning to learn the optimal ITR using the training data and compute the expected value in the testing sample under this estimated rule. The averages of the cross-validated value functions from the four methods are presented in Table 3.5 and their distributions over cross-validations are plotted in Figure 3.5. With a value function of 8.91, the SOM learning achieves a lower HAM-D compared to Q-learning, OVA, OVO, and any of the non-personalized, random assignment rules.

There are 99, 103, and 116 patients predicted to have MEDS+BSP, MEDS+CBASP, and MEDS alone as the optimal treatment, respectively. Table 3.6 presents the coefficients of the $3! - 1 = 5$ models derived from SOM learning rule in REVAMP study. Model 1 and model 2 corresponds to the 2 permutations of inner loop, determining whether a subject should be optimally assigned to MEDS only group or not. After eliminating the possibility of being assigned to MEDS only group, model 3 assigns a subject into MEDS+BSP or MEDS+CBASP treatment group. Let $\widehat{\beta}_{11}, \widehat{\beta}_{12}, \widehat{\beta}_{21}, \widehat{\beta}_{22}, \widehat{\beta}_3$ be the estimated coefficients of model 1(1), 1(2), 2(1), 2(2) and 3, respectively. A patient will be assigned

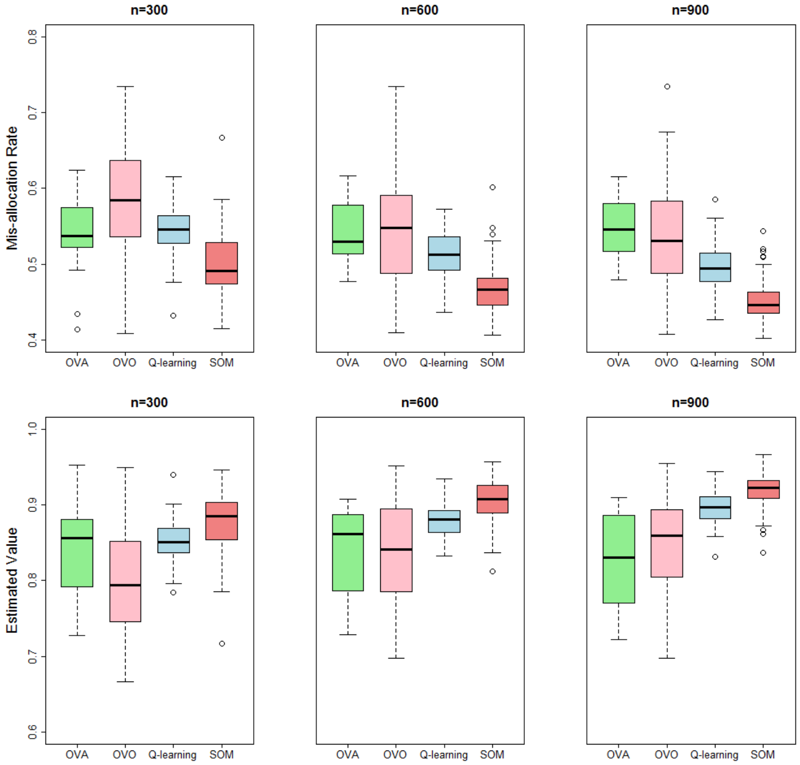


Figure 3.3: Box plots of the optimal treatment mis-allocation rates and estimated value functions of SOM, Q-learning, OVA and OVO for setting 1 with sample size of 300, 600 and 900. The optimal value of setting 3 is 1.1438.

with: MEDS if he has $\{X^T\widehat{\beta}_{11} < 0, X^T\widehat{\beta}_{12} < 0\}$, or $\{X^T\widehat{\beta}_{21} < 0, X^T\widehat{\beta}_{22} < 0\}$; MEDS+CBASP if he has not been assigned to MEDS and $X^T\widehat{\beta}_3 < 0$; MEDS+BSP if he has not been assigned to MEDS and $X^T\widehat{\beta}_3 > 0$. The column "Norm" reports the overall effect of feature variables on the optimal treatment decision rule as the $L_2$-norm of all coefficients for predicting each model.

The overall most predictive variable as determined by the norm of the coefficients in estimating the optimal ITR is phase I QISD rate of change, followed by phase I HAM-D rate of change. Both variables are most predictive of patients with MEDS alone as the optimal choice compared to two other combined pharmacotherapy and psychotherapy. Gender, response at phase I, patients expectancy of treatment efficacy, and CBASP expectation are also informative with an overall effect size greater than 0.7. Gender is also most predictive of MEDS alone versus two combined therapies alternatives with females favoring the latter. Other predictive variables include history of drug abuse and current alcohol use. No feature variable has a substantially large effect in model 3, implies that potentially many variables are in play to distinguish MEDS only versus the other two

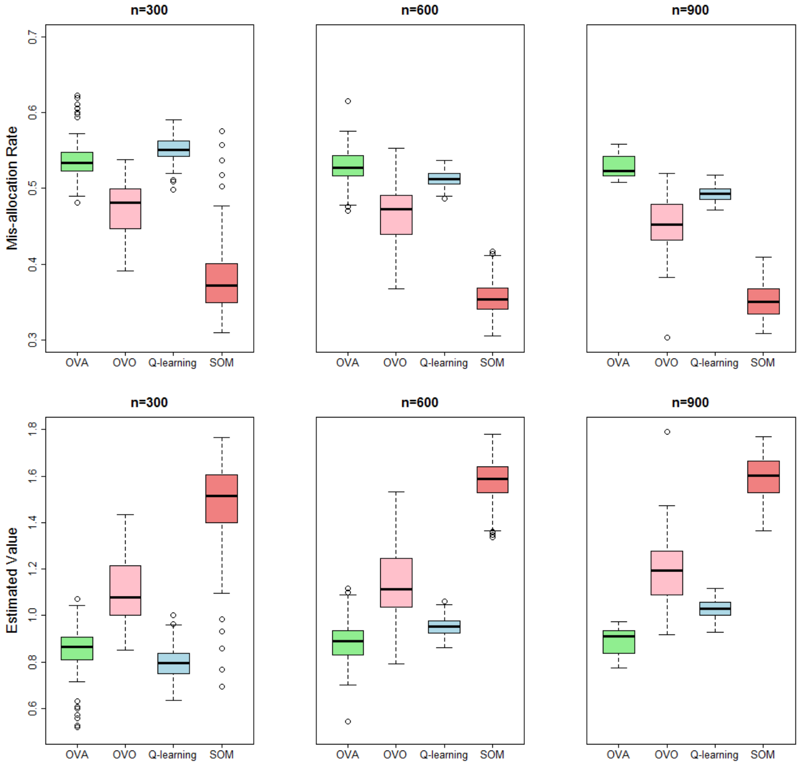


Figure 3.4: Box plots of the optimal treatment mis-allocation rates and estimated value functions of SOM, Q-learning, OVA and OVO for setting 1 with sample size of 300, 600 and 900. The optimal value of setting 4 is 2.9999.

combined therapies. In a recent analysis ofanother randomized trial on major depressive disorder comparing Nefazodone, CBASP or the combination of the two treatments, obsessive compulsive and past history of alcohol dependence (Gunter et al., 2011), race, and education level (Klein et al., 2011) are identified as predictive by Q-learning, which partially corroborates our findings. Our analyses identify several additional feature variables as informative.

To further visualize the relationship between feature variables and the optimal treatment for each individual, in Figure 3.6, we present the heatmap of 17 standardized feature variables by predicted optimal treatment on all subjects. We can see that history of drug abuse has a different pattern between patients with MEDS+BSP as optimal choice and the other two groups (more prevalent in the former versus the latter two groups), and thus may be informative of distinguishing MEDS+BSP versus others; dysfunctional attitudes, and patient's treatment efficacy expectation, frequency of sides effects, HAM-D rate of change during phase I, QIDS and HAM-D end of phase I score are informative for distinguishing all three treatments. It is clear that no single variable has a dominating

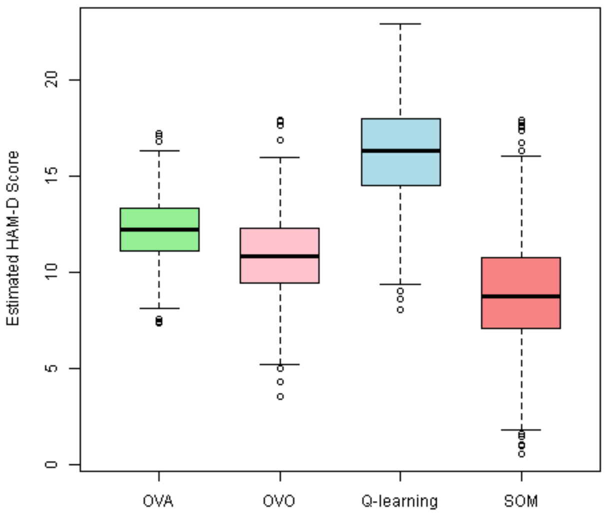


Figure 3.5: Box plot of the value function for the optimal ITR estimated by various methods from 2-fold cross-validation with 500 repetitions using REVAMP data: HAM-D score after phase II treatment (a smaller score indicates a better outcome).

effect on estimating the optimal ITR, and all feature variables in combination may be more effective.

## 3.6 Conclusions

We have proposed a sequential outcome weighted learning, SOM learning, to estimate the optimal ITRs with multicategory treatment studies, where each step solves a weighted binary classification problem via SVMs. By carefully choosing weights in each SVM step and combining the treatment decision functions from all steps, we showed that the derived rule from the proposed learning algorithm is Fisher consistent. In contrast to other Fisher consistent mutli-category learning algorithms (e.g., Liu and Shen, 2006), our method does not require non-convex optimization. In both numeric simulations and data application, SOM learning yields a more desirable value function as compared to the method based on a standard regression model or other straightforward methods such as one-vs-one and one-vs-all.

SOM learning can be extended in several directions. First, for some chronic diseases with multi-stage therapy, dynamic treatment regimes (DTRs) can be more powerful in obtaining favorable outcomes than a simple combination of single-stage treatment rules. Various approaches have been developed to estimate optimal DTR, such as Murphy (2003); Robins (2004); Moodie et al. (2007);

Table 3.6: Coefficients estimated from SOM learning in REVAMP study (ranked by the overall effect of a feature variable).

| Feature Variable | Model 1(1) | Model 1(2) | Model 2(1) | Model 2(2) | Model 3 | Norm* |
|---|---|---|---|---|---|---|
| QISD phase I change | -0.1281 | 0.0765 | 0.1200 | 2.2308 | -0.1338 | 2.2430 |
| HAM-D phase I change | -0.0063 | 0.0050 | 0.0285 | 1.7769 | -0.0252 | 1.7773 |
| Gender (Male) | -0.2726 | -0.0565 | -0.0880 | -1.0370 | 0.0753 | 1.0800 |
| Phase I response | -0.2269 | -0.0147 | 0.0166 | -0.7567 | -0.0382 | 0.7913 |
| Tx efficacy expectation | 0.4010 | 0.1026 | 0.1232 | 0.5791 | 0.0268 | 0.7229 |
| CBASP expectation | -0.2767 | -0.0153 | 0.0093 | -0.6371 | -0.1444 | 0.7097 |
| History of drug abuse | 0.3187 | 0.0199 | 0.0017 | 0.5644 | 0.0648 | 0.6517 |
| Current alcohol use | 0.3159 | 0.0100 | -0.0801 | 0.0877 | 0.2214 | 0.4038 |
| Social adjustment | 0.0813 | 0.0481 | 0.1075 | -0.3202 | -0.0207 | 0.3513 |
| BSP expectation | 0.1498 | -0.0349 | -0.0542 | 0.2703 | 0.1087 | 0.3338 |
| Freq of side effects | -0.0189 | 0.1816 | 0.1394 | 0.0909 | -0.1199 | 0.2746 |
| QISD end of phase I | 0.1999 | -0.0216 | -0.0796 | 0.1259 | 0.1005 | 0.2697 |
| Anxious Arousal | 0.0957 | 0.1316 | 0.1227 | 0.0850 | -0.0069 | 0.2209 |
| General Distress Anxious | -0.0975 | -0.0900 | -0.0954 | -0.0844 | -0.0085 | 0.1842 |
| HAMD end of phase I | -0.0449 | -0.0101 | 0.0108 | -0.0798 | -0.0520 | 0.1063 |
| Dysfunctional Attitudes | -0.0099 | 0.0041 | 0.0058 | -0.0108 | -0.0049 | 0.0170 |
| Age | 0.0017 | -0.0001 | -0.0009 | 0.0011 | 0.0018 | 0.0029 |

*: “Norm" measures the overall effect of a variable on the optimal treatment assignment rule as the $L_2$ norm of all coefficients for predicting each model.

Zhao et al. (2011); Zhang et al. (2012, 2013); Liu et al. (2014). While our method has focused on single-stage studies only, the proposed procedure can be easily generalized to handle multicategory DTR for multiple stage trials. Second, although the proposed method was only applied to a finite number of categories, it can be naturally extended to find optimal personalized dose, where treatment is in a continuous scale, after discretizing the dosage into categories. However, one challenge is to determine the number of the categories and the threshold of discretization. A possibility is to include these uncertainties as parameters to by estimated in SOM learning. Further research is worth pursuing.

A major computational cost for SOM learning is to screen all possible permutations of the treatment categories. Since the sequential learning for each permutation can be carried out independently of one another, an improvement in implementation is to incorporate distributed computing to leverage this natural parallel computing structure.

Finally, although we suggested to treat the most prevalent treatment as the first target optimal treatment in SOM learning procedure, this may result in few cases for later treatments in consideration and cause a large mis-allocation rate for patients whose optimal treatments are less prevalent. In practice, when different treatments have different importance, for instance, due to the need to

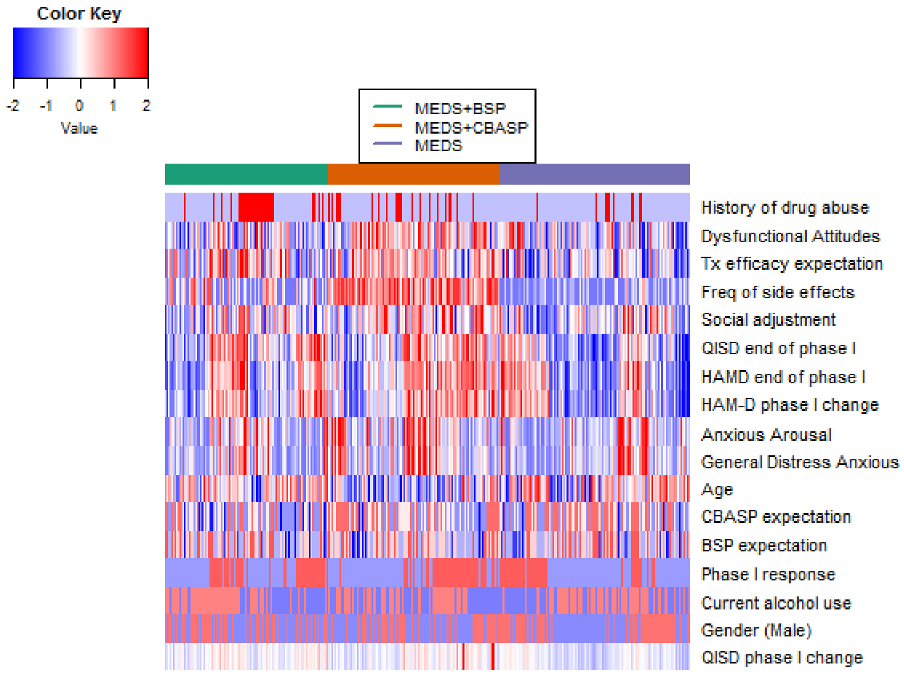


Figure 3.6: Heatmap of 17 standardized feature variables on all patients grouped by predicted optimal treatment. Row corresponds to feature variable and column corresponds to patients stratified by predicted optimal treatment.

balance efficacy and risk, the order of the targeted treatments should take into account the practical importance.

# CHAPTER 4: MULTICATEGORY LEARNING FOR ESTIMATING OPTIMAL ITR FROM ELECTRONIC HEALTH RECORD DATA

## 4.1 Introduction

### 4.1.1 Adverse drug reaction and EHR data

Adverse drug reactions (ADR) due to drug side effects and drug-drug interactions has become an increasing burden and threat to global public health. A recent study showed that in the United Kingdom, ADRs account for 1 in 16 of all hospital admissions (Pirmohamed et al., 2004). Moreover, drug-drug interactions particularly elevate the risk of ADRs. Although many of the implicated drugs have proved benefit, measures need to be put into place to reduce the burden of ADRs and thereby further improve the benefit:harm ratio of the drugs. It has been well documented that, in some cases, a treatment that is effective in the overall population may be highly ineffective in a subgroup of patients with specific characteristics, or no longer beneficial after patients develop resistance (Trivedi Madhukar et al., 2008; Lipska and Krumholz, 2014). In recent years, there is a growing interest in understanding treatment heterogeneity and discovering individualized treatment rules (ITRs) to maximize efficacy and achieve personalized medicine (Kosorok and Moodie, 2015). Particularly, individualized treatment strategy aims to recommend optimal treatment decision for an individual patient using information on a combination of his or her characteristics such as genomic features, medical history etc. Given patients' characteristics, searching for their optimal treatments could be one of the solutions to reduce the risk of having ADRs during or after drug use.

To estimate the optimal ITRs and handle high-dimensional patient-level feature variables, numerous machine learning methods are proposed. One popular method is Q-learning (Watkins, 1989; Murphy, 2005; Qian and Murphy, 2011), which relies on regression models to incorporate the interactions between treatment and covariates. However, because Q-learning relies heavily on the correctness of models, it may perform poorly when the model is misspecified. Another frequently used approach is outcome weighted learning (O-learning) proposed by Zhao et al. (2012). O-learning is an algorithm based on weighted support vector machines, to choose optimal ITRs by drawing a connection to a classification problem and directly optimizing the expected clinical outcome under the rule. We propose sequential outcome-weighted multicategory learning (SOM-learning) in Chapter 3, which generalizes O-learning to settings with more than two treatment options, and also allows negative outcome values. We also provide the convergence rate of SOM-learning and prove the Fisher consistency of the estimated ITR.

Electronic health records (EHR) provide rich resources to enable understanding of various clinical problems in real world setting including adverse drug reactions. Because of a greater availability at lower costs and technological advances that made computational processing on large-scale data more feasible, EHR databases have been actively used in pharmacoepidemiology especially for the past decade (Madigan et al., 2014). Relative to individual case safety reports, EHR data cover extended parts of the underlying medical histories, include more complete information on potential risk factors, and are not restricted to patients who have experienced a suspected ADR. Patient records contain a wide variety of time-stamped events that may serve as the basis for temporal pattern discovery: drug prescriptions, laboratory test results, hospital referrals and admissions, and notes of clinical symptoms, signs, and diagnoses (Norén et al., 2010). Long-term longitudinal capture of data in these sources can also enable studies that monitor the performance of risk management programs or other interventions over time (Weatherby et al., 2002).

While EHR databases provide valuable and unique opportunities for researchers to study the risk of a large number of drugs on ADRs simultaneously, there exist limitations such as data can be incomplete, or sometimes be artificially manipulated to serve clinical care. The principal challenge of analyzing EHR data is the potential bias in observational studies such as confounding and selection bias. Specifically, in the context of drug safety analyses, one of the most challenging issues is confounding by indication, i.e., a situation in which the indication for the medical product is also an independent risk factor for the outcome. Thus, a medical product can spuriously appear to be associated with the outcome when no appropriate control for the underlying condition exists, and confounding may persist despite advanced methods for adjustment (Walker, 1996; Bosco et al., 2010). Therefore, powerful statistical methods are required to better take the advantage of the richness of EHR data as well as overcome the downsides. Powerful statistical learning methods are able to deal with large scaled EHR data and provide robust results. However, there are very few literatures adapting statistical learning techniques to estimate optimal ITRs in EHR data.

### 4.1.2 Statin-induced myopathy

The Indiana Network for Patient Care (INPC) includes information from five major hospital systems (fifteen separate hospitals), the county and state public health departments, and Indiana Medicaid and RxHub, containing health records for over 15 millions patients (Biondich and Grannis, 2004; McDonald et al., 2005). The Common Data Model (CDM) data, which is a derivation of the INPC for this work, is to standardize the format and content of the observational data, so standardized applications, tools and methods can be applied to them. The CDM of the INPC contains over 60 million drug dispensing events, 140 million patient diagnoses, and 360 million clinical observations such as laboratory results, diagnose codes, and prescription medications (Zhang et al., 2015).

Statins, also known as HMG-CoA reductase inhibitors have been found to reduce cardiovascular disease (CVD) and mortality in those who are at high risk (Taylor et al., 2011). They are currently the most effective and widely prescribed drugs available for the reduction of low-density lipoprotein cholesterol, a critical therapeutic target for primary and secondary prevention of CVD (Chatzizisis et al., 2010). There are currently seven types of statins approved by the U.S. Food and Drug Administration (FDA), which include: Atorvastatin, Fluvastatin, Lovastatin, Pitavastatin, Pravastatin, Rosuvastatin, and Simvastatin (Sweetman and Blake, 2011).

Side effects of statins include muscle pain, increased risk of diabetes mellitus, and abnormalities in liver enzyme tests, among which statin-induced myopathy is a very common adverse effect. Myopathy is a muscle disease which results in muscular weakness. In observational studies $10-15\%$ of people who take statins experience muscle problems, in most cases these consist of muscle pain (Abd and Jacobson, 2011). Besides, statin-induced myopathy is likely to induce severe and potentially fatal cases of rhabdomyolysis, a rapid destruction of skeleton muscle (Chatzizisis et al., 2010). It is reported that Lovastatin and Pravastatin generally causes fewer side effects than others, high dose Simvastatin may be more likely to cause muscle pain, and Rosuvastatin has the highest rates of reported side effects. Historical data have shown that there are some certain factors to have side effects from taking statin, such as a patient is taking more than one medication to reduce cholesterol, is a female, is 65 years or older, or has kidney or liver disease. In this work, we consider adverse reactions of both myopathy and rhabdomyolysis and aim to discover potential risk factors based on available EHR data.

In this chapter, we apply SOM-learning to estimate optimal ITRs in EHR data collected from the Indiana Network for Patient Care (INPC) database. The data contain basic information and medical history of patients who have ever exposed to statin drugs. Assuming that all possible statin drugs have equivalent efficacy, we aim to recommend personalized treatment option to each patient such that their risks of having myopathy and rhabdomyolysis can be as low as possible, given potential confounding factors. We first create a large set of relevant features from the medical information provided in EHR data, and then estimate propensity scores using regularized multinomial regression. To assure an more accurate result as well as computational capacity, we conduct repeated stratified sampling and draw conclusions by combining estimated ITRs from all subsets. Propensity score along with sampling proportion will be adjusted in the SOM-learning procedure.

The rest of this chapter is structured as follows. Section 2 introduces the main idea and the mathematical framework for multicategory ITRs using SOM learning, and preparation details of data. In Section 3, we provide results of ITR estimation using SOM-learning and compare them with the ones using Q-learning. Conclusions and possible future works are discussed in Section 4.

## 4.2 Methodology

### 4.2.1 SOM-learning

Suppose the data contain $n$ patients and $k$ different treatment options. For each patient $i$, we observe a $d$-dimensional vector of feature variables, denoted by $X_i \in \mathcal{X}$, a treatment assignment $A_i \in \mathcal{A} = \{1, 2, ..., k\}$, $i = 1, ..., n$, and the clinical outcome after treatment denoted by $R_i$ (also referred as the "reward"), with larger values of $R_i$ being more desirable. A multicategory ITR, denoted by $\mathcal{D}$, is a mapping from the space of feature variables, $\mathcal{X}$, to the domain of treatments, $\mathcal{A}$. An optimal ITR is a treatment assignment rule that maximizes the mean reward $E[R(\mathcal{D}(X))|X]$, where $R(a)$ is the potential outcome had treatment $a$ been given. According to Qian and Murphy (2011), for randomized trials and assuming consistency of the potential outcomes, the optimal ITR maximizes the following value function:

$$E\left[\frac{I\{A = \mathcal{D}(X)\}}{\pi_A(X)} R\right], \tag{4.1}$$

where $\pi_a(x) = pr(A = a|X = x)$ is the randomization probability for treatment $a$, $a = 1, ..., k,$, assumed to be bounded by a positive constant from below, so $\sum_{a=1}^{k} \pi_a(x) = 1$. The goal is to learn the optimal treatment rule using empirical observations $(R_i, A_i, X_i), i = 1, ..., n$, and an optimal ITR is

$$\mathcal{D}^*(x) = \text{argmax}_a E[R|A = a, X = x].$$

One approach to estimate the optimal ITR is Q-learning, which builds a regression model to estimate the conditional means on the right-hand side. However, it heavily relies on the correctness of the postulated model, and model misspecification can lead to substantially non-optimal ITR even for a binary treatment situation (Zhao et al., 2012). Alternatively, O-learning proposed by Zhao et al. (2012) directly maximized the empirical version of the value function but replaced $I(A = \mathcal{D}(x))$ by $1 - \max(0, 1 - Af(x))$, where $f(x)$ is the decision function such that $\mathcal{D}(x) = \text{sign}(f(x))$. The latter corresponds to a weighted support vector machine where the weight for each observation is proportional to $R_i$. They demonstrated that O-learning outperformed the regression model based method in small sample. However, the proposed method can only be applied to estimate binary treatment decisions, and thus not directly applicable when more than two treatment options are of interest.

SOM-learning generalizes O-learning to fitting multi-treatment settings and transforms multi-category ITR learning problem into a sequence of binary ones. The underlying idea is to perform a sequence of binary treatment decision learning, where each step in the sequence determines whether the optimal treatment for a patient should be a candidate treatment category or the other treatments. Thus O-learning with weighted support vector machines (Zhao et al., 2012) is adapted in each subsequent binary step with adjusted weights. The selection of weights in SOM-learning ensures that the final rules are Fisher consistency.

Unlike randomized trials, the probability of receiving a treatment, $\pi_A(X)$, in observational studies is not known at stage of study design anymore, but has to be estimated from the data. Because of the observational nature of EHR data, we derive informative covariates from the observed measurements to estimate the probabilities and adjust for confounding. Thus, the estimated probability (also known as “propensity score"), $\hat{\pi}_A(X)$ will replace $\pi_A(X)$ in the value function (4.1). Therefore, the value function in this case is:

$$E\left[\frac{I\{A = \mathcal{D}(X)\}}{\hat{\pi}_A(X)} R\right], \tag{4.2}$$

where $\hat{\pi}_A(X) = \hat{P}(A = a|X = x)$ denote the estimated propensity score.

## 4.3 Data preparation

The available medical history and occurrence time of adverse reaction in the data expand from year 2000 to 2016. Patients who experienced ADRs–myopathy or rhabdomyolysis from year 2012 to 2016 were selected as cases, and those didn't experience any ADRs during 2012 to 2016 served as controls. The ADR was then scored as 10, 1, and 0 for rhabdomyolysis, myopathy or control, respectively. The outcome value of interest is the number of ADR occurrences multiplies the score of ADR during year 2012 to 2016. Demographic variables include age at 2012 and gender. There are a total of 270,118 patients in the dataset.

The data contain information of more than 1000 drugs, among which we excluded very rarely used ones cutting at the 5th percentile of frequencies and 187 drugs left for further analysis. There are five different statin medicines among all the 187 drugs: Lovastatin, Pravastatin, Simvastatin, Atorvastatin, and Rosuvastatin. Based on the marketing popularity and potency ranking, we combine Lovastatin and Pravastatin into one group. Thus, there are 4 possible treatment options in this problem. Table 4.1 summarizes number of patients who have mostly exposed to each drug.

Table 4.1: Information of Statins

| Group | Statin Medicine | Number of Patients |
|---|---|---:|
| Statin 1 | Lovastatin & Pravastatin | 45,912 |
| Statin 2 | Simvastatin | 196,657 |
| Statin 3 | Atorvastatin | 17,604 |
| Statin 4 | Rosuvastatin | 9,945 |
| Total | | 270,118 |

For the remaining 182 drugs, we computed the duration and frequency for each drug use and adjusted them by the length of patients' drug record history during 2000-2012. We also included the rate of ADR occurrence. Therefore, for each patient, 364 drug related variables, in addition with age at 2012, gender, and previous rate of ADRs were included in further analyses.

Regularized multinomial regression was carried out to estimate propensity scores for each patient with outcome being the 4 statin treatment options. The optimal tuning parameter for LASSO penalty in the regression model is chosen by cross validation. The model also selected 322 out of the 367 feature variables, among which age, gender and previous case rate were all kept.

### 4.3.1 Clustering of variables

Due to the sparsity of feature variables, we conducted clustering to group them into homogeneous clusters and reduce dimension using R package ClustOfVar (Chavent et al., 2011). This algorithm is to find a partition of a set of variables such that the variables within a cluster are related to each other as strongly as possible.

According to Chavent et al. (2011), a synthetic variable $c_k$ of a cluster $C_k$ is defined as the variable “most linked" to all variables in $C_k$ based on their Pearson correlation. Let $\mathbf{X}_k$ denote the matrix of $C_k$, extracting from our data matrix $\mathbf{X}$. Specifically, $c_k$ is the first principal component of $\mathbf{X}_k$. Finally, 20 synthetic variables from 20 clusters, age and gender were selected to learn optimal ITRs.

### 4.3.2 Stratified sampling and SOM-learning

One main concern raised by the richness of EHR data is that it requires high computational capacity. To lighten the burden of computation for SOM learning and make sure there are enough ADRs being selected, we conducted stratified simple random sampling. There were 14 strata which were formed based on patients categorized age and gender. Because there are much more controls than cases, in each stratum we kept all cases and randomly selected the same number of controls. These subsets of the strata are then pooled to form a random sample.

Let $p_h(x)$ denote the probability of a patient being selected from its strata $h$, where $h = 1, ..., 14$. From our setting of stratified sampling, $p_h(x) = 1$ for cases and $p_h(x) = n_h(\text{case})/n_h(\text{control})$ for controls, where $n_h(\text{case}), n_h(\text{control})$ denote the original number of cases and controls in the stratum. The sample size of each stratum and the probability of a control being selected are summarized in Table 4.2.

Then we used the selected sample to train SOM-learning rules. Within each step of SVMs in

SOM-learning, the inverse of $p_h(x)$ were multiplied to the inverse of propensity scores as weights after being truncated at the 99th percentile. To control for sampling bias, we repeated stratified sampling and computed expected outcome values for the whole data using the rules learned from different samples.

Table 4.2: Sample sizes of each stratum

| Female, Age | $<= 40$ | $41-50$ | $51-60$ | $61-70$ | $71-80$ | $> 80$ |
|---|---|---|---|---|---|---|
| Number of cases | 1005 | 2438 | 4014 | 3297 | 1967 | 1593 |
| Pr(control) | 0.15 | 0.16 | 0.13 | 0.09 | 0.08 | 0.09 |
| **Male, Age** | $<= 40$ | $41-50$ | $51-60$ | $61-70$ | $71-80$ | $> 80$ |
| Number of cases | 346 | 981 | 1577 | 1309 | 855 | 504 |
| Pr(control)[1] | 0.05 | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 |

[1] The probability of a control being selected.

## 4.4 Analysis of confounding

In addition to the preparation steps listed above, we also did confounding analysis and model selection based on several parametric approaches.

A confounding variable is the one influences both treatment and outcome. To determine the potential confounder, we first computed the Spearman's correlations of the 322 variables with treatment indicators and outcome. Because of the large sample size, 0.05 is not a suitable threshold of significance. A correlation with absolute value below 0.01 can have a p-value$<$ .0001. Thus, we selected confounding variables for which the absolute value of the correlation with outcome and at least one of the treatment indicators was greater than 0.02. There were 33 variables met the criteria, which included age, gender, historical case rate and 30 drug related variables.

Next, we tried to select the variables that are predictable to outcome. Because of the fact that 92.6% of outcome value is 0, we applied a continuous odds ratio model with categorized outcome. The outcome $y$ was categorized at 4 thresholds: 1, 4, 8, and 16. We created three subsets from the original dataset which contained patients with outcome greater than 1, 4 and 8, respectively. The sample sizes of the subsets were 19813, 8556, and 5311. A new dichotomous response $v$ was defined based on the following rules. We stacked the four datasets and built a logistic regression model with $v$ as the dependent variable and the 33 variables as covariates and forward selection method was applied. Furthermore, drug information was dichotomized to used or not. After all, gender, case rate, and 5 medication use (Pregabalin, Acetaminophen, Loratadine, Acetylsalicylic acid, and Gabapentin) were selected. These drugs are all related to muscle weakness or pain. Last, we regress on the dichotomous response $v$ using the selected 7 variables and treatment as covariates for the original dataset.

Definition of response $v$:

Original dataset: $v = 1$ if $y >= 1, v = 0$ if $y = 0$;

Subset 1: $v = 1$ if $y >= 4, v = 0$ otherwise ;

Subset 2: $v = 1$ if $y >= 8, v = 0$ otherwise ;

Subset 3: $v = 1$ if $y >= 16, v = 0$ otherwise .

## 4.5 Results

### 4.5.1 Descriptive statistics

Among all the 270,118 patients, there were 250,234 (92.6%) controls, 18,701 (7.0%) were with myopathy cases and 1,183 (0.4%) were rhabdomyolysis cases. There are 144,819 (54%) females and 125,301 (46%) males. 224,587 (83%) patients had 0 case occurrence from 2000 to 2012. More descriptive statistics by statin groups are summarized in Table 4.3. Moreover, the association between value, case rate and statin groups are both statistically significant ($P < .0001$). Figure 4.1 shows the histograms of patients' age at 2012. From the histogram of estimated propensity scores (see Figure 4.2), we can tell that the probability of a patient being assigned to a statin medicine is not completely at random and most of the patients were assigned to their treatment with probability about 0.75.

Table 4.3: Descriptive statistics

| Variable | Statistic | Statin 1 | Statin 2 | Statin 3 | Statin 4 | Total |
|---|---|---|---|---|---|---|
| Outcome | mean (sd) | 0.6213 (5.0419) | 0.6524 (5.6157) | 0.9894 (6.5368) | 0.8179 (5.7632) | 0.6752 (5.5942) |
|  | (min, max) | (0, 320) | (0, 640) | (0, 470) | (0, 250) | (0, 640) |
| Case rate | mean (sd) | 0.0178 (0.1281) | 0.0204 (0.1427) | 0.0284 (0.0642) | 0.0223 (0.0564) | 0.0205 (0.1342) |
|  | (min, max) | (0, 22) | (0, 24) | (0, 0.825) | (0,0.812) | (0, 24) |
| Control | $n$ | 42,590 | 182,836 | 15,742 | 9,066 | 250,234 |
| Myopathy | $n$ | 3,141 | 12,974 | 1,766 | 820 | 18,701 |
| Rhabdomyolysis | $n$ | 181 | 847 | 96 | 59 | 1,183 |

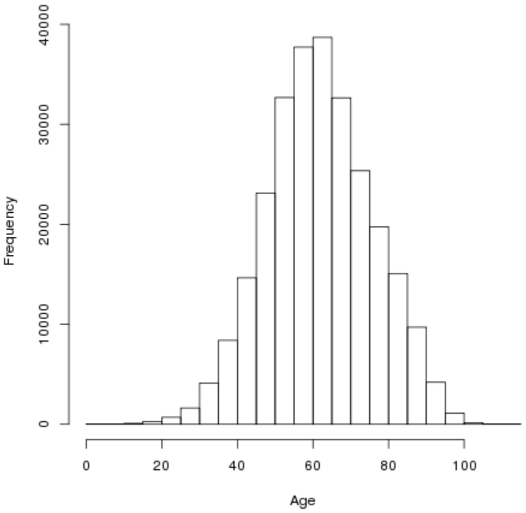


Figure 4.1: Histogram of age at 2012.

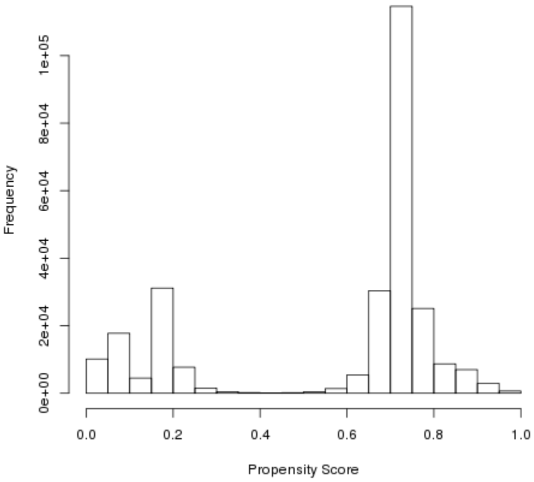


Figure 4.2: Histogram of estimated propensity scores.

### 4.5.2 Predicted ITR

To compare the results of SOM-learning, we also conducted random forest based Q-learning using age, gender, 20 synthetic variables from clustering, and their interactions with statin groups as feature variables. The original number of patients in each treatment group is 45912, 196657, 17604, and 9945, respectively. Table 4.4 gives the outcome value of “one-size-fits-all" method. Results of SOM-learning from 10 stratified samples are summarized in Table 4.5. Results of Q-learning using the same samples are provided in Table 4.6 for reference.

We see that Q-learning recommended Lovastatin or Pravastatin to majority of patients and gained a smaller predicted value compared to the “one-size-fits-all" method. However, since the training data is a subsample of test data, random forest based Q-learning may lead to overfitting. Thus, we are not able to conclude that the optimal ITRs obtained by random forest based Q-learning are superior than non-individualized treatment rules. The results for SOM-learning from different samples are fairly consistent. The estimated ITRs are more likely to allocate patients to Simvastatin and very few to Atorvastatin or Pravastatin. The predicted outcome value from SOM-learning is a little bit larger than the value of Simvastatin in “one-size-fits-all" rule.

Table 4.4: Outcome of “one-size-fits-all" (a smaller value indicates a better outcome)

| Statin 1 | Statin 2 | Statin 3 | Statin 4 |
|---|---|---|---|
| 0.6495 | 0.6570 | 0.8197 | 0.7470 |

Table 4.5: Summary of predicted labels and outcome values from different samples of SOM-learning (a smaller value indicates a better outcome)

| | Statin 1 | Statin 2 | Statin 3 | Statin 4 | Value |
|---|---|---|---|---|---|
| Sample 1 | 3857 | 265187 | 857 | 217 | 0.6580 |
| Sample 2 | 1174 | 268398 | 350 | 196 | 0.6633 |
| Sample 3 | 4145 | 264215 | 1556 | 202 | 0.6625 |
| Sample 4 | 11169 | 254231 | 4318 | 400 | 0.6657 |
| Sample 5 | 5963 | 261947 | 1267 | 941 | 0.6670 |
| Sample 6 | 1864 | 267648 | 359 | 247 | 0.6652 |
| Sample 7 | 1153 | 268468 | 277 | 220 | 0.6634 |
| Sample 8 | 5673 | 262824 | 622 | 999 | 0.6640 |
| Sample 9 | 8569 | 256798 | 4446 | 305 | 0.6607 |
| Sample 10 | 6040 | 262996 | 848 | 234 | 0.6652 |
| Std deviation | 3241.23 | 4745.35 | 1577.83 | 308.77 | 0.0026 |

Table 4.6: Summary of predicted labels and outcome values from different samples of Q-learning (a smaller value indicates a better outcome)

| | Statin 1 | Statin 2 | Statin 3 | Statin 4 | Value |
|---|---|---|---|---|---|
| Sample 1 | 171573 | 97892 | 440 | 213 | 0.2465 |
| Sample 2 | 175661 | 93928 | 405 | 124 | 0.2420 |
| Sample 3 | 176694 | 92768 | 476 | 180 | 0.2394 |
| Sample 4 | 176476 | 93036 | 470 | 136 | 0.2404 |
| Sample 5 | 177513 | 92010 | 349 | 246 | 0.2426 |
| Sample 6 | 174723 | 94731 | 300 | 364 | 0.2465 |
| Sample 7 | 172543 | 97169 | 288 | 118 | 0.2386 |
| Sample 8 | 173597 | 95962 | 399 | 160 | 0.2447 |
| Sample 9 | 173906 | 95726 | 350 | 136 | 0.2444 |
| Sample 10 | 173589 | 95868 | 446 | 215 | 0.2411 |
| Std deviation | 1927.20 | 1949.13 | 67.88 | 75.23 | 0.0028 |

### 4.5.3 Confounding analysis

The estimated odds ratio with 95% confidence interval (CI) of the logistic regression model introduced in section 4.4 are summarized in Table 4.7. We can see that Lovastatin or Pravastatin is significantly associated with a severer outcome compared to Simvastatin. Female is more likely to end up with myopathy or rhabdomyolysis after taking statin medicines.

Table 4.7: Summary of odds ratio estimates

| Variable | Odds Ratio | 95% CI | p-value |
|---|---|---|---|
| Female | 1.304 | (1.251, 1.358) | $< .0001$ |
| Statin 1 vs 2 | 1.092 | (1.037, 1.151) | 0.0008 |
| Statin 3 vs 2 | 0.994 | (0.931, 1.061) | 0.8564 |
| Statin 4 vs 2 | 1.010 | (0.921, 1.108) | 0.8248 |
| Drug 1[1] | 1.494 | (1.411, 1.581) | $< .0001$ |
| Durg 2 | 0.853 | (0.805, 0.904) | $< .0001$ |
| Drug 3 | 1.111 | (1.062, 1.163) | $< .0001$ |
| Drug 4 | 0.984 | (0.943, 1.027 | 0.4616 |
| Drug 5 | 1.151 | (1.104, 1.200) | $< .0001$ |

[1] Drug 1-5 are Pregabalin, Acetaminophen, Loratadine, Acetylsalicylic acid, and Gabapentin, respectively.

## 4.6 Discussion

We have applied the proposed SOM-learning method to estimate the optimal ITRs for EHR data by solving a multicategory treatment selection problem via sequential weighted support vector machines. We considered statin-induced myopathy and rhabdomyolysis as ADRs and tried to learn the optimal ITRs with minimum average scores of such ADRs. The treatment choice recommended by Q-learning was more desirable compared to the non-individualized “one-size-fits-all" rules. For SOM-learning, we applied sampling technique and inverse probability weighting scheme to assure the computational capability. However, due to the observational nature of this dataset and potential bias caused by sampling, the results of SOM-learning is not as expected. Therefore, one possible approach to improve the performance of SOM-learning is to develop more efficient function with parallel computing such that it is able to handle high-dimensional data as is in this case.

Machine learning methods are powerful in analyzing high-dimensional and sparse data from EHR databases as is showed in this work. Tailored by patients' characteristics and medical histories, the estimated ITRs are able to recommend treatment with lower risk of having ADRs induced by medicines. Because of the observational feature of EHR data, inverse probability weighting of propensity score was adopted to adjust for bias of observed features. However, unobserved confounding and bias may still exist thus require more attention for ITRs that are estimated from non-experimental studies.

One limitation of this work is the definition of outcome and features. Especially for the outcome, we took score 0, 1, and 10 for different levels of ADR severity. However, these scores may not be clinically meaningful, and because it's critical to results, a misleading definition may end up with incorrect estimation of ITRs. A more rational definition of the ADR outcome is needed. Another possibly problematic part is the estimation of propensity scores. As shown in Figure 4.2, there are two obvious clusters for the estimated propensity scores, which is an unusual situation. Directly adopting these propensity scores may lead to mis-estimated value functions. An alternative approach is to split all the patients into two groups based on their propensity scores and redo the analysis separately for the two groups. Because the estimated propensity scores are distributed roughly normally in each of the two clusters (see Figure 4.2), this approach may help to improve the final decision of optimal treatment rules.

While statin-induced myopathy has been shown to be dose dependent (Pasternak et al., 2002), one potential extension of our work is to make use of the information of drug dose and choose more accurate and effective treatment for patients. Other risk factors, for instance, history of disease such as hypertension, hypercholesterolemia and diabetes, alcohol abuse, ethnicity, and are worth exploring when available (Golomb and Evans, 2008). Additionally, on the basis of temporal information, we could search for a different baseline other than 2012 to make more practical sense and make fully use of the time information. Moreover, for diseases with multiple-stage therapy, powerful methodologies of estimating dynamic treatment regimes (DTRs) are explored to obtain favorable outcomes (Murphy, 2003; Robins, 2004; Moodie et al., 2007; Zhao et al., 2011; Zhang et al., 2012, 2013; Liu et al., 2014). In this work, we have focused on single-stage medical decision making, the proposed procedure can be easily generalized to handle multicategory DTRs for multiple-stage trials. Except for machine learning techniques we have focused on, parametric approaches can be helpful in analyzing this type of data. For example, logistic regression models with dichotomized or categorized outcome are useful to detect the prediction performance of interesting covariates and a model selection method can help with dimension reduction when there is a large amount of covariates.

# CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation focused on developing new and computationally efficient statistical learning methods and algorithms for multicategory classification and multicategory personalized medical decision making. We are motivated by two goals: create consistent rules for multicategory classification problems, and develop computationally efficient algorithms based on binary support vector machines (SVM). We first proposed forward-backward SVM method to consistently and efficiently learn multicategory rules for classification. The learning algorithm was then adapted to estimating multicategory individualized treatment rules via a connection with outcome weighted learning proposed by Zhao et al. (2012). In this work, we established theoretical properties for the proposed method such as Fisher consistency and risk bound of the derived learning rules. In addition, we showed that our methods performed promisingly in both extensive simulation studies and real-world data analysis, with low misclassification rates and significantly improved computational speed when compared to existing methods. Nevertheless, some potential improvements and extensions of this work merit future exploration.

First of all, while a major computational cost for the proposed learning algorithms is to screen all possible permutations of the categories, their computational performance can be immediately improved. Because the sequential learning for each permutation can be carried out independently of one another, we can incorporate distributed computing to leverage this natural parallel computing structure (Almasi and Gottlieb, 1988). This improvement is worth pursuing especially for the applications with a large number of categories. Another approach is to explore a more efficient and effective methodology for a data-driven choice of tuning parameter in the loss functions. These potential improvements could end up with a more sophisticated computational software or package.

Second, although our algorithm was originally developed based on SVMs, other binary classifiers can be easily implemented and applied alternatively. In each binary step, we can replace SVM with other large-margin classifiers, such as adaboost (Freund and Schapire, 1997), import vector machines (Zhu and Hastie, 2005), $\psi$-learning (Shen et al., 2003), and large-margin unified machines (Liu et al., 2011). The multicategory rules derived from the proposed algorithm will remain consistent as long as the input binary classifiers are consistent. Moreover, we compared the performance of our method only with other multicategory classification methods of SVMs. Indeed, it may be more critical if we expand the comparisons to other methods who deal with categorical outcomes, such as parametric models like multinomial regression, or other machine learning approaches like random forests (Ho, 1995, 1998), and neural networks (Specht, 1990; Khan et al., 2001) etc. Another possible improvement of our algorithm is to develop a build-in function to handle missing data when analyzing real data.

Third, we suggested to treat the most prevalent category as the first target in the sequences of learning procedure, however, this may result in few cases for later categories in consideration and cause a large misclassification rate for subjects whose optimal categories are less prevalent. In practice, when different categories have different importance, for instance, multiple treatment options with the need to balance efficacy and risk, the order of the targeted categories should take the practical importance into account. This arises another possible extension that we can broaden our method to resolve classification problems with ordinal categories.

Furthermore, for some chronic diseases with multi-stage therapy, dynamic treatment regimes (DTRs) can be more powerful in obtaining favorable outcomes than a simple combination of single-stage treatment rules. Various approaches have been developed to estimate optimal DTRs, such as Murphy (2003); Robins (2004); Moodie et al. (2007); Zhao et al. (2011); Zhang et al. (2012, 2013); Liu et al. (2014). While our method focused on single-stage studies only, the proposed procedure can be easily generalized to handle multicategory DTRs for multiple stage trials.

Finally, statistical analysis for observational data can be extremely challenging, especially for the large scale data collected from EHR database. Different from intervention studies, subjects in these databases are rarely assigned to randomized treatment groups. Instead, they receive treatments based on the known characteristics, medical history and onset symptoms. Propensity scores methods (Rosenbaum and Rubin, 1983) such as matching, stratification and inverse probability weighting are mostly widely used to adjust for potential confounding and bias when making inference of causality. More effective and accurate approaches for matching and weighting are desirable. Recently, in drug developing industry, there is an increasing interest of comparing different drugs from different data sources with various inclusive criteria. Thus, matching techniques are worth exploring not only for observational studies, but also for clinical trials. Besides, people could combine parametric methods and machine learning to seek for more reasonable and reliable results when analyzing big observational data from EHR databases.

# APPENDIX

## A.1 Proof of Theorem 2.1

We start from class label $k$ and follow the order in FB-SVM. First, we show $\mathcal{D}^*(\mathbf{X}) = k$ if and only if $P_k(\mathbf{X}) = \max_{h=1}^k P_h(\mathbf{X})$. For any $\mathbf{X}$ with $\mathcal{D}^*(\mathbf{X}) = k$, by the definition of $\mathcal{D}^*$, there exists a permutation $(j_1, ..., j_{k-1})$ of $\{1, ..., k-1\}$ such that $\mathcal{D}_l^{*(k)}(\mathbf{X}) = -1$ for $l = j_1, ..., j_{k-1}$. That is,

$$f_{j_1}^*(\mathbf{X}) < 0, f_{j_2}^*(\mathbf{X}) < 0, \ldots, f_{j_{k-1}}^*(\mathbf{X}) < 0. \tag{1}$$

On the other hand, from the estimation of $\widehat{f}_{j_1}$, it is clear that $f_{j_1}^*$ is the minimizer of the expectation of a weighted hinge loss corresponding to $V_{n,j1}$, which is given by

$$\begin{aligned} &E\left[\frac{k-1}{k} I(Y = j_1)[1 - f(\mathbf{X})]_+ + \frac{1}{k} I(Y \neq j_1)[1 + f(\mathbf{X})]_+\right] \\ =&E\left[\frac{k-1}{k} P_{j_1}(X)[1 - f(\mathbf{X})]_+ + \frac{1}{k}(1 - P_{j_1}(X))[1 + f(\mathbf{X})]_+\right]. \end{aligned}$$

Simple algebra following standard SVM theory gives

$$\text{sign}(f_{j_1}^*(\mathbf{X})) = \text{sign}\left(P_{j_1}(\mathbf{X})(k-1) - (1 - P_{j_1}(\mathbf{X}))\right) = \text{sign}\left(P_{j_1}(\mathbf{X}) - \frac{1}{k}\right).$$

That is, $f_{j_1}^*(\mathbf{X}) < 0$ is equivalent to $P_{j_1}(\mathbf{X}) < 1/k$. Now, in the next step, because we restrict to data with $Y_i \neq j_1$ and $f_{j_1}^*(X_i) < 0$, it is clear that $f_{j_2}^*$ minimizes

$$\begin{aligned} &E\Bigg[\frac{k-2}{k-1} I(Y = j_2)[1 - f(\mathbf{X})]_+ + \frac{1}{k-1} I(Y \neq j_2)[1 + f(\mathbf{X})]_+ \\ &\qquad\qquad \Bigg| Y \neq j_1, f_{j_1}^*(X) < 0\Bigg] \\ =&E\Bigg[\frac{1}{k-1}\frac{1}{1 - P_{j_1}(X)}\{P_{j_2}(X)(k-2)[1 - f(\mathbf{X})]_+ \\ &\quad +(1 - P_{j_1}(X) - P_{j_2}(X))[1 + f(\mathbf{X}_i)]_+\} \Bigg| P_{j_1}(X) < 1/k\Bigg]. \end{aligned}$$

Thus, we conclude that

$$\text{sign}(f_{j_2}^*(\mathbf{X})) = \text{sign}\left(P_{j_2}(\mathbf{X})(k-2) - (1 - P_{j_1}(\mathbf{X}) - P_{j_2}(\mathbf{X}))\right) \times I\{P_{j1}(\mathbf{X}) < 1/k\}.$$

That is, $f^*_{j2}(\mathbf{X}) < 0$ if and only if

$$\frac{P_{j_2}(\mathbf{X})}{1 - P_{j_1}(\mathbf{X})} < \frac{1}{k-1}.$$

We continue the same arguments and establish the relationship between $f^*_{j_l}$ and $P_{j_l}$ as

$$\text{sign}(f^*_{j_l}(\mathbf{X})) = \text{sign}\left(\frac{P_{j_l}(\mathbf{X})}{\sum_{h=l}^k P_{j_h}(\mathbf{X})} - \frac{1}{k-l+1}\right).$$

In other words, we obtain that for this subject with $f^*_{j_1}(\mathbf{X}) < 0, ..., f^*_{j_{k-1}}(\mathbf{X}) < 0$, it holds that

$$P_{j_1}(\mathbf{X}) < \frac{1}{k}, \quad \frac{P_{j_2}(\mathbf{X})}{1 - P_{j_1}(\mathbf{X})} < \frac{1}{k-1}, \cdots, \tag{2}$$

$$\frac{P_{j_{k-2}}(\mathbf{X})}{P_{j_{k-2}}(\mathbf{X}) + P_{j_{k-1}}(\mathbf{X}) + P_k(\mathbf{X})} < \frac{1}{3}, \quad \frac{P_{j_{k-1}}(\mathbf{X})}{P_{j_{k-1}}(\mathbf{X}) + P_k(\mathbf{X})} < \frac{1}{2}. \tag{3}$$

Starting from the last inequality, we have

$$P_{j_{k-1}}(\mathbf{X}) < P_k(\mathbf{X}), \quad P_{j_{k-2}}(\mathbf{X}) < \frac{1}{3}(P_{j_{k-2}}(\mathbf{X}) + P_{j_{k-1}}(\mathbf{X}) + P_{j_k}(\mathbf{X})),$$

$$\ldots, P_{j_2}(\mathbf{X}) < \frac{\sum_{h=2}^k P_{j_h}(\mathbf{X})}{k-1}, \quad P_{j_1}(\mathbf{X}) < \frac{1}{k}$$

so obtain

$$P_{j_{k-1}}(\mathbf{X}) < P_k(\mathbf{X}), \quad P_{j_{k-2}}(\mathbf{X}) < P_k(\mathbf{X}), \ldots, P_{j_2}(\mathbf{X}) < P_k(\mathbf{X}), \quad P_{j_1}(\mathbf{X}) < \frac{1}{k}.$$

Therefore, $P_k(\mathbf{X}) = \max_{l=1}^k P_l(\mathbf{X})$.

For the other direction, suppose that $P_k(\mathbf{X}) = \max_{l=1}^k P_l(\mathbf{X})$. We order $P_1(\mathbf{X}), ..., P_{k-1}(\mathbf{X})$ to obtain a sequence with $P_{j1}(\mathbf{X}) \leq P_{j2}(\mathbf{X}) \leq \cdots \leq P_{j(k-1)}(\mathbf{X}) \leq P_{jk}(\mathbf{X})$. We consider the corresponding classification rules for the same sequence. Because all inequalities in (4) and (5) hold, from the equivalence between $f^*_{j_l}$ and $P_{j_l}$, it is straightforward to see that

$$f^*_{j_1}(\mathbf{X}) < 0, ..., f^*_{j_{k-1}}(\mathbf{X}) < 0.$$

In other words, $\mathcal{D}^*(\mathbf{X}) = k$. Hence, we have proved that the FB-SVM rule is the Bayesian rule that correctly classifies subjects into class $k$ .

To prove the consistency of the remaining classes, FB-SVM obtains the rule for class $(k-1)$ conditional on $Y \neq k$ and $\mathcal{D}^*(X) \neq k$. Using the same proof as above, we conclude that

$$\mathcal{D}^*(\mathbf{X}) = (k-1) \text{ if and only if } (k-1) = \text{argmax}_{l=1}^{k-1} \widetilde{P}_{k-1,l}(\mathbf{X}),$$

where $\widetilde{P}_{k-1,l}(\mathbf{X})$ is the conditional probability of $Y = l$ given $X = \mathbf{X}$, $Y \neq k$, and $\mathcal{D}^*(X) \neq k$. Clearly, $\widetilde{P}_{k-1,l}(\mathbf{X})$ is proportional to $P_l(\mathbf{X})$ for $l = 1, ..., k-1$. Moreover, $\mathcal{D}^*(\mathbf{X}) \neq k$ implies that $P_k(\mathbf{X})$ cannot be the maximum. Therefore,

$$(k-1) = \text{argmax}_{l=1}^{k-1} P_{k-1,l}(\mathbf{X}) = \text{argmax}_{l=1}^{k} P_l(\mathbf{X}).$$

That is,

$$\mathcal{D}^*(\mathbf{X}) = (k-1) \text{ if and only if } (k-1) = \text{argmax}_{l=1}^{k} P_l(\mathbf{X}).$$

We continue this proof for the remaining classes and finally obtain Theorem 2.1.

### A.2 Proof of Theorem 2.2

We first examine the difference

$$\Delta_k = P(Y = k, \widehat{\mathcal{D}}(\mathbf{X}) \neq k) - P(Y = k, \mathcal{D}^*(\mathbf{X}) \neq k).$$

Clearly,

$$\Delta_k \leq P(Y = k, \widehat{\mathcal{D}}(\mathbf{X}) \neq k, \mathcal{D}^*(\mathbf{X}) = k).$$

From Theorem 1, for any $\boldsymbol{x}$ in the domain of $\mathbf{X}$, we let $j_1(\boldsymbol{x}), j_2(\boldsymbol{x}), ..., j_{k-1}(\boldsymbol{x})$ be the permutation of $\{1, ..., k-1\}$ such that

$$P(Y = j_1(\boldsymbol{x}) | \mathbf{X} = \boldsymbol{x}) < ... < P(Y = j_{k-1}(\boldsymbol{x}) | \mathbf{X} = \boldsymbol{x}).$$

Then, $\mathcal{D}^*(\boldsymbol{x}) = k$ implies that $f^*_{j_l(\boldsymbol{x})}(\boldsymbol{x}) < 0$ for any $l = 1, .., k-1$. On the other hand, $\widehat{\mathcal{D}}(\mathbf{X}) \neq k$ implies that for this particular permutation, there exists some $l = 1, ..., k-1$ such that $\widehat{f}_{j_l}(\boldsymbol{x}) > 0$ so

$\widehat{f}_{j_l}(\boldsymbol{x})f^*_{j_l}(\boldsymbol{x}) < 0$. Therefore, we obtain

$$\begin{aligned}
\Delta_k &\leq P\left(\cup_j \left\{Y = k, \text{there exists some } l = 1, ..., k-1 \text{ such that } \widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X}) < 0\right\}\right) \\
&\leq \sum_j P\left(Y = k, \text{there exists some } l = 1, ..., k-1 \text{ such that } \widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X}) < 0\right) \\
&\leq \sum_j P\left(Z_{j_1} = -1, ..., Z_{j_{k-1}} = -1, \widehat{f}_{j_l}(\mathbf{X})f^*_{j_l}(\mathbf{X}) < 0\right).
\end{aligned}$$

Hence, it suffices to bound each term on the right-hand side of the above inequality.

When $l = 1$, under conditions (2.1)-(2.4), from Theorem 8.25 in Steinwart and Christmann (2008), there exists a probability at least $1 - 3e^{-\epsilon}$ and a constant $C_1$ such that

$$P(Z_{j_1}\widehat{f}_{j_1}(\mathbf{X}) < 0) - P(Z_{j_1}f^*_{j_1}(\mathbf{X}) < 0) \leq C_1 Q_n(\epsilon),$$

where

$$Q_n(\epsilon) = \left\{\lambda_n^{\frac{\tau}{2+\tau}}\sigma_n^{-\frac{d\tau}{d+\tau}} + \sigma_n^{-\beta} + \epsilon\left(n\lambda_n^p\sigma_n^{\frac{1-p}{1+\epsilon_0 d}}\right)^{-\frac{q+1}{q+2-p}}\right\}$$

with any constant $\epsilon_0 > 0$ and $d/(d+\tau) < p < 2$. According to Lemma 5 in Bartlett et al. (2006) and condition (2.2), this gives

$$P(\widehat{f}_{j_1}(\mathbf{X})f^*_{j_1}(\mathbf{X}) < 0) \leq [C_1 Q_n(\epsilon)]^\alpha,$$

where $\alpha = q/(1+q)$.

When $l = 2$, because $Z_{ij_2}$ is no longer defined if $Z_{ij_1} = 1$, we extend to define $Z_{ij_2} = 1$ if $Z_{ij_1} = 1$. We then consider the following minimization

$$n^{-1}\sum_{i=1}^n I(Z_{ij_1}\widehat{f}_{j_1}(\mathbf{X}_i) > 0)(1 - Z_{ij_2}g(Z_{ij_1}, \mathbf{X}_i))_+ + \lambda_n(\|g(1, \boldsymbol{x})\| + \|g(-1, \boldsymbol{x})\|),$$

which is equivalent to minimizing

$$n^{-1}\sum_{i=1}^n I(Z_{ij_1} = 1, \widehat{f}_{j_1}(\mathbf{X}_i) > 0)(1 - g(1, \mathbf{X}_i))_+ + \lambda_n\|g(1, \boldsymbol{x})\|$$

and

$$n^{-1}\sum_{i=1}^{n} I(Z_{ij_1}=-1, \widehat{f}_{j_1}(\mathbf{X}_i)<0)(1-Z_{ij_2}g(-1,\mathbf{X}_i))_+ + \lambda_n\|g(-1,\boldsymbol{x})\|.$$

Thus, it is obvious that the optimal estimator for $g$, denoted by $\widehat{g}$, is given as

$$\widehat{g}(1,\boldsymbol{x})=1, \quad \widehat{g}(-1,\boldsymbol{x})=\widehat{f}_{j_2}(\boldsymbol{x}).$$

Similarly, the optimal estimator that minimizes the limit is given as

$$g^*(1,\boldsymbol{x})=1, \quad g^*(-1,\boldsymbol{x})=f^*_{j_2}(\boldsymbol{x}).$$

We then apply to $\widehat{g}$ the same arguments used by Steinwart and Christmann (2008) to prove Theorem 8.25 and obtain

$$P(Z_{j_2}\widehat{g}(Z_{j_1},\mathbf{X})<0)-P(Z_{j_2}g^*(Z_{j_1},\mathbf{X})<0)$$

$$\leq C_2\left\{Q_n(\epsilon)+|P(Z_{j_1}\widehat{f}_{j_1}(\mathbf{X})>0)-P(Z_{j_1}f^*_{j_1}(\mathbf{X})>0)|\right\}$$

with a probability at least $1-3e^{-\epsilon}$ for a constant $C_2$. The second term in the right-hand side is due to that the estimation is conditional on a random set with $Z_{ij_1}\widehat{f}_{j_1}(\mathbf{X}_i)>0$. On the other hand, from the previous result at $l=1$, this term is bounded by $C_1Q_n(\epsilon)$ with probability at least $1-3e^{-\epsilon}$. We conclude that with a probability at least $1-6e^{-\epsilon}$, it holds that

$$P(Z_{j_2}\widehat{g}(Z_{j_1},\mathbf{X})<0)-P(Z_{j_2}g^*(Z_{j_1},\mathbf{X})<0)\leq C_3Q_n(\epsilon)$$

for $C_3=C_2(1+C_1)$. From the fact that $\widehat{g}=g^*=1$ if $Z_{j_1}=1$, we have that with a probability at least $1-3e^{-\epsilon}$,

$$P(Z_{j_1}=-1, Z_{j_2}\widehat{f}_{j_2}(\mathbf{X})<0)-P(Z_{j_1}=-1, Z_{j_2}f^*_{j_2}(\mathbf{X})<0)$$

$$\leq C_3Q_n(\epsilon).$$

Thus, Lemma 5 in Bartlett et al. (2006) gives

$$P(Z_{j_1}=-1, \widehat{f}_{j_2}(\mathbf{X})f^*_{j_2}(\mathbf{X})<0)\leq[C_3Q_n(\epsilon)]^{\alpha}.$$

We contableinue the same arguments for $l = 3, ..., k-1$ to obtain

$$E\left[I\left\{Z_{j_l}\widehat{f}_{j_l}(\mathbf{X}) < 0, Z_{j_{l-1}} = -1, ..., Z_{j_1} = -1\right\}\right.$$

$$\left.-I\left\{Z_{j_l}f^*_{j_l}(\mathbf{X}) < 0, Z_{j_{l-1}} = -1, ..., Z_{j_1} = -1\right\}\right] \leq C_l Q_n(\epsilon)$$

with a probability at least $1 - 3le^{-\epsilon}$. Hence, with a probability $1 - [3k(k-1)/2]e^{-\epsilon}$, $\Delta_k \leq CQ_n(\epsilon)^\alpha$ for a constant $C$.

Similarly, we can examine the difference

$$\Delta_{k-1} \quad = \quad P(Y = k-1, \widehat{\mathcal{D}}(\mathbf{X}) \neq k-1) - P(Y = k-1, \mathcal{D}^*(\mathbf{X}) \neq k-1).$$

We follow exactly the same arguments as before by considering all possible permutations from $\{1, ..., k-2\}$ and $l = 1, ..., k-2$. The only difference in the argument is that the random set is restricted to subjects with $Y_i \neq k$ and $\widehat{\mathcal{D}}^{(k)} = -1$. However, the probability of the latter differs from the probability $Y_i \neq k$ and $\mathcal{D}^{*(k)} = -1$ by $CQ_n(\epsilon)$ from the previous conclusion. Therefore, we obtain that with a probability at least $1 - [3k(k-1)/2 + 3(k-1)(k-2)/2]e^{-\epsilon}$, $\Delta_{k-1} \leq CQ_n(\epsilon)^\alpha$ for another constant $C$. Continue the same arguments for $\Delta_l, l = k-2, ..., 1$, where $\Delta_l = P(Y = l, \widehat{\mathcal{D}}(\mathbf{X}) \neq l) - P(Y = l, \mathcal{D}^*(\mathbf{X}) \neq l)$. Finally, by combining all these results, we conclude that

$$P(Y \neq \widehat{\mathcal{D}}(\mathbf{X})) \leq P(Y \neq \mathcal{D}^*(\mathbf{X})) + CQ_n(\epsilon)^\alpha$$

with a probability at least $1 - C'e^{-\epsilon}$, where $C'$ is a constant depending on $k$. Theorem 2.1 holds.

### A.3 Proof of Theorem 3.1

We start from class label $k$ following the order in SOM. First, we show $\mathcal{D}^*(x) = k$ if and only if $E(R|X = x, A = k) = \max_{l=1}^k E(R|X = x, A = l)$. For any $x$ with $\mathcal{D}^*(x) = k$, by the definition of $\mathcal{D}^*$, there exists a permutation $(j_1, ..., j_{k-1})$ of $\{1, ..., k-1\}$ such that $\mathcal{D}_l^{*(k)}(x) = -1$ for $l = j_1, ..., j_{k-1}$. That is,

$$f^*_{j_1}(x) < 0, f^*_{j_2}(x) < 0, \ldots, f^*_{j_{k-1}}(x) < 0,$$

where $f^*_{j_l}$ is the counterpart of $\widehat{f}_{j_1}$ when $n = \infty$.

On the other hand, from the estimation of $\widehat{f}_{j_1}$, it is clear that $f^*_{j_1}$ is the minimizer of the expectation of a weighted hinge loss corresponding to $V_{n,j1}$, which is

$$
\begin{aligned}
&E\left\{\frac{k-1}{k}\frac{R^+I(A=j_1)}{\pi_{j_1}(x)}\{1-f(X)\}_+\middle| X=x\right\}+E\left\{\frac{1}{k}\sum_{l=2}^{k}\frac{R^-I(A=j_l)}{\pi_{j_l}(x)}\{1-f(X)\}_+\middle| X=x\right\}\\
&+E\left\{\frac{1}{k}\sum_{l=2}^{k}\frac{R^+I(A=j_l)}{\pi_{j_l}(x)}\{1+f(X)\}_+\middle| X=x\right\}+E\left\{\frac{k-1}{k}\frac{R^-I(A=j_1)}{\pi_{j_1}(x)}\{1+f(X)\}_+\middle| X=x\right\}\\
=&E\left(\frac{k-1}{k}R^+\middle| X=x,A=j_1\right)\{1-f(X)\}_++\sum_{l=2}^{k}E\left(\frac{R^-}{k}\middle| X=x,A=j_l\right)\{1-f(X)\}_+\\
&+\sum_{l=2}^{k}E\left(\frac{R^+}{k}\middle| X=x,A=j_l\right)\{1+f(X)\}_++E\left(\frac{k-1}{k}R^-\middle| X=x,A=j_1\right)\{1+f(X)\}_+
\end{aligned}
$$

where $R^+=RI(R>0)$, $R^-=-RI(R\leq 0)$, and $R=R^+-R^-$.

We first consider the case when $f(x)\in(-\infty,-1]$, the equation above can be reduced to

$$
\left\{E\left(\frac{k-1}{k}R^+\middle| X=x,A=j_1\right)+\sum_{l=2}^{k}E\left(\frac{R^-}{k}\middle| X=x,A=j_l\right)\right\}\{-f(X)\}+constant \qquad (4)
$$

It's clear that we cannot find a minimizer for (1). Similarly, the minimizer cannot be in the interval $f(x)\in[1,\infty)$. Therefore, we only consider $f(X)\in(-1,1)$. Then the expectation of a weighted hinge loss corresponding to $V_{n,j1}$ above is:

$$
\begin{aligned}
&E\left(\frac{k-1}{k}R^+\middle| X=x,A=j_1\right)\{1-f(X)\}_++\sum_{l=2}^{k}E\left(\frac{R^-}{k}\middle| X=x,A=j_l\right)\{1-f(X)\}_+\\
&+\sum_{l=2}^{k}E\left(\frac{R^+}{k}\middle| X=x,A=j_l\right)\{1+f(X)\}_++E\left(\frac{k-1}{k}R^-\middle| X=x,A=j_1\right)\{1+f(X)\}_+\\
=&\left\{\sum_{l=2}^{k}E\left(\frac{R}{k}\middle| X=x,A=j_l\right)-E\left(\frac{k-1}{k}R\middle| X=x,A=j_1\right)\right\}f(X)+constant
\end{aligned}
$$

That is, $f^*_{j_1}(X)<0$ is equivalent to

$$
E\left(\frac{k-1}{k}R\middle| X=x,A=j_1\right)<\sum_{l=2}^{k}E\left(\frac{R}{k}\middle| X=x,A=j_l\right),
$$

which is equivalent to

$$E\left(R|\,X=x,A=j_1\right) < \frac{1}{k-1}\sum_{l=2}^{k} E\left(R|\,X=x,A=j_l\right).$$

Next, we restrict to data with $A \neq j_1$ and $f^*_{j_1}(X) < 0$, it is clear that $f^*_{j_2}$ minimizes

$$\begin{aligned}
&E\left\{\frac{k-2}{k-1}\frac{R^+}{\pi_{j_2}(x)}I(A=j_2)\{1-f(X)\}\middle|\, X=x, A\neq j_1, f^*_{j_1}(X)<0\right\}\\
&+E\left\{\frac{1}{k-1}\sum_{l=3}^{k}\frac{R^-}{\pi_{j_l}(x)}I(A=j_l)\{1-f(X)\}\middle|\, X=x, A\neq j_1, f^*_{j_1}(X)<0\right\}\\
&+E\left\{\frac{1}{k-1}\sum_{l=3}^{k}\frac{R^+}{\pi_{j_l}(x)}I(A=j_l)\{1+f(X)\}\middle|\, X=x, A\neq j_1, f^*_{j_1}(X)<0\right\}\\
&+E\left\{\frac{k-2}{k-1}\frac{R^-}{\pi_{j_2}(x)}I(A=j_2)\{1+f(X)\}\middle|\, X=x, A\neq j_1, f^*_{j_1}(X)<0\right\}\\
=&E\left\{\frac{k-2}{k-1}R^+\middle|\, X=x, A=j_2, f^*_{j_1}(X)<0\right\}\{1-f(X)\}\\
&+\sum_{l=3}^{k}E\left\{\frac{R^-}{k-1}\middle|\, X=x, A=j_l, f^*_{j_1}(X)<0\right\}\{1-f(X)\}\\
&+\sum_{l=3}^{k}E\left\{\frac{R^+}{k-1}\middle|\, X=x, A=j_l, f^*_{j_1}(X)<0\right\}\{1+f(X)\}\\
&+E\left\{\frac{k-2}{k-1}R^-\middle|\, X=x, A=j_2, f^*_{j_1}(X)<0\right\}\{1+f(X)\}\\
=&\left[\sum_{l=3}^{k}E\left\{\frac{R}{k-1}\middle|\, X=x, A=j_l, f^*_{j_1}(X)<0\right\}-E\left\{\frac{k-2}{k-1}R\middle|\, X=x, A=j_2, f^*_{j_1}(X)<0\right\}\right]f(X)\\
&+\mathit{constant}
\end{aligned}$$

Thus, we conclude that

$$\text{sign}\{f^*_{j_2}(X)\} = \text{sign}[E\{(k-2)R|X=x,A=j_2\}-\sum_{l=3}^{k}E\{R|X=x,A=j_l\}]I\{f^*_{j_1}(X)<0\}.$$

That is, $f^*_{j2}(x) < 0$ if and only if

$$E(R|X=x,A=j_2) < \frac{1}{k-2}\sum_{l=3}^{k}E(R|X=x,A=j_l)$$

Continue the same arguments so we establish the relationship between $f^*_{j_l}$ and $E(R|X=x,A=j_l)$

as

$$\text{sign}\{f_{j_l}^*(x)\} = \text{sign}\left\{E(R|X=x, A=j_l) - \frac{1}{k-l}\sum_{h=l+1}^{k} E(R|X=x, A=j_h)\right\}$$

In other words, we obtain that for this subject with $f_{j_1}^*(x) < 0, ..., f_{j_{k-1}}^*(x) < 0$, it holds

$$\begin{aligned}
E(R|X=x, A=j_1) < & \quad \frac{1}{k-1}\sum_{l=2}^{k} E(R|X=x, A=j_l), \\
E(R|X=x, A=j_2) < & \quad \frac{1}{k-2}\sum_{l=3}^{k} E(R|X=x, A=j_l), \\
\vdots & \\
E(R|X=x, A=j_{k-2}) < & \quad 1/2\{E(R|X=x, A=j_{k-1}) + E(R|X=x, A=k)\}, \\
E(R|X=x, A=j_{k-1}) < & \quad E(R|X=x, A=k).
\end{aligned}$$

Starting from the last inequality in the above, in turn, we have

$$\begin{aligned}
E(R|X=x, A=j_{k-1}) & < E(R|X=x, A=k) \\
E(R|X=x, A=j_{k-2}) & < 1/2\{E(R|X=x, A=j_{k-1}) + E(R|X=x, A=k)\} \\
& < E(R|X=x, A=k), \\
\vdots & \\
E(R|X=x, A=j_1) & < \frac{1}{k-1}\sum_{l=2}^{k} E(R|X=x, A=j_1) < E(R|X=x, A=k).
\end{aligned}$$

Therefore,

$$E(R|X=x, A=k) = \max_{l=1}^{k} E(R|X=x, A=l).$$

For the other direction, we suppose that

$$E(R|X=x, A=k) = \max_{l=1}^{k} E(R|X=x, A=l).$$

We order the expectations to obtain

$$E(R|X = x, A = j_1) \leq E(R|X = x, A = j_2) \leq ... \leq E(R|X = x, A = k)$$

Thus all the inequalities in (2)-(7) hold, from equivalence between $f^*_{j_l}$ and $E(R|X = x, A = j_l)$'s, it is straightforward to see that

$$f^*_{j_1}(x) < 0, ..., f^*_{j_{k-1}}(x) < 0.$$

In other words, $\mathcal{D}^*(x) = k$. Hence, we have proved that SOM learning correctly assigns subjects whose conditional mean outcomes are maximal in treatment $k$ into the optimal treatment $k$.

To prove the consistency of the remaining classes, obtains the rule for class $(k-1)$ conditional on $A \neq k$ and $\mathcal{D}^*(X) \neq k$. Using the same proof as above, we conclude

$$\mathcal{D}^*(x) = (k-1) \text{ if and only if } (k-1) = \text{argmax}_{l=1}^{k-1} \widetilde{E}(R|X = x, A = l),$$

where $\widetilde{E}(R|X = x, A = j_l)$ is the conditional expectation of $R$ given $X = x$, $A \neq k$ and $\mathcal{D}^*(X) \neq k$. Moreover, $\mathcal{D}^*(x) \neq k$ implies that $E(R|X = x, A = k)$ cannot be the maximum. Therefore,

$$(k-1) = \text{argmax}_{l=1}^{k-1} E(R|X = x, A = l) = \text{argmax}_{l=1}^{k} E(R|X = x, A = l).$$

That is,

$$\mathcal{D}^*(x) = (k-1) \text{ if and only if } (k-1) = \text{argmax}_{l=1}^{k} E(R|X = x, A = l).$$

We continue this proof for the remaining classes and finally obtain Fisher consistency.

### A.4 Proof of Theorem 3.2

We first note

$$\begin{aligned}
&\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}(\mathcal{D}^*) \\
=&\sum_{l=1}^{k} \left[ E\left\{ \frac{R}{\pi_l(X)} I(A = l, \widehat{\mathcal{D}}(X) \neq l) \right\} - E\left\{ \frac{R}{\pi_l(X)} I(A = l, \mathcal{D}^*(X) \neq l) \right\} \right] \\
=&\sum_{l=1}^{k} \left[ E\left\{ \frac{R}{\pi_l(X)} I(A = l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} - E\left\{ \frac{R}{\pi_l(X)} I(A = l, \mathcal{D}^*(X) \neq l, \widehat{\mathcal{D}}(X) = l) \right\} \right].
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}(\mathcal{D}^*) \\
=&\sum_{l=1}^{k} \left[ E\left\{ \frac{R}{\pi_l(X)} I(A=l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} - E\left\{ \frac{R}{\pi_A(X)} I(A \neq l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} \right]. \\
\leq&\sum_{l=1}^{k} \left[ E\left\{ \frac{R^+}{\pi_A(X)} I(A=l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} + E\left\{ \frac{R^-}{\pi_A(X)} I(A \neq l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} \right]
\end{aligned}
$$

We let $\Delta_l$ to denote each term on the right-hand side of the above equation. That is,

$$
\begin{aligned}
\Delta_l &= E\left\{ \frac{R^+}{\pi_A(X)} I(A=l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} + E\left\{ \frac{R^-}{\pi_A(X)} I(A \neq l, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\} \\
&= E\left\{ \frac{|R|}{\pi_A(X)} I(Z_l \text{sign}(R) = 1, \widehat{\mathcal{D}}(X) \neq l, \mathcal{D}^*(X) = l) \right\},
\end{aligned}
$$

where we recall $Z_l = 2I(A = l) - 1$.

We first examine $\Delta_k$. For any $x$ in the domain of $X$, we let $j_1, j_2, ..., j_{k-1}$ be the permutation of $\{1, ..., k-1\}$ such that

$$
E(R|A = j_1, X = x) < ... < E(R|A = j_{k-1}, X = x).
$$

Then according to SOM learning, $\mathcal{D}^*(x) = k$ implies that $f^*_{j_l(x)}(x) < 0$ for any $l = 1, .., k-1$, while $\widehat{\mathcal{D}}(X) \neq k$ implies that for this particular permutation, there exists some $l = 1, ..., k-1$ such that $\widehat{f}_{j_l}(x) > 0$ so $\widehat{f}_{j_l}(x) f^*_{j_l}(x) < 0$. Recall that $f^*_{j_l}(x) = \eta_{j_l, S}$ with $S = \{j_{l+1}, ..., k\}$ and it is the limit of $\widehat{f}_{j_l}$ from Theorem 3.1. Therefore, we obtain

$$
\begin{aligned}
\Delta_k &\leq E\left[ \frac{|R|}{\pi_A(X)} \left\{ \sum_{(j_1, ..., j_{k-1})} I(Z_k \text{sign}(R) = 1, \text{there exists } l \leq k-1 \text{ s.t. } \widehat{f}_{j_l}(X) f^*_{j_l}(X) < 0) \right\} \right] \\
&\leq \sum_{(j_1, ..., j_{k-1})} E\left[ \frac{|R|}{\pi_A(X)} I\left\{ Z_{j_1} \text{sign}(R) = -1, ..., Z_{j_{l-1}} \text{sign}(R) = -1, \widehat{f}_{j_l}(X) f^*_{j_l}(X) < 0 \right\} \right] \\
&\leq \sum_{(j_1, ..., j_{k-1})} E\left[ \frac{|R|}{\pi_A(X)} \{ I(A = j_l)(k - l + 1) + I(A \neq j_l) \} \right. \\
&\qquad \left. \times I\left\{ Z_{j_1} \text{sign}(R) = -1, ..., Z_{j_{l-1}} \text{sign}(R) = -1, \widehat{f}_{j_l}(X) f^*_{j_l}(X) < 0 \right\} \right].
\end{aligned}
$$

Hence, it suffices to bound each term on the right-hand side of the above inequality.

When $l = 1$, under conditions 3.1-3.3, we use the same proof of Theorem 3.2 in Zhao et al. (2012), which extends the result in Steinwart and Christmann (2008) to a weighted support vector machine. Particularly, in their proof, we let the weight for subject $i$ be

$$|R_i|/\pi_{A_i}(X_i)\left\{(k-1)I(A_i = j_1) + I(A_i \neq j_1)\right\}$$

and the class label be $Z_{j_1}\mathrm{sign}(R_i)$. Furthermore, from the proof of Theorem 3.1, $f^*_{j_1}(x)$ has the same sign as $\eta_{j_1,\{j_2,\ldots,j_k\}}(x)$. Thus, from condition (3.1), we conclude that there exists at least probability $1-3e^{-\epsilon}$ and a constant $C_1$ such that it holds

$$E\left[\frac{|R|}{\pi_A(X)}\left\{(k-1)I(A = j_1) + I(A \neq j_1)\right\} I(Z_{j_1}\mathrm{sign}(R)\widehat{f}_{j_1}(X) < 0)\right]$$

$$-E\left[\frac{|R|}{\pi_A(X)}\left\{(k-1)I(A = j_1) + I(A \neq j_1)\right\} I(Z_{j_1}\mathrm{sign}(R)f^*_{j_1}(X) < 0)\right] \leq C_1 Q_n(\epsilon),$$

where

$$Q_n(\epsilon) = \left\{\lambda_n^{\frac{\tau}{2+\tau}}\sigma_n^{-\frac{d\tau}{d+\tau}} + \sigma_n^{\beta} + \epsilon\left(n\lambda_n^p\sigma_n^{\frac{1-p}{1+\epsilon_0 d}}\right)^{-\frac{q+1}{q+2-p}}\right\}$$

with any constant $\epsilon_0 > 0$ and $d/(d+\tau) < p < 2$. Then according to the proof of Lemma 5 in Barlette et al. (2006) and conditions 3.1 and 3.2, this gives

$$pr\{\widehat{f}_{j_1}(X)f^*_{j_1}(X) < 0\} \leq \{C_1' Q_n(\epsilon)\}^{\alpha},$$

where $\alpha = q/(1+q)$ and $C_1'$ is a constant.

When $l = 2$, the step at $j_2$ in SOM is to minimize

$$n^{-1}\sum_{i=1}^{n} I\{Z_{ij_1} = -1, Z_{ij_1}\mathrm{sign}(R_i)\widehat{f}_{j_1}(X_i) < 0\} w_i \{1 - Z_{ij_2}\mathrm{sign}(R_i)f(X_i)\}_+ + \lambda_{n,j_2}\|f\|^2,$$

where $w_i = |R_i|/\pi_{A_i}(X_i)\left\{(k-2)I(A_i = j_2) + I(A_i \neq j_2)\right\}$. Thus, we can proceed the same proof of Theorem 3.2 in Zhao et al. (2012) except that only subjects in the random set

$$\left\{i : Z_{ij_1} = -1, Z_{ij_1}\mathrm{sign}(R_i)\widehat{f}_{j_1}(X_i) < 0\right\}$$

are used in the derivation. We obtain that

$$
\begin{aligned}
& E\left[\frac{|R|}{\pi_A(X)}\{(k-2)I(A=j_2)+I(A\neq j_2)\}I\{Z_{j_1}=-1, Z_{j_2}\text{sign}(R)\widehat{f}_{j_2}(X)<0\}\right] \\
& \qquad -E\left[\frac{|R|}{\pi_A(X)}\{(k-2)I\{A=j_2)+I(A\neq j_2)\}I\{Z_{j_1}=-1, Z_{j_2}\text{sign}(R)f^*_{j_2}(X)<0\}\right] \\
\leq \;& C_2\left\{Q_n(\epsilon)+|pr(Z_{j_1}\text{sign}(R)\widehat{f}_{j_1}(X)>0)-pr(Z_{j_1}\text{sign}(R)f^*_{j_1}(X)>0)|\right\} \\
\leq \;& C_2\{Q_n(\epsilon)+Q_n(\epsilon)^\alpha\}
\end{aligned}
$$

with a probability at least $1-3e^{-\epsilon}$ for a constant $C_2$. Note that the second term on the right-hand side is due to the estimated random set in this step. Again, the proof of Lemma 5 in Barlette et al. (2006) gives

$$
pr\{Z_{j_1}=-1, \widehat{f}_{j_2}(X)f^*_{j_2}(X)<0\}\leq\{C'_2Q_n(\epsilon)\}^\alpha.
$$

We continue the same arguments for $l=3,...,k-1$ to obtain

$$
\begin{aligned}
& E\left[\frac{|R|}{\pi_A(X)}\{(k-l+1)I(A=j_l)+I(A\neq j_l)\}\right. \\
& \quad \left.\times\, I\left\{Z_{j_l}\text{sign}(R)\widehat{f}_{j_l}(X)<0, Z_{j_{l-1}}=-1,...,Z_{j_1}=-1\right\}\right] \\
& -E\left[\frac{|R|}{\pi_A(X)}\{(k-l+1)I(A=j_l)+I(A\neq j_l)\}\right. \\
& \quad \left.\times\, I\left\{Z_{j_l}f^*_{j_l}(X)<0, Z_{j_{l-1}}=-1,...,Z_{j_1}=-1\right\}\right] \\
& \leq C_l\{Q_n(\epsilon)+Q_n(\epsilon)^\alpha\}
\end{aligned}
$$

with a probability at least $1-3le^{-\epsilon}$ for some constant $C_l$, and

$$
pr\{Z_{j_1}=-1,...,Z_{j_{l-1}}=-1, \widehat{f}_{j_l}(X)f^*_{j_l}(X)<0\}\leq\{C'_lQ_n(\epsilon)\}^\alpha
$$

for a constant $C'_l$. Hence, with a probability $1-\{3k(k-1)/2\}e^{-\epsilon}$, $\Delta_k\leq CQ_n(\epsilon)^\alpha$ for a constant $C$.

Similarly, we can examine the difference for $\Delta_{k-1}$. We follow exactly the same arguments as before by considering all possible permutations from $\{1,...,k-2\}$ and $l=1,...,k-2$. The only difference in the argument is that the random set is restricted to subjects with $A\neq k$ and $\widehat{\mathcal{D}}^{(k)}(X)=-1$. However, the probability of the latter differs from the probability $A\neq k$ and

$\mathcal{D}^{*(k)}(X) = -1$ by $CQ_n(\epsilon)^\alpha$ from the previous conclusion. Therefore, we obtain that with probability at least $1 - \{3k(k-1)/2 + 3(k-1)(k-2)/2\}e^{-\epsilon}$, $\Delta_{k-1} \leq CQ_n(\epsilon)^\alpha$ for another constant $C$. Continue the same arguments for $\Delta_l, l = k-2, ..., 1$ so we finally conclude

$$\mathcal{R}(\widehat{\mathcal{D}}) - \mathcal{R}^* \leq CQ_n(\epsilon)^\alpha$$

with probability at least $1 - C'e^{-\epsilon}$ where $C'$ is a constant depending on $k$. Thus Theorem 3.2 holds.

# REFERENCES

Abd, T. T., and Jacobson, T. A. (2011), “Statin-induced myopathy: a review and update,” *Expert opinion on drug safety*, 10, 373–387.

Allwein, E. L., Schapire, R. E., and Singer, Y. (2001), “Reducing multiclass to binary: A unifying approach for margin classifiers,” *The Journal of Machine Learning Research*, 1, 113–141.

Almasi, G. S., and Gottlieb, A. (1988), “Highly parallel computing,” , .

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006), “Convexity, classification, and risk bounds,” *Journal of the American Statistical Association*, 101, 138–156.

Biondich, P. G., and Grannis, S. J. (2004), “The Indiana network for patient care: an integrated clinical information system informed by over thirty years of experience,” *Journal of Public Health Management and Practice*, 10, S81–S86.

Bishop, C. M. (2006), “Pattern Recognition,” *Machine Learning*, .

Bosco, J. L., Silliman, R. A., Thwin, S. S., Geiger, A. M., Buist, D. S., Prout, M. N., Yood, M. U., Haque, R., Wei, F., and Lash, T. L. (2010), “A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies,” *Journal of clinical epidemiology*, 63, 64–74.

Bredensteiner, E. J., and Bennett, K. P. (1999), “Multicategory classification by support vector machines,” in *Computational Optimization* Springer, pp. 53–79.

Carini, C., Menon, S. M., and Chang, M. (2014), *Clinical and Statistical Considerations in Personalized Medicine*, CRC Press.

Chang, C.-C., and Lin, C.-J. (2011), “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 27.

Chatzizisis, Y. S., Koskinas, K. C., Misirli, G., Vaklavas, C., Hatzitolios, A., and Giannoglou, G. D. (2010), “Risk factors and drug interactions predisposing to statin-induced myopathy,” *Drug safety*, 33, 171–187.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, L. (2011), “ClustOfVar: an R package for the clustering of variables,” *arXiv preprint arXiv:1112.0295*, .

Crammer, K., and Singer, Y. (2002), “On the algorithmic implementation of multiclass kernel-based vector machines,” *The Journal of Machine Learning Research*, 2, 265–292.

Cristianini, N., and Shawe-Taylor, J. (2000), *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J. S. et al. (2011), “Identification of conversion from mild cognitive impairment to Alzheimer’s disease using multivariate predictors,” *PLoS ONE*, 6, e21896.

Dietterich, T. G., and Bakiri, G. (1995), “Solving multiclass learning problems via error-correcting output codes,” *Journal of Artificial Intelligence Research*, pp. 263–286.

Freund, Y., and Schapire, R. E. (1997), “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of computer and system sciences*, 55, 119–139.

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The Elements of Statistical Learning,*, Vol. 1 Springer series in statistics Springer, Berlin.

Golomb, B. A., and Evans, M. A. (2008), “Statin adverse effects,” *American Journal of Cardiovascular Drugs*, 8, 373–418.

Gunter, L., Zhu, J., and Murphy, S. (2011), “Variable selection for qualitative interactions,” *Statistical Methodology*, 8, 42–55.

Hill, S. I., and Doucet, A. (2007), “A Framework for Kernel-Based Multi-Category Classification.,” *J. Artif. Intell. Res.(JAIR)*, 30, 525–564.

Ho, T. K. (1995), “Random decision forests,” in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, IEEE, pp. 278–282.

Ho, T. K. (1998), “The random subspace method for constructing decision forests,” *IEEE transactions on pattern analysis and machine intelligence*, 20, 832–844.

Hsu, C.-W., and Lin, C.-J. (2002), “A comparison of methods for multiclass support vector machines,” *Neural Networks, IEEE Transactions on*, 13, 415–425.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. et al. (2001), “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature medicine*, 7, 673–679.

Klein, D. N., Leon, A. C., Li, C., D’Zurilla, T. J., Black, S. R., Vivian, D., Dowling, F., Arnow, B. A., Manber, R., Markowitz, J. C. et al. (2011), “Social problem solving and depressive symptoms over time: A randomized clinical trial of cognitive-behavioral analysis system of psychotherapy, brief supportive psychotherapy, and pharmacotherapy.,” *Journal of Consulting and Clinical Psychology*, 79, 342.

Kocsis, J. H., Gelenberg, A. J., Rothbaum, B. O., Klein, D. N., Trivedi, M. H., Manber, R., Keller, M. B., Leon, A. C., Wisniewski, S. R., Arnow, B. A. et al. (2009), “Cognitive behavioral analysis system of psychotherapy and brief supportive psychotherapy for augmentation of antidepressant nonresponse in chronic depression: the REVAMP Trial,” *Archives of General Psychiatry*, 66, 1178–1188.

Kosorok, M. R., and Moodie, E. E. (2015), *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine,*, Vol. 21 SIAM.

Kreßel, U. H.-G. (1999), “Pairwise classification and support vector machines,” in *Advances in Kernel Methods*, MIT Press, pp. 255–268.

Lauer, F., and Guermeur, Y. (2011), “MSVMpack: a multi-class support vector machine package,” *The Journal of Machine Learning Research*, 12, 2293–2296.

Lee, Y., Lin, Y., and Wahba, G. (2004), “Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data,” *Journal of the American Statistical Association*, 99, 67–81.

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

Lipska, K. J., and Krumholz, H. M. (2014), "Comparing diabetes medications: where do we set the bar?," *JAMA Internal Medicine*, 174, 317–318.

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

Liu, Y., and Shen, X. (2006), "Multicategory $\psi$-learning," *Journal of the American Statistical Association*, 101, 500–509.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2014), "Robust Hybrid Learning for Estimating Personalized Dynamic Treatment Regimens," *arXiv preprint arXiv:1611.02314*, .

Liu, Y., and Yuan, M. (2011), "Reinforced multicategory support vector machines," *Journal of Computational and Graphical Statistics*, 20, 901–919.

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

Madigan, D., Stang, P. E., Berlin, J. A., Schuemie, M., Overhage, J. M., Suchard, M. A., Dumouchel, B., Hartzema, A. G., and Ryan, P. B. (2014), "A systematic statistical approach to evaluating evidence from observational studies," *Annual Review of Statistics and Its Application*, 1, 11–39.

McDonald, C. J., Overhage, J. M., Barnes, M., Schadow, G., Blevins, L., Dexter, P. R., Mamlin, B., Committee, I. M. et al. (2005), "The Indiana network for patient care: a working local health information infrastructure," *Health affairs*, 24, 1214–1220.

Moodie, E. E., Richardson, T. S., and Stephens, D. A. (2007), "Demystifying optimal dynamic treatment regimes," *Biometrics*, 63, 447–455.

Murphy, S. A. (2003), "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 331–355.

Murphy, S. A. (2005), "A generalization error for Q-learning," *Journal of Machine Learning Research*, 6, 1073–1097.

Norén, G. N., Hopstadius, J., Bate, A., Star, K., and Edwards, I. R. (2010), "Temporal pattern discovery in longitudinal electronic patient records," *Data Mining and Knowledge Discovery*, 20, 361–387.

Pasternak, R. C., Smith, S. C., Bairey-Merz, C. N., Grundy, S. M., Cleeman, J. I., Lenfant, C. et al. (2002), "ACC/AHA/NHLBI clinical advisory on the use and safety of statins," *Circulation*, 106, 1024–1028.

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K., and Breckenridge, A. M. (2004), "Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients," *Bmj*, 329, 15–19.

Qian, M., and Murphy, S. A. (2011), "Performance guarantees for individualized treatment rules," *Annals of Statistics*, 39, 1180.

Rifkin, R., and Klautau, A. (2004), "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, 5, 101–141.

Robins, J. M. (2004), "Optimal structural nested models for optimal sequential decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, Springer, pp. 189–326.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Schölkopf, B., and Smola, A. J. (2002), *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003), "On $\psi$-learning," *Journal of the American Statistical Association*, 98, 724–734.

Specht, D. F. (1990), "Probabilistic neural networks," *Neural networks*, 3, 109–118.

Steinwart, I., and Christmann, A. (2008), *Support vector machines*, Springer Science & Business Media.

Sweetman, S. C., and Blake, P. (2011), "Martindale," *The complete drug reference*, 33.

Taylor, F., Ward, K., Moore, T., Burke, M., Davey Smith, G., Casas, J. P., and Ebrahim, S. (2011), "Statins for the primary prevention of cardiovascular disease," *Cochrane database syst rev*, 1.

Tewari, A., and Bartlett, P. L. (2007), "On the consistency of multiclass classification methods," *The Journal of Machine Learning Research*, 8, 1007–1025.

Trivedi Madhukar, H. et al. (2008), "Treatment strategies to improve and sustain remission in major depressive disorder," *Dialogues in Clinical Neuroscience*, 10, 377–384.

Vapnik, V. N., and Vapnik, V. (1998), *Statistical Learning Theory*,, Vol. 1 Wiley New York.

Walker, A. M. (1996), "Confounding by indication.," *Epidemiology*, 7, 335–336.

Watkins, C. J. C. H. (1989), Learning from delayed rewards, PhD thesis, University of Cambridge England.

Weatherby, L. B., Nordstrom, B. L., Fife, D., and Walker, A. M. (2002), "The impact of wording in "Dear doctor" letters and in black box labels," *Clinical Pharmacology & Therapeutics*, 72, 735–742.

Weiner, M. W., Aisen, P. S., Jack, C. R., Jagust, W. J., Trojanowski, J. Q., Shaw, L., Saykin, A. J., Morris, J. C., Cairns, N., Beckett, L. A. et al. (2010), "The Alzheimer's disease neuroimaging initiative: progress report and future plans," *Alzheimer's & Dementia*, 6, 202–211.

Weston, J., Watkins, C. et al. (1999), "Support vector machines for multi-class pattern recognition.," in *ESANN*, pp. 219–224.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A robust method for estimating optimal treatment regimes," *Biometrics*, 68, 1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013), "Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions," *Biometrika*, 100.

Zhang, P., Du, L., Wang, L., Liu, M., Cheng, L., Chiang, C.-W., Wu, H.-Y., Quinney, S., Shen, L., and Li, L. (2015), "A Mixture Dose–Response Model for Identifying High-Dimensional Drug Interaction Effects on Myopathy Using Electronic Medical Record Databases," *CPT: pharmacometrics & systems pharmacology*, 4, 474–480.

Zhang, T. (2004), "Statistical analysis of some multi-category large margin classification methods," *The Journal of Machine Learning Research*, 5, 1225–1251.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating individualized treatment rules using outcome weighted learning," *Journal of the American Statistical Association*, 107, 1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011), "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer," *Biometrics*, 67, 1422–1433.

Zhu, J., and Hastie, T. (2005), "Kernel logistic regression and the import vector machine," *Journal of Computational and Graphical Statistics*, 14, 185–205.