Published in final edited form as:

J Stat Plan Inference. 2016 May 1; 172: 23–25. doi:10.1016/j.jspi.2015.12.006.

A general approach to categorizing a continuous scale according to an ordinal outcome

Limin Peng^{a,*}, Amita Manatunga^a, Ming Wang^b, Ying Guo^a, and AKM Fazlur Rahman^a

^a Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Road NE., Atlanta, GA 30322, USA

^b Department of Public Health Sciences, Penn State University, 500 University Drive, Hershey, PA 17033, USA

Abstract

In practice, disease outcomes are often measured in a continuous scale, and classification of subjects into meaningful disease categories is of substantive interest. To address this problem, we propose a general analytic framework for determining cut-points of the continuous scale. We develop a unified approach to assessing optimal cut-points based on various criteria, including common agreement and association measures. We study the nonparametric estimation of optimal cut-points. Our investigation reveals that the proposed estimator, though it has been ad-hocly used in practice, pertains to nonstandard asymptotic theory and warrants modifications to traditional inferential procedures. The techniques developed in this work are generally adaptable to study other estimators that are maximizers of nonsmooth objective functions while not belonging to the paradigm of M-estimation. We conduct extensive simulations to evaluate the proposed method and confirm the derived theoretical results. The new method is illustrated by an application to a mental health study.

Keywords

Agreement; Association; Empirical process; M-estimation; Non-smooth objective function; Subsampling

1. Introduction

In many biomedical and behavioral studies, to identify a certain disease in human body, different instruments or rating scales are utilized. Typically, measurements are made on a continuous scale; however, researchers are often interested in dividing a continuous scale into ordered categories for reasons such as clinical interpretations of results and simplification of the instrument (O'Brien, 2004). For example, depression is a common problem in medically-ill patients with diabetes and other chronic diseases (Moussavi et al., 2007). The psychiatric diagnostic interview instruments, such as the Mini International Neuropsychiatric Interview (MINI) diagnostic interview (Sheehan et al., 1998), in general

*Corresponding author. Tel.: +1 404 727 7701. lpeng@emory.edu (L. Peng)..

provide accurate psychiatric diagnoses in the medically healthy individuals. However, the MINI interview instrument is too time-consuming for sick patients and requires trained psychiatric interviewers, which are not always affordable or available. On the other hand, dimensional psychometric instruments designed to measure the same state of the disease, such as the 20-item Zung Depression rating scale (Zung, 1965), require less time and could be self-administered by patients. The established total observer-rated MINI score has been interpreted with well-accepted graded severity of depression such as no depression, mild depression, and markedly severe depression (Sheehan et al., 1998). While the self-reported Zung scale has many advantages, there are no routine, reliable cut-points of Zung provided to reflect the degree of the severity of depression. Establishment of such cut-points would enhance the utility of the convenient Zung rating scale in the screening or diagnosis of depression, particularly in large medically-ill patient populations.

Analytic methods for determining cut-points of a continuous scale based on validated categorical measurements, despite their practical importance, have not been well studied. For example, Youden index (Youden, 1950) and its variants based on receiver operating curve (ROC) (Pepe, 2003) were studied for identifying the “best” cut-point to dichotomize a continuous scale (Kraemer, 1988; Schisterman et al., 2005; Perkins and Schisterman, 2006). However, this type of approach can only deal with a single cut-point, and moreover, lack formal inference procedures about the estimated cut-point. For the general cases possibly involving two or more cut-points, existing approaches are mostly based on ad hoc arguments and generally lack statistical rigor. For example, the following methods were used in the literature: (a) considering arbitrary cut-points or a certain sample quantile like the median or a cut-point that corresponds to the highest proportions of correct classification with a gold standard (Altman et al., 1994; Mazumdar and Glassman, 2000); (b) finding out cut-points that result in disease rates consistent with a known population disease prevalence (Altman, 1991); and (c) depending on clinicians’ experience (Altman et al., 1994). The fundamental deficiency of all these methods is that the criterion levels or cut-points are generally decided subjectively on the basis of the probabilities of Type I/II misclassification, and “judgment” or “experience”. Several authors (James, 1978; Brownie and Habicht, 1984) proposed to minimize the variance estimator of the prevalence of a disease by assuming a mixture of normal distributions; however, this criterion is restricted by the distribution and is valid only for the cases with a single cut-point. Baughman et al. (2006) also used a mixture model but the maximum likelihood estimators may not be accurate when the distributions are not well separated, the sample size is small, or the mixture model is misspecified.

In this work, we seek to develop a general and objective analytic framework for addressing the problem of determining cut-points in a continuous scale according to an established ordinal scale. To this end, a fundamental question is, what are the most desirable or optimal cut-points? A common viewpoint in practice is that meaningful cut-points in a continuous scale should produce high agreement or association between the newly categorized continuous scale and established categories. Therefore, we propose to evaluate each set of cut-points by some criterion that reflects a desirable relationship between two ordinal scales (e.g. high agreement or association). From doing so, we obtain a function of cut-points, which we shall refer to as a criterion function. We then define the optimal set of cut-points as the one that optimizes the criterion function.

We propose a general formulation of criterion functions, which are expressed as a smooth function of cell probabilities of the contingency table formed from cross-tabulating an established categorical scale and the newly categorized continuous scale based on a given set of cut-points. Our definition encompasses many important special cases, including those where cut-points are evaluated by weighted kappa (Cohen, 1960; Agresti, 1990), Kendall's τ_b , correct classification rate (Kendall, 1938; Agresti, 1990), or Youden's index in ROC. The general specification of criterion functions forms the foundation of the proposed unified framework for investigating cut-points based on various criteria.

We study the estimation of optimal cut-points without imposing any parametric assumptions on the distributions of the data. We consider a natural approach, which is to optimize a nonparametric estimator of the criterion function, called empirical criterion function hereafter. For instance, one may adopt weighted kappa statistic when the criterion for optimal cut-points is weighted kappa coefficient. While the basic idea is conceptually intuitive and has been adopted in practice in an ad-hoc way, our detailed investigation indicates that such a method is subject to nonstandard theory and requires special attention to its inference procedures. For example, the resulting estimator has a slower convergence rate lower than the usual root n , and may not possess asymptotic normality. The main issue is that an empirical criterion function is usually not smooth, and more specifically, involves cut-points through indicator functions. The challenge resembles the difficulty in M-estimation with non-smooth objective function (Chernoff, 1964, for example). Nevertheless, the proposed estimator is yet not a M-estimator. As a result, existing methods that deal with non-smooth M-estimation, for example, Kim and Pollard (1990), are not directly applicable.

In this work, we employ the technique of empirical processes and conduct rigorous asymptotic studies for the proposed nonparametric cut-point estimator. It is important to point out that our theoretical framework is quite general and may be adapted to many other estimation settings that involve maximization of non-smooth objective functions. Provided the nonstandard asymptotic properties of the proposed estimator, bootstrapping fails to work properly (Kosorok, 2008). We propose to use subsampling (Politis et al., 1999) as a well justified device for inferences including variance estimation and confidence intervals.

We elaborate in Section 2 the proposed method for determining cut-points in a continuous scale. We present the general problem formulation, the proposed nonparametric estimation, and the corresponding asymptotic results and inference. Extensive simulation studies reported in Section 3 demonstrate satisfactory finite-sample performance of our proposals, and also help confirm some of our theoretical results. In Section 4, we illustrate our method via an application to a mental health study. Our analysis suggests a refinement of current empirical rules for categorizing depression among diabetic subjects based on the Zung rating scale. A few concluding remarks are provided in Section 5.

2. The proposed method

2.1. A general formulation of optimal cut-points

Let Y denote an ordinal measurement that takes ordinal values, $1 < \dots < L$. Let X denote a continuous measurement bounded between x_L and x_U . All possible cut-points for X

according to the L categories of Y form a compact parameter space, denoted by

$\Theta = \{(d_1, \dots, d_{L-1}) : d_0 = x_L < d_1 < \dots < d_{L-1} < d_L = d_U\}$. For $\vec{d} = (d_1, \dots, d_{L-1}) \in \Theta$, we define $\tilde{X}(\vec{d}) = \sum_{k=1}^L k I(d_{k-1} \leq X < d_k)$, where $I(\cdot)$ is the indicator function.

It is clear that $\tilde{X}(\vec{d})$ and Y form a $L \times L$ contingency table, which has cell probabilities, as arranged in a vector, $\mathbf{P}(\vec{d}) = (p_{1,1}(\vec{d}), \dots, p_{1,L}(\vec{d}), \dots, p_{L,1}(\vec{d}), \dots, p_{L,L}(\vec{d}))^T$, where $p_{ij}(\vec{d}) = Pr(\tilde{X}(\vec{d}) = i, Y = j)$. Let $\vartheta(\cdot)$ denote a smooth function from $[0, 1]^{L^2}$ to $\mathbf{R}$, at least twice differentiable almost everywhere. We define the optimal set of cut-points, $\vec{d}_0$, as

$$\vec{d}_0 = \arg \max_{\vec{d} \in \Theta} \vartheta(\mathbf{P}(\vec{d})). \quad (1)$$

Here $\vartheta(\mathbf{P}(\vec{d}))$ serves as a general form of the criterion function, which offers great flexibility in scientific applications. If the view taken to determine cut-points is that the newly categorized scale $\tilde{X}(\vec{d})$ should produce high agreement or association with the ordinal categories of Y , we can properly choose the function $\vartheta(\cdot)$ so that $\vartheta(\mathbf{P}(\vec{d}))$ represents an agreement or association measure of interest. For instance, given the cell probabilities $p_{ij}(\vec{d})$ and the marginal probabilities $p_{i\cdot}(\vec{d}) \equiv \sum_{j=1}^L p_{ij}(\vec{d})$ and $p_{\cdot j}(\vec{d}) \equiv \sum_{i=1}^L p_{ij}(\vec{d})$ ($i, j = 1, \dots, L$), the weighted kappa coefficient, as a popular agreement measure, can be expressed as

$$\vartheta_{\kappa}(\mathbf{P}(\vec{d})) = \frac{\sum_{i=1}^L \sum_{j=1}^L \omega_{ij} p_{ij}(\vec{d}) - \sum_{i=1}^L \sum_{j=1}^L \omega_{ij} p_{i\cdot}(\vec{d}) p_{\cdot j}(\vec{d})}{1 - \sum_{i=1}^L \sum_{j=1}^L \omega_{ij} p_{i\cdot}(\vec{d}) p_{\cdot j}(\vec{d})}, \quad (2)$$

where the weights $\{\omega_{ij}\}_{i,j=1}^L$ are specified and represent the degree of discrepancy between two categories. Two common choices for ω_{ij} include linear weights $\omega_{ij} = 1 - \frac{|i-j|}{L-1}$ or quadratic weights $\omega_{ij} = 1 - \frac{(i-j)^2}{(L-1)^2}$. In addition, a simpler uncorrected agreement measure is correct classification rate, which corresponds to

$$\vartheta_{CC}(\mathbf{P}(\vec{d})) = \sum_{i=1}^L p_{ii}(\vec{d}). \quad (3)$$

Kendall's τ_b (Agresti, 1990) is a well-known association measure for ordinal measurements, and can be written as

$$\vartheta_{\tau}(\mathbf{P}(\vec{d})) = \frac{\sum_{i=1}^{L-1} \sum_{j=1}^{L-1} p_{ij}(\vec{d}) \left(\sum_{t=i+1}^L \sum_{m=j+1}^L p_{tm}(\vec{d}) \right) - \sum_{i=1}^{L-1} \sum_{j=2}^L p_{ij}(\vec{d}) \left(\sum_{t=i+1}^L \sum_{m=1}^{j-1} p_{tm}(\vec{d}) \right)}{\frac{1}{2} \left\{ \left(1 - \sum_{i=1}^L p_{i\cdot}(\vec{d})^2 \right) \left(1 - \sum_{j=1}^L p_{\cdot j}(\vec{d})^2 \right) \right\}^{1/2}}. \quad (4)$$

In the special case with $L = 2$, we can also show that the specification of $\vartheta_{\tau}(\mathbf{P}(\vec{d}))$ can encompass Youden's index, a measure for determining the cut-point in the context of ROC, which equals,

$$J(\mathbf{P}(\vec{d})) = \frac{p_{11}(\vec{d})}{p_{\cdot 1}(\vec{d})} + \frac{p_{22}(\vec{d})}{p_{\cdot 2}(\vec{d})} - 1.$$

These examples illustrate the general applicability of the proposed framework for determining cut-points in a continuous scale. Our general formulation of optimal cut-points in (1) accommodates many different ways that researchers decide and interpret cut-points in practice.

2.2. Nonparametric estimation of optimal cut-points

We study the nonparametric estimation of $\vec{d}_0$ that does not require parametric assumptions about data distributions. The basic idea is to estimate the criterion function $\vartheta(\mathbf{P}(\vec{d}))$ by its empirical counterpart, $\vartheta(\mathbf{P}_n(\vec{d}))$, and then find the maximizer of the empirical criterion function $\vartheta(\mathbf{P}_n(\vec{d}))$. Here

$\hat{\mathbf{P}}_n(\vec{d}) = (\hat{p}_{11}(\vec{d}), \dots, \hat{p}_{1L}(\vec{d}), \dots, \hat{p}_{L1}(\vec{d}), \dots, \hat{p}_{LL}(\vec{d}))^T$ with

$$\hat{p}_{ij}(\vec{d}) = \frac{1}{n} \sum_{k=1}^n I(\tilde{X}_k(\vec{d}) = i, Y_k = j).$$

More specifically, we propose to estimate $\vec{d}_0$ by

$$\widehat{\vec{d}} = \arg \max_{\vec{d} \in \Theta} \vartheta(\mathbf{P}_n(\vec{d})).$$

Suppose that observable data consist of n i.i.d. replicates of (X, Y) , denoted by $\{X_t, Y_t\}_{t=1}^n$.

We can obtain $\widehat{\vec{d}}$ through the following steps:

1. For $\vec{d} \in \Theta$, transform the continuous measures X_t 's to ordinal scales $\tilde{X}_t(\vec{d})$'s;

2. For each $\vec{d} \in \Theta$, calculate the empirical criterion function, $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$ using the data $\left\{ \left(\tilde{X}_t \left(\vec{d} \right), Y_t \right) \right\}_{t=1}^n$;
3. Search in Θ for $\vec{d}$ which maximizes $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$.

It is worth noting that, for a given dataset, $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$ is a piecewise constant function of $\vec{d}$, which only jumps at the observed values of X . As a result, the search for the maximizer of $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$ only requires the evaluation of the empirical criterion function at a finite number of $\vec{d} \in \{x_1, \dots, x_n\}^{L-1} \subset \mathbf{R}^{L-1}$, where x_j denotes the observed value of X_j ($j = 1, \dots, n$). We can also see that the set of all maximizers of $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$ takes the form of either a product of $L - 1$ left open and right closed intervals or a union of multiple such disjoint product intervals. In the large sample sense, one may choose any value in the maximizer set as the estimator of $\vec{d}_0$. In our numerical studies, we define $\widehat{\vec{d}}$ as the midpoint of the leftmost maximizer product interval (e.g. the first interval when $L = 2$). The number of solutions is defined as the number of disjoint product intervals. We regard the case where the maximizer set contains two or more disjoint product intervals as the case with multiple solutions.

2.3. Asymptotic properties and inference

While the proposed estimator $\widehat{\vec{d}}$ is conceptually simple, studying its asymptotic properties however is not trivial. The main challenge comes from the fact that $\mathbf{P}_n \left(\vec{d} \right)$ is not a smooth function of $\vec{d}$ and so is $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$. The nature of the difficulty mimics that in M-estimation when the objective function is not smooth and thus the standard linearization technique (van der Vaart and Wellner, 1996) does not work. On the other hand, $\widehat{\vec{d}}$ is not a M-estimator given that $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$, in general, is not an empirical measure of any known function. As a result, existing results on irregular M-estimation (Kim and Pollard, 1990; van der Vaart and Wellner, 1996) are not directly applicable.

To address these challenges, we take the following steps. First, we “linearize” $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$ based on the smoothness of $\mathcal{A}(\cdot)$. Next, we examined

$\vec{d}^* = \arg \max \left[\vartheta \left\{ \mathbf{P} \left(\vec{d} \right) \right\} + \vartheta^{(1)} \left\{ \mathbf{P} \left(\vec{d} \right) \right\} \left\{ \mathbf{P}_n \left(\vec{d} \right) - \mathbf{P} \left(\vec{d} \right) \right\} \right]$, the maximizer of the linear approximation of $\vartheta \left(\mathbf{P}_n \left(\vec{d} \right) \right)$. Here $\mathcal{A}^{(1)}(\cdot)$ denotes the first derivative of $\mathcal{A}(\cdot)$. From the definition of $\mathbf{P}_n \left(\vec{d} \right)$, we see that $\vec{d}^*$ is a M-estimator, for which we can use empirical process techniques to tackle its asymptotic behaviors. We find that the non-smoothness of $\vartheta \left\{ \mathbf{P} \left(\vec{d} \right) \right\} + \vartheta^{(1)} \left\{ \mathbf{P} \left(\vec{d} \right) \right\} \left\{ \mathbf{P}_n \left(\vec{d} \right) - \mathbf{P} \left(\vec{d} \right) \right\}$ in $\vec{d}$ causes a “sharp-edge effect” (Kim and Pollard, 1990; Delgado et al., 2001). As a result, the convergence rate of $\vec{d}^*$ slows to

$n^{1/3}$. The limiting distribution of $\vec{d}^*$ is not necessarily normal but a random vector that maximizes a Gaussian process. Finally, we are able to show the asymptotic equivalence between $n^{1/3}(\widehat{\vec{d}} - \vec{d}_0)$ and $n^{1/3}(\vec{d}^* - \vec{d}_0)$, and therefore the large-sample properties of $\widehat{\vec{d}}$ follow those derived for $\vec{d}^*$.

We first introduce necessary notation and regularity conditions. Define

$$\begin{aligned}\vartheta^{(1)}(\mathbf{P}) &= \frac{\partial \vartheta(\mathbf{P})}{\partial \mathbf{P}^T}, \\ \mathbf{D}(\vec{d}) &= \frac{\partial \vartheta^{(1)}(\mathbf{P}(\vec{d}))}{\partial \vec{d}}, \\ \mathbf{V}(\vec{d}) &= \frac{\partial^2 \vartheta(\mathbf{P}(\vec{d}))}{\partial \vec{d} \partial \vec{d}^T}, \\ \psi(\delta) &= \left\{ x = (x_1, \dots, x_{L^2-1})^T : \inf_{\vec{d} \in \Theta} \|x - \mathbf{P}(\vec{d})\| \leq \delta \right\},\end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm. The regularity conditions include:

- C1.** (i) $\vartheta\{\mathbf{P}(\vec{d})\}$ is twice differentiable with respect to $\vec{d}$; (ii) $\vec{d}_0$ is the unique maximizer of $\vartheta\{\mathbf{P}(\vec{d})\}$ with bounded nonsingular second-derivative matrix $\mathbf{V}(\vec{d}_0)$;
- C2.** There exists δ_0 such that $\mathcal{J}^{(1)}(\mathbf{P})$ exists and is bounded for $\mathbf{P} \in \mathcal{P}(\delta_0)$;
- C3.** The conditional density of X given $Y = l$ ($l = 1, \dots, L$), denoted by $f_{X|Y=l}(x)$, is uniformly bounded in x .

These regularity conditions pose rather mild assumptions on $\vartheta(\cdot)$ and the conditional distribution of X given $Y = l$ ($l = 1, \dots, L$). More specifically, C1(i) and C2 requires $\vartheta(\cdot)$ and the conditional distribution of X given Y be sufficiently smooth, and C3 assumes bounded conditional density functions for X given Y . These assumptions are expected to hold for many choices of $\vartheta(\cdot)$, such as those corresponding to weighted Kappa, Kendall's tau, and correct classification rate, and common continuous distributions, such as Normal distributions. Note that, C1(ii) is warranted by our formulation of the optimal cut-points.

That is, $\vec{d}_0$ is not well defined unless $\vartheta\{\mathbf{P}(\vec{d})\}$ has a unique maximizer. By C1(ii), $\vec{d}_0$ is further assumed to be an interior point of Θ . Figs. 1–3 in our empirical studies suggest the plausibility of this assumption.

Note that $\mathcal{J}(\cdot)$ is a pre-specified function in the proposed framework. When the joint distribution of X and Y is known (like in Monte-Carlo simulations), we can derive the analytic forms for $\mathbf{P}(\vec{d})$ and $\vartheta\{\mathbf{P}(\vec{d})\}$; thus we can verify conditions C1 and C2 analytically. In practice, the joint distribution of X and Y is typically unknown. Therefore, we do not expect the regularity conditions can be fully verified. A practical recommendation is to first derive reasonable parametric estimates for $\mathbf{P}(\vec{d})$ and then plug them into

$\vartheta \left\{ \mathbf{P} \left(\vec{d} \right) \right\}$ and $\vartheta^{(1)} \left\{ \mathbf{P} \left(\vec{d} \right) \right\}$. This would help evaluate whether a selected $\vartheta(\cdot)$ meets the required technical assumptions.

We state the asymptotic properties of $\widehat{\vec{d}}$ in the following theorems. Detailed proofs are provided in Appendix.

Theorem 1—Under the regularity conditions C1–C2, the proposed cut-point estimator, $\widehat{\vec{d}}$, is consistent. That is, $\widehat{\vec{d}} \rightarrow_P \vec{d}_0$ as $n \rightarrow \infty$.

Theorem 2—Under the regularity conditions C1–C3, $n^{1/3} \left\{ \widehat{\vec{d}} - \vec{d}_0 \right\}$ converges in distribution to the unique maximizer of the stochastic process, $G(\mathbf{h}) + \frac{1}{2} \mathbf{h}^\top \mathbf{V}(\vec{d}_0) \mathbf{h}$. Here $G(\mathbf{h})$ is a zero-mean Gaussian process of continuous sample paths, and satisfies (A.3) in Appendix A.3.

Our theoretical investigations show that $n^{1/3} \left(\widehat{\vec{d}} - \vec{d}_0 \right)$ converges to a tight limit distribution, which however does not have a closed form. Given the non-standard asymptotics, using conventional bootstrap for inference is questionable (Abrevaya and Huang, 2005; Kosorok, 2008). We propose to adopt random subsampling (Politis et al., 1999) that uses without-replacement subsamples instead of with-replacement bootstrap samples to approximate the distribution of $n^{1/3} \left(\widehat{\vec{d}} - \vec{d}_0 \right)$. More specifically, let superscript (j) denote the j th component of a vector. The subsampling-based inference can be carried out as follows.

Step 1: Choose a subsample size b . Set $s = 1$.

Step 2: Randomly draw a subsample of size b without replacement from $\{X_t, Y_t\}_{t=1}^n$. Apply the proposed method to estimate $\vec{d}_0$ based on this random subsample. Denote the resulting estimator by $\widehat{\vec{d}}_{(s)}$.

Step 3: Increase s by 1. If s is less than a prespecified large number S , then go back to Step 2.

Step 4: Compute the empirical variance of $\left\{ \widehat{\vec{d}}_{(s)}^{(j)} \right\}_{s=1}^S$, which provides a variance estimator for $\widehat{\vec{d}}^{(j)}$.

Step 5: Compute the empirical $100(1 - \alpha)$ th percentile of $\left\{ \left| \widehat{\vec{d}}_{(s)}^{(j)} - \widehat{\vec{d}}^{(j)} \right| \right\}_{s=1}^S$, denoted by $\nu_{\alpha}^{(j)}$. The $100(1 - \alpha)$ confidence interval for $\vec{d}_0^{(j)}$ can be constructed by $\left[\widehat{\vec{d}}^{(j)} - \nu_{\alpha}^{(j)}, \widehat{\vec{d}}^{(j)} + \nu_{\alpha}^{(j)} \right]$.

The validity of the above subsampling procedure follows from the results of Politis and Romano (1994), coupled with our Theorem 2, which implies $n^{1/3} \left(\widehat{\vec{d}} - \vec{d}_0 \right)$ converges weakly to a limit distribution. Note that the subsample size s is subject to the theoretical constraints, $s \rightarrow \infty$ as $n \rightarrow \infty$ and $s = o(n)$. Discussions about practical selection of s can be found in Politis and Romano (1994) and Delgado et al. (2001).

Other modifications of the conventional bootstrap method, such as m out of n bootstrap and smooth bootstrap, have also been investigated in various estimation settings with cubic root convergence (Lee and Pun, 2006; Léger and MacGibbon, 2006; Sen et al., 2010; Sen and Xu, 2015, among others). These methods can be adapted to make inference about $\vec{d}_0$ under stronger assumptions. For example, by the results of Léger and MacGibbon (2006), resampling without replacement from a smooth and symmetric estimator of $f_{X|Y=l}(x)$ ($l = 1, \dots, L$) can lead to a consistent bootstrap procedure when $f_{X|Y=l}(x)$ is a symmetric function of x . Such a smooth bootstrap procedure is described in detail in Appendix A.3. In practice, it can serve as a useful alternative inference procedure for $\vec{d}_0$ when real data suggest symmetric conditional distributions of X given $Y = l$ ($l = 1, \dots, L$).

3. Simulation study

We conducted extensive simulations to evaluate the proposed method for determining cut-points in a continuous scale. We considered criterion functions constructed based on weighted kappa with linear weight (ϑ_{κ}), Kendall's τ_b (ϑ_{τ}), and correct classification rate (ϑ_{CC}). The cut-point estimators corresponding to these criterion functions are denoted by $\widehat{\vec{d}}_{\kappa}$, $\widehat{\vec{d}}_{\tau}$ and $\widehat{\vec{d}}_{CC}$ respectively.

We first examined the situation with $L = 2$, where the cut-point to estimate is a scalar. We generated Y as a binary random variable taking values 3 and 4 with equal probabilities. Conditional on Y , the continuous X was obtained as $X = 2Y + \varepsilon$, where ε follows $N(0, \sigma^2)$ distribution. We set $\sigma = 0.6$ and 1.0 to reflect high-to-moderate and moderate-to-low separation in X between different Y groups. In Fig. 1, we demonstrate the three criterion functions as well as the corresponding empirical criterion functions computed based on one simulated dataset with sample size $n = 100$. It is shown that the optimal cut-point based on ϑ_{κ} , ϑ_{τ} or ϑ_{CC} is $d_0 = 7$. All criterion functions are unimodal. The maxima of a criterion function is less “prominent” with a larger σ , which means less separation in X among different Y categories. The criterion functions, ϑ_{κ} and ϑ_{τ} have steeper curvatures around their maxima than ϑ_{CC} . It is also observed that the empirical criterion functions approximate the true criterion function quite well with a moderate sample size $n = 100$.

We conducted 1000 Monte-Carlo simulations under each set-up with sample sizes $n = 50, 100$, and 200 . For subsampling, we set $S = 100$, and selected b roughly as $n^{0.7}$. That is, $b = 15, 25$, and 40 respectively when $n = 50, 100$, and 200 . We evaluated the empirical bias, empirical standard deviations, and average estimated standard deviations of the proposed estimator, and empirical coverage probabilities of the proposed 95% confidence intervals. In case of multiple solutions, we selected our estimator as the midpoint of the leftmost maximizer interval (e.g. the first interval when $L = 2$). From Table 1, we observe that cut-point estimation based on criterion, weighted-kappa or Kendall's τ_b , may perform better than that based on corrected classification rate. The cut-point estimators, $\widehat{d}_\kappa$ and $\widehat{d}_\tau$, have small bias even with a sample size as small as $n = 50$. In comparison, $\widehat{d}_{cc}$ seems to have considerably larger bias. For all three cut-point estimators, the estimated standard deviations are fairly close to the empirical standard deviations, and their agreement improves with sample size n . As expected, the standard deviations increase with σ , reflecting the elevated estimation variability with more overlap in X between Y categories. The 95% confidence intervals have pretty accurate coverage probabilities.

We also examined cases with $L = 3$, where $\vec{d}_0$ is a 2×1 vector. The ordinal Y was generated as multinomial random variable that takes values of 3, 4, and 5 with equal probabilities. The continuous X was obtained based on the same equation used for the cases with $L = 2$. Based on 1000 simulations, we summarize in Table 2 the results in the same format as in Table 1. Under each set-up, the first row corresponds to the smaller cut-point in X and the second row corresponds to the larger cut-point. In Table 2, we observe larger empirical bias and standard deviations compared to those in Table 1. The estimator $\widehat{d}_{cc}$ again exhibits the largest bias. For all three estimators, the estimated standard deviations and empirical standard deviations still match pretty well, and the empirical coverage probabilities are close to the nominal level.

It is important to note from Tables 1–2 that the standard deviations of each estimator decrease with sample size at a rate that conforms to the theoretical convergence rate of $n^{1/3}$ stated in Theorem 2. More specifically, by assessing the ratios of empirical standard deviations between two different sample sizes (e.g. 100 versus 200 or 50 versus 200), we find that these ratios are generally around the cube root of the corresponding sample size ratios.

We also evaluated the performance of the alternative smooth bootstrap inference procedure (Léger and MacGibbon, 2006) described in Appendix A.3. Note that, the conditional distributions of X given $Y = l$ are normal distributions in all simulation set-ups, and hence the smooth bootstrap approach is expected to yield consistent inference. We present the additional simulation results in Tables 3 and 4, parallel to Tables 1 and 2 respectively.

Our simulation results demonstrate satisfactory performance of the smooth bootstrap procedure. The smooth bootstrap procedure gives standard error estimates quite close to empirical standard errors. The resulting confidence intervals can be slightly under-covered when the sample size is small, and the coverage probabilities become closer to the nominal

level as the sample size increases. We further compare the lengths of 95% confidence intervals between the subsampling method and the smooth bootstrap approach. Our simulation results (not reported here) show that subsampling based confidence intervals tend to be wider than the smooth bootstrap based confidence intervals. This is consistent with the observation from Tables 1–4 that the smooth bootstrap approach often yields coverage probabilities lower than the nominal level while the opposite trend holds for the subsampling method. In addition, our numerical experiences suggest that the subsampling procedure is computationally faster. For example, for a set-up with $\sigma = 0.6$ and $n = 100$, computation time ratio between the smooth bootstrap procedure and the subsampling procedure is 4.2 when $L = 2$, and the ratio is increased to 9.5 when $L = 3$. The shorter computation time and the wider confidence intervals with the subsampling approach may be explained by the smaller sample size used in each resampling step of the subsampling procedure.

As discussed in Section 2.2, an empirical criterion function usually has more than one maximizers that form a connected interval (i.e. one solution) or multiple disjoint intervals (i.e. multiple solutions). While all criterion functions considered in the simulations are unimodal, multiple solutions can occur with the finite sample estimation of optimal cut-points. In our simulation studies, we also investigated the extent to which the multiple solution problem exists. We report in Table 5 the average number of solutions (ANS) as well as the proportion of simulations that produced multiple solutions (PMS). It is shown that adopting correct classification rate as the criterion for determining cut-points may be prone to the most severe multiple solution problems as compared to the other two criteria, weight kappa and Kendall's τ_b . For example, when $L = 2$, the percentages of getting multiple solutions from using ϑ_{CC} range from 30% to 40%, which are much higher than those from ϑ_κ , around 3%–4%, or those from ϑ_τ about 1%–3%. While the frequency of encountering multiple solutions seems to decrease as L increases from 2 to 3 when weighted kappa or Kendall's τ_b was used as the criterion for cut-point determination, it appears to be a reversed situation for ϑ_{CC} . When $L = 3$, multiple solutions were encountered in over 50% simulations in most set-ups with ϑ_{CC} . In contrast, using ϑ_κ or ϑ_τ only resulted in less than 1% simulations with multiple solutions. One possible explanation for these observations may relate to the observed flatter curvature around $\vec{d}_0$ in $\vartheta_{CC}(\vec{d})$ than that in $\vartheta_\kappa(\vec{d})$ or $\vartheta_\tau(\vec{d})$. The flatter curvature in ϑ_{CC} around maxima may suggest more ambiguities in the identification of optimal cut-points with finite sample sizes, and consequently lead to the larger bias of $\widehat{\vec{d}}_{CC}$ observed in Tables 1–2 (or 3–4) and more frequent occurrences of multiple solutions observed in Table 5.

4. Diabetes and depression study

Diabetes and Depression study was conducted to determine the prevalence of depression among African-Americans with type II diabetes and its association with socioeconomic determinants, adherence to treatment and glycemic control. There are 1430 African American diabetes patients in the study. Two psychometric instruments were administered for each patient: the MINI diagnostic instrument (Sheehan et al., 1998) and the Zung rating scale (Zung, 1965). In order to develop an efficient diagnosis technique to identify major

depression in a large population of minority diabetic patients, we seek to determine the cut-points of Zung rating scale for measuring the severity of the depression. Of note is that the MINI, a structured psychiatric interview to diagnose the syndrome of major depression with well-established graded severity guidelines, is designed to be used by licensed professionals or well-trained interviewers (Sheehan et al., 1998) and is time-consuming. In contrast, the Zung rating scale, a short self-administered survey with 20 items to quantify depressive symptoms with total scores ranging from 25 to 100, requires only 5–10 min to complete and does not require a high level of literacy. In this dataset, the average and median Zung scores are 45.4 and 43.8 respectively, and the inter-quartile range for Zung is (35.0, 55.0).

It has been recognized that physical illness of diabetes patients may falsely elevated the scores of certain items in scales especially because of those related to physical symptoms including fatigue, cognitive dysfunction. First, we consider the classification of patients into two categories by combining mild depressed patients with no depression patients as a non-major depression group and the other patients as a major depression group. According to the MINI, 1112 (77.7%) subjects have no or mild depression, and 318 (22.3%) subjects are moderately or severely depressed. Fig. 2 displays three criterion functions based on weighted kappa, Kendall's τ_b , and correct classification rate versus all possible cut-points for dichotomizing Zung rating scale according to the MINI depression status. It can be seen that the empirical criterion functions can be approximated by a smooth concave curve where the maximum is achieved when the cut-point is around 57 on the Zung rating scale.

When weighted kappa (with linear weight) is adopted as the criterion function, the estimated cut-point for the Zung scale is 57.0 ($SE = 1.43$) with 95% CI (53.94, 60.13). Here and hereafter, SE and CI stand for abbreviations for standard error and confidence interval respectively. The corresponding maximum weighted kappa is 0.64. The criterion function based on Kendall's tau provides the same estimate of 57.0 with a slightly larger $SE = 1.60$ and wider 95% CI, (53.49, 60.58). Using the empirical criterion function based on correct classification rate, the estimated cut-point is 57.5 ($SE = 1.63$) with 95% CI, (53.94, 60.13). Overall, the three different criterion functions consistently suggest using a cut-point around 57 in Zung rating scale to differentiate no or minor depression versus major depression.

Next, we consider the classification of the Zung scale into three categories: no depression, mild depression, and moderate or severe depression. According to the MINI, 1043 (72.9%) subjects have no depression, 69 (4.8%) are in mild depression and 318 (22.3%) are moderately or severely depressed. In Fig. 3, the three three-dimensional plots depict the empirical criterion functions defined on weighted kappa, Kendall's τ_b , and correct classification rate, respectively. It is clear that the empirical criterion function based on weighted kappa demonstrates the most desirable profile for identifying maximizers. Using weighted kappa (with linear weight) as the criterion for determining cut-points, we estimate the first cut-point in Zung separating no depression versus mild depression by 55.7 ($SE = 1.42$) with 95% CI (52.67, 58.84) and estimate the second one, separating mild versus moderate or severe depression, by 57.0 ($SE = 1.15$) with 95% CI (54.54, 59.53). Based on Kendall's τ_b , the first cut-point estimate is 55.7 ($SE = 1.99$) with 95% CI (50.86, 60.65) and the second one is 63.1 ($SE = 6.07$) with 95% CI (44.74, 69.33). Adopting the criterion function based on correct classification rate, we obtain the first cut-point estimate, 56.4 (SE

= 1.18) with 95% CI, (53.28, 58.23); and the second one, 57.0 ($SE = 1.20$) with 95% confidence interval, (54.54, 59.53).

We note that the second cut-point estimate based on Kendall's τ_b is quite different from those based on weighted kappa or correct classification rate, and also has a pretty large standard error. This may be explained by the bi-modal trend in the second cut-point demonstrated by the empirical criterion function based on Kendall's τ_b (see Fig. 3). Such a bi-modal feature could have caused the instability in the estimation of the second cut-point. The estimated second cut-point based on weighted kappa or correct classification rate is in a close proximity of our finding from the case, where only one cut-point is considered for separating no/minor depression from major depression. We also notice that by all methods, the estimated second cut-point is quite close to the estimated first cut-point. Such a result is consistent with the low prevalence rate of mild depression in diabetic subjects, around 4.8%, suggested by the MINI scale of this dataset. The relatively small sample size for the mild depression category, 69, may explain the observed overlaps between the confidence intervals for the first cut-point and those for the second cut-point.

Based on a national survey, Zung (1973) presented the following guidelines: patients with no depression yields score indices of between 25 to 49, patients with global ratings of mild to moderate depressions have indices of 50 to 59 and those with moderate to severe depressions have indices of 60 and over (Zung, 1973). Such rules implicate the two cut-points in the continuous Zung scale, 49 and 59. These values are roughly in the range with our estimated cut-point values. Using the dataset from the Diabetes and Depression study, the weighted kappa, Kendall's τ_b and correct classification rate resulted from applying such empirical cut-points, $\vec{d} = (49, 59)$, are 0.61, 0.65, and 0.76 respectively. Adopting the proposed cut-point estimator from maximizing weighted kappa, $\widehat{\vec{d}}_\kappa = (55.7, 57)$, these agreement or association statistics are higher, equal to 0.65, 0.66, and 0.84 respectively.

When comparing $\widehat{\vec{d}}_\kappa$ with the current empirical cut-points, we find that our results do not indicate a significant modification to the Zung cut-point for differentiating minor depression from major depression among diabetic subjects. However our analysis may suggest a shifting-up of the first cut-point that separates no depression from mild depression. This probably reflects a reasonable adjustment in diagnosing minor depression in subjects with chronic disease, as some mild self-reported depressive symptoms may due to medical illness. Such a finding may contribute a refinement of existing diagnosis rules for mental health disorders in medically unhealthy populations.

5. Discussion

In this work, we propose a general and rigorous framework for determining cut-points in a continuous scale according to an ordinal outcome. We formulate our approach as a problem of maximizing a predetermined criterion function of cut-points. Our general specification of criterion functions can accommodate many common agreement or association measures. It also allows for considerations about the relative costs of different types of classification errors and the chance of making these different errors, for example, via formulating the criterion function based on weighted correct classification rate.

Our theoretical studies uncover nonstandard asymptotics attached to the proposed method. Similar nonstandard results also arise in other problem settings, such as classification tree. We adapt the inference procedures accordingly, identifying subsampling as a valid technique for obtaining variance estimation and confidence intervals. Our simulation studies demonstrate appreciable finite sample performance of the proposed method. Via an application to a real mental health dataset, we demonstrate the high potential of the proposed method to become a sensible, flexible, and well justified analytic tool for categorizing a continuous scale in practice.

Acknowledgments

This research project was supported by grants from National Institute of Health (R01MH079448 and R01HL113548). We thank Dr. Musselman for discussions related to Diabetes and Depression study.

Appendix. Regularity conditions and proofs of Theorems 1–2

A.1. Proof of Theorem 1

Applying Taylor expansion, we have

$$|\vartheta \{P_n(\vec{d})\} - \vartheta \{P(\vec{d})\}| = |\vartheta^{(1)} \{\check{P}(\vec{d})\} \{P_n(\vec{d}) - P(\vec{d})\}| \quad (\text{A.1})$$

where $\check{P}(\vec{d})$ is some vector on the between $P_n(\vec{d})$ and $P(\vec{d})$. By the definition of $P_n(\vec{d})$ and the Glivenko-Cantelli theorem (van der Vaart and Wellner, 1996), we have $\sup_{\vec{d} \in \Theta} \|P_n(\vec{d}) - P(\vec{d})\| \rightarrow_P 0$. Under condition C2, $\sup_{\vec{d} \in \Theta} \vartheta^{(1)}(\check{P}(\vec{d}))$ is bounded as n is large enough, and thus we have

$$\sup_{\vec{d} \in \Theta} |\vartheta \{P_n(\vec{d})\} - \vartheta \{P(\vec{d})\}| \rightarrow_P 0. \quad (\text{A.2})$$

By the Argmax continuous mapping theorem (Theorem 3.2.2 of van der Vaart and Wellner (1996)), (A.2) and condition C1 indicate that $\widehat{\vec{d}} \rightarrow_P \vec{d}_0$. That is, $\widehat{\vec{d}}$ is a consistent estimator of $\vec{d}_0$.

A.2. Proof of Theorem 2

For a vector $\vec{d}$, let $\vec{d}^{(k)}$ denote its k th component. Define

$$m_{\vec{d}}(x, y) = \vartheta \{P(\vec{d})\} + \vartheta^{(1)} \{P(\vec{d})\} \cdot \left(I(x < \vec{d}^{(1)}, y=1), \dots, I(x < \vec{d}^{(1)}, y=L), I(\vec{d}^{(1)} \leq x < \vec{d}^{(2)}, y=1) \right. \\ \left. - \vartheta^{(1)} \{P(\vec{d})\} \cdot P(\vec{d}) \right)$$

. Note that, although the function $m_{\vec{d}}(\cdot)$ is not Lipschitz in the parameter $\vec{d}$, according to condition C1, $E(m_{\vec{d}}(X, Y))$, which equals $\vartheta \{P(\vec{d})\}$, is twice differentiable at the unique maximizer $\vec{d}_0$ with nonsingular second derivative matrix $V(\vec{d}_0)$.

Let $\vec{d}^* = \arg \max \left[\vartheta \left\{ \mathbf{P}(\vec{d}) \right\} + \vartheta^{(1)} \left\{ \mathbf{P}(\vec{d}) \right\} \left\{ \mathbf{P}_n(\vec{d}) - \mathbf{P}(\vec{d}) \right\} \right]$. By the definition of $\mathbf{P}_n$ and $n_{\vec{d}}(\cdot)$, it is equivalent to define $\vec{d}^*$ as $\vec{d}^* = \arg \max_{\vec{d} \in \Theta} \sum_{i=1}^n m_{\vec{d}}(X_i, Y_i)$. Therefore, we can view $\vec{d}^*$ as a standard M-estimator while treating $\mathbf{P}(\cdot)$ as a known function. Our basic idea to prove Theorem 2 is to first study the convergence rate and the asymptotic distribution of $\vec{d}^*$ and then establish the large sample properties of $\widehat{\vec{d}}$ based on the connection between $\vec{d}^*$ and $\widehat{\vec{d}}$.

For a given $\delta > 0$, define a function class $\mathcal{M}_\delta = \left\{ m_{\vec{d}} - m_{\vec{d}_0} : \|\vec{d} - \vec{d}_0\| < \delta \right\}$. It can be shown that the function class $\mathcal{M}_\delta$ is a VC (van der Vaart and Wellner, 1996) with the envelope function, which under conditions C1–C3, takes the form,

$$V_\delta(x, y) \equiv C_1 \delta + C_2 \left[I(x \in A_{1,\delta}, y=1) + \cdots + I(x \in A_{1,\delta}, y=L) + \{I(x \in A_{1,\delta}, y=1) + I(x \in A_{2,\delta}, y=1)\} \cdots + \{I(x \in A_{1,\delta}, y=L) + I(x \in A_{2,\delta}, y=L)\} \right]$$

where C_1 and C_2 are positive constants, and $A_{k,\delta} = \left[\vec{d}_0^{(k)} - \delta, \vec{d}_0^{(k)} + \delta \right]$, $k=1, \dots, L-1$.

As δ approaches zero, by condition C3, $\sqrt{E \left\{ V_\delta(X, Y)^2 \right\}}$ is bounded above by a constant times $\sqrt{\delta}$. Thus, the conditions of Theorem 3.2.10 of van der Vaart and Wellner (1996) are satisfied with $\phi(\delta) = C_3 \sqrt{\delta}$ for a constant C_3 . This leads to a rate of convergence $n^{1/3}$ for $\vec{d}^* - \vec{d}_0$. Moreover, Theorem 3.2.10 of van der Vaart and Wellner (1996) implies that $n^{1/3}(\vec{d}^* - \vec{d}_0)$ converges in distribution to the maximizer for the process

$M(\mathbf{h}) \equiv G(\mathbf{h}) + \frac{1}{2} \mathbf{h}^T \mathbf{V}(\vec{d}_0) \mathbf{h}$, where $G(\mathbf{h})$ is a zero-mean Gaussian process, which has continuous sample paths and satisfies

$$E \left[\{G(\mathbf{g}) - G(\mathbf{h})\}^2 \right] = \sum_{k=1}^{L^2-1} \left\{ a_k(\vec{d}_0) \right\}^2 b_k(\vec{d}_0, \mathbf{g}, \mathbf{h}). \quad (\text{A.3})$$

Here $a_k(\vec{d}_0)$ is the k th component of

$$\begin{aligned} \vartheta^{(1)}(\vec{d}_0), b_{(s-1)L+t}(\vec{d}_0, \mathbf{g}, \mathbf{h}) &= f_{X|Y=t}(\vec{d}_0^{(s)}) |g^{(s)} - h^{(s)}| \text{ when } s=1 \text{ and } t \in \{1, \dots, L\}, \\ b_{(s-1)L+t}(\vec{d}_0, \mathbf{g}, \mathbf{h}) &= f_{X|Y=t}(\vec{d}_0^{(s-1)}) |g^{(s-1)} - h^{(s-1)}| + f_{X|Y=t}(\vec{d}_0^{(s)}) |g^{(s)} - h^{(s)}|, \text{ when } \\ s &\in \{2, \dots, L-1\} \text{ and } t \in \{1, \dots, L\}, \text{ and } b_{(s-1)L+t}(\vec{d}_0, \mathbf{g}, \mathbf{h}) &= f_{X|Y=t}(\vec{d}_0^{(s)}) |g^{(s)} - h^{(s)}| \\ \text{when } s &= L \text{ and } t \in \{1, \dots, L-1\}. \end{aligned}$$

Next, we shall complete the proof by showing the asymptotic equivalence between

$$n^{1/3}(\widehat{\vec{d}} - \vec{d}_0) \text{ and } n^{1/3}(\vec{d}^* - \vec{d}_0). \text{ Define}$$

$$M_n^*(\mathbf{h}) = \vartheta \left\{ \mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\} + \vartheta^{(1)} \left\{ \mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\} \cdot \left\{ \mathbf{P}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) - \mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\}.$$

Let $\hat{h}_n^* = n^{1/3} \left(\vec{d}^* - \vec{d}_0 \right)$. By the definition of $\vec{d}^*$, $\hat{h}_n^*$ is the maximizer of the stochastic process $M_n^*(\mathbf{h}) - M_n^*(\mathbf{0})$. The fact that $m_{\vec{d}}$ satisfies the conditions of Theorem 3.2.10 of van der Vaart and Wellner (1996) implies that $n^{2/3} \{M_n^*(\mathbf{h}) - M_n^*(\mathbf{0})\}$ is asymptotically tight in $\ell^\infty(\mathbf{h}; \|\mathbf{h}\| < K)$ and weakly converges to the Gaussian process $M(\mathbf{h})$ for every K .

Define $\tilde{M}_n(\mathbf{h}) = \vartheta \left\{ \mathbf{P}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\}$. Let $\hat{h}_n = n^{1/3} \left(\widehat{\vec{d}} - \vec{d}_0 \right)$, $\epsilon_n = n^{2/3} \left[\left\{ \tilde{M}_n(\hat{h}_n) - \tilde{M}_n(\mathbf{0}) \right\} - \left\{ M_n^*(\hat{h}_n) - M_n^*(\mathbf{0}) \right\} \right]$, and $\epsilon_n^* = n^{2/3} \left[\left\{ \tilde{M}_n(\hat{h}_n^*) - \tilde{M}_n(\mathbf{0}) \right\} - \left\{ M_n^*(\hat{h}_n^*) - M_n^*(\mathbf{0}) \right\} \right]$. Applying Taylor's expansion to $\vartheta \left\{ \hat{\mathbf{P}}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\}$ around $\mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right)$, we get

$$\tilde{M}_n(\mathbf{h}) = M_n^*(\mathbf{h}) + \frac{1}{2} \left\{ \mathbf{P}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) - \mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\}^\top \vartheta^{(2)} \left(\mathbf{P}^\dagger \right) \left\{ \mathbf{P}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) - \mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right) \right\},$$

where $\vartheta^{(2)}(\mathbf{P}) = \frac{\partial^2 \vartheta(\mathbf{P})}{\partial \mathbf{P} \partial \mathbf{P}^\top}$, and $\mathbf{P}^\dagger$ is between $\mathbf{P}_n \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right)$ and $\mathbf{P} \left(\vec{d}_0 + n^{-1/3} \mathbf{h} \right)$. By the Donsker theorem (van der Vaart and Wellner, 1996), we have

$\sup_{\vec{d} \in \Theta} n^{1/3} \left\| \mathbf{P}_n \left(\vec{d} \right) - \mathbf{P} \left(\vec{d} \right) \right\| \rightarrow_p 0$. Therefore, under condition C1(ii),

$\sup_{\mathbf{h} \in D_\Theta} n^{2/3} \left\{ \tilde{M}_n(\mathbf{h}) - M_n^*(\mathbf{h}) \right\} = o_p(1)$ where $D_\Theta = \left\{ \mathbf{h}: \vec{d}_0 + n^{-1/3} \mathbf{h} \in \Theta \right\}$. This immediately implies that $\epsilon_n = o_p(1)$ and $\epsilon_n^* = o_p(1)$. Therefore,

$$o_p(1) = \epsilon_n^* - \epsilon_n = n^{2/3} \left[M_n^*(\hat{h}_n) - M_n^*(\hat{h}_n^*) + \vartheta \left\{ \mathbf{P}_n \left(\vec{d}^* \right) \right\} - \vartheta \left\{ \mathbf{P}_n \left(\widehat{\vec{d}} \right) \right\} \right].$$

This implies $n^{2/3} \left\{ M_n^*(\hat{h}_n) - M_n^*(\mathbf{0}) \right\} \geq n^{2/3} \left\{ M_n^*(\hat{h}_n^*) - M_n^*(\mathbf{0}) \right\} - o_p(1)$ because

$\vartheta \left\{ \mathbf{P}_n \left(\vec{d}^* \right) \right\} - \vartheta \left\{ \mathbf{P}_n \left(\widehat{\vec{d}} \right) \right\} \leq 0$ and $M_n^*(\hat{h}_n) - M_n^*(\hat{h}_n^*) \leq 0$. It then follows from

Argmax continuous mapping theorem (van der Vaart and Wellner, 1996) that $\hat{h}_n$, like $\hat{h}_n^*$, converges in distribution to the unique maximizer of the stochastic process $M(\mathbf{h})$. This completes the proof of Theorem 2.

A.3. A smooth bootstrap inference procedure

A smooth bootstrap inference procedure adapted from Léger and MacGibbon's (2006) work is described as follows.

Step 1: Let $\tilde{X}_{l,k}$ denote the k th observation of X given $Y = l$ ($j = 1, \dots, n_l$). Estimate the conditional density functions, $f_{X|Y=l}$ ($l = 1, 2, \dots, L$), by the following smooth and symmetric kernel based estimators,

$$\hat{f}_{X|Y=l}(x) = \frac{1}{2n_l h_l} \left[\sum_{k=1}^{n_l} K\left(\frac{x - \tilde{X}_{l,k}}{h_l}\right) + \sum_{k=1}^{n_l} K\left(\frac{x + \tilde{X}_{l,k} - 2\hat{\theta}_l}{h_l}\right) \right],$$

where $K(\cdot)$ is the Gaussian kernel function, h_l is a smoothing parameter standing for bandwidth, $\hat{\theta}_l$ is the sample median of X given $Y = l$. We proceed bandwidth selection as follows:

- i. Following Sen et al.'s (2010) rule of thumb, we start with an initial bandwidth, $h_0 = 0.9An^{-1/n}$, with $A = \min\{s, IQR/1.34\}$, where s and IQR are the sample standard deviation and inter-quartile of $\{\tilde{X}_{l,k}, k=1, 2, \dots, n_l\}$.
- ii. We evaluate a sequence of candidate bandwidth values in the neighborhood of h_0 , say $h_0 - 0.1, h_0 - 0.05, h_0, h_0 + 0.05, h_0 + 0.1$, based on the integrated least square cross-validation criterion (Sheather, 2004),

$$LSCV(h) = \int \left\{ \hat{f}_{X|Y=l}(x) \right\}^2 dx - \frac{2}{n_l} \sum_{k=1}^{n_l} \hat{f}_{X|Y=l}^{-(l,k)}(\tilde{X}_{l,k}),$$

where $\hat{f}_{X|Y=l}^{-(l,k)}(\cdot)$ denotes the kernel density estimator constructed from the data given $Y = l$ excluding $\tilde{X}_{l,k}$.

- iii. We set the bandwidth as $h_{l,opt} = \operatorname{argmin}_{h_l} [LSCV(h_l)]$.

Step 2: Set $s = 1$.

Step 3: Randomly select $Y = l$ from $\{Y_1, \dots, Y_n\}$. Given $Y = l$, we resample X based on the estimated density function $\hat{f}_{X|Y=l}(x)$ using Acceptance-Rejection method. Repeat this procedure for n times to obtain a bootstrap sample of size n .

Step 4: Apply the proposed method to estimate $\vec{d}_0$ based on the bootstrap sample obtained from Step 3. Denote the resulting estimator by $\widehat{\vec{d}}_{(s)}$ and increase s by 1. Go back to Step 3 unless $s > S$.

Step 5: Compute the empirical variance of $\left\{ \widehat{\vec{d}}_{(s)}^{(j)} \right\}_{s=1}^S$, which provides a variance estimate for $\vec{d}^{(j)}$.

Step 6: Compute $\hat{\vec{d}}^{(j)} = \sum_{s=1}^S \hat{\vec{d}}_{(s)}^{(j)} / S$ and then obtain the empirical $100(\alpha/2)$ th and $100(1 - \alpha/2)$ th percentile of $\left\{ n^{1/3} \left(\hat{\vec{d}}_{(s)}^{(j)} - \hat{\vec{d}}^{(j)} \right) \right\}_{s=1}^S$, denoted by $\nu_{\alpha/2}^{(j)}$ and $\nu_{1-\alpha/2}^{(j)}$ respectively. The $100(1 - \alpha)\%$ (basic) confidence interval for $\vec{d}_0^{(j)}$ can be constructed by

$$\left[\hat{\vec{d}}^{(j)} - n^{-1/3} \nu_{1-\alpha/2}^{(j)}, \hat{\vec{d}}^{(j)} - n^{-1/3} \nu_{\alpha/2}^{(j)} \right].$$

The consistency of the above smooth bootstrap procedure follows from Theorem 1 and Corollary 1 of Léger and MacGibbon (2006).

References

- Abrevaya J, Huang J. On the bootstrap of the maximum score estimator. *Econometrica*. 2005; 73:1175–1204.
- Agresti, A. *Categorical Data Analysis*. Wiley; New York: 1990.
- Altman DG. Categorizing continuous variables. *Br. J. Cancer*. 1991; 64:975. [PubMed: 1931629]
- Altman DG, Lausen B, Sauerbrei W. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *J. Natl. Cancer Inst.* 1994; 86(11):829–835. [PubMed: 8182763]
- Baughman AL, Bisgard KM, Lynn F, Meade BD. Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels. *Stat. Med.* 2006; 25(17):2994–3010. [PubMed: 16345022]
- Brownie C, Habicht JP. Selecting a screening cut-off point or diagnostic criterion for comparing prevalences of disease. *Biometrics*. 1984; 40(3):675–684. [PubMed: 6518245]
- Chernoff H. Estimation of the mode. *Ann. Inst. Math. Stat.* 1964; 16:31–41.
- Cohen J. A coefficient of agreement for nominal scales. *Educ. Psychol. Measure.* 1960; 20(1):37–46.
- Delgado M, Rodriguez-Poo J, Wolf M. Subsampling inference in cube-root asymptotics with an application to manski's maximum score estimator. *Econom. Lett.* 2001; 73:241–250.
- James IR. Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. *Biometrics*. 1978; 34:265–275. [PubMed: 667273]
- Kendall MG. A new measure of rank correlation. *Biometrika*. 1938; 30(1–2):81–93.
- Kim J, Pollard D. Cube root asymptotics. *Ann. Statist.* 1990; 18:191–210.
- Kosorok, M. *Introduction to Empirical Process and Semiparametric Inference*. Springer; New York: 2008.
- Kraemer H. Assessment of 2×2 associations: Generalization of signal-detection methodology. *Amer. Statist.* 1988; 42:37–49.
- Lee SMS, Pun MC. On m out of n bootstrapping for nonstandard M-estimation with nuisance parameters. *J. Amer. Statist. Assoc.* 2006; 101:1185–1197.
- Léger C, MacGibbon B. On the bootstrap in cube root asymptotics. *Canad. J. Statist.* 2006; 34:29–44.
- Mazumdar M, Glassman JR. Categorizing a prognostic variable: Review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat. Med.* 2000; 19:113–132. [PubMed: 10623917]
- Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic disease, and decrements in health: results from the world health surveys. *The Lancet*. 2007; 370:851–858.
- O'Brien SM. Cutpoint selection for categorizing a continuous predictor. *Biometrics*. 2004; 60:504–509. [PubMed: 15180677]
- Pepe, M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press Inc.; New York: 2003.

- Perkins N, Schisterman E. The inconsistency of optimal cut-points using two roc based criteria. *Amer. J. Epidem.* 2006; 163:67–675.
- Politis D, Romano J. Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* 1994; 22:2031–2050.
- Politis, D.; Romano, J.; Wolf, M. *Subsampling*. Springer; New York: 1999.
- Schisterman E, Perkins N, Liu A, Bondell H. Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. *Epidemiology.* 2005; 16:73–81. [PubMed: 15613948]
- Sen B, Banerjee M, Woodroffe M. Inconsistency of bootstrap: the Grenander estimator. *Ann. Statist.* 2010; 38:1953–1977.
- Sen B, Xu G. Model based bootstrap methods for interval censored data. *Comput. Statist. Data Anal.* 2015; 81:121–129.
- Sheather SJ. Density estimation. *Statist. Sci.* 2004; 19:588–597.
- Sheehan D, Janavs J, Baker R, Harnett-Sheehan K, Knapp E, Sheehan M, Lecrubier Y, Weiller E, Hergueta T, Amorim P, Bonora LI, Lépine JP. M.i.n.i.-mini international neuropsychiatric interview-english version 5.0.0-dsm-iv. *J. Clinial Psychiartry.* 1998; 59:34–57.
- Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC. The mini-international neuropsychiatric interview (m.i.n.i.): the development and validation of a structured diagnostic psychiatric interview for d sm-iv and icd-10. *J. Clinial Psychiartry.* 1998; 59(S20):22–33.
- van der Vaart, A.; Wellner, J. *Weak Convergence and Empirical Processes: With Applications in Statistics*. Springer-Verlag; New York: 1996.
- Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950; 3:32–35. [PubMed: 15405679]
- Zung WWK. A self-rating depression scale. *Archiv. Gen. Psychiartry.* 1965; 12:63–70.
- Zung WWK. From art to science: The diagnosis and treatment of depression. *Archiv. Gen. Psychiartry.* 1973; 29:328–337.

Fig. 1. The plots of criterion functions (left column) and empirical criterion functions (right column) based on weighted kappa (solid lines), Kendall's τ_b (dashed lines) and correct classification rate (dotted lines).

Fig. 2.

The plots of empirical criterion functions based on weighted kappa $\left(\hat{\vartheta}_{\kappa}\right)$, Kendall's τ_b $\left(\hat{\vartheta}_{\tau}\right)$ and correct classification rate $\left(\hat{\vartheta}_{cc}\right)$ versus all possible one-dimensional cut-points for the Diabetes and Depression study data ($n = 1430$).

Fig. 3. The plots of empirical criterion functions based on weighted kappa (ϑ_{κ} , left top), Kendall's τ_b (ϑ_{τ} , right top) and correct classification rate ($\hat{\vartheta}_{CC}$, bottom) versus all possible two-dimensional cut-points for the Diabetes and Depression study dataset ($n = 1430$).

Table 1

Simulation results when $L = 2$ (Bias: Empirical bias $\times 10^3$; SE: empirical standard deviations $\times 10^3$; ASE: average of estimated standard deviation $\times 10^3$; CP: empirical coverage probabilities of 95% confidence intervals $\times 10^2$).

σ	n	ϑ_{κ}				ϑ_{τ}				ϑ_{CC}			
		Bias	SE	ASE	CP	Bias	SE	ASE	CP	Bias	SE	ASE	CP
0.6	50	6	207	188	93.4	5	219	191	93.2	47	214	191	92.5
	100	4	166	155	95.0	4	180	162	95.7	41	166	157	93.5
	200	1	133	129	97.7	2	149	137	96.9	30	131	130	95.2
1.0	50	8	358	370	95.6	2	415	370	95.5	91	366	358	94.7
	100	7	276	272	98.1	9	326	314	98.0	65	280	291	97.7
	200	5	227	220	97.6	3	271	257	97.4	28	234	235	97.0

Table 2

Simulation results when $L = 3$ (Bias: Empirical bias $\times 10^3$; SE: empirical standard deviations $\times 10^3$; ASE: average of estimated standard deviation $\times 10^3$; CP: empirical coverage probabilities of 95% confidence intervals $\times 10^2$).

σ	n	ϑ_{κ}			ϑ_{τ}			ϑ_{CC}					
		Bias	SE	ASE	CP	Bias	SE	ASE	CP	Bias	SE	ASE	CP
0.6	50	32	246	215	92.1	15	244	227	93.0	50	239	215	92.9
		42	236	217	94.0	0	240	226	94.2	52	236	222	94.3
	100	36	196	181	95.8	5	211	185	95.3	60	189	180	94.7
		33	188	176	94.6	2	199	178	94.9	42	188	176	94.8
	200	29	152	146	96.8	4	166	151	97.2	39	153	147	95.8
		26	151	147	95.8	5	163	151	96.2	35	151	147	95.4
1.0	50	34	396	384	96.9	58	463	434	97.0	96	416	400	94.8
		12	400	387	96.0	56	466	432	96.8	98	415	427	95.4
	100	7	311	321	96.6	29	399	367	96.5	79	322	335	96.8
		3	327	322	96.9	51	405	370	96.2	69	342	343	94.9
	200	5	251	258	98.4	31	323	301	97.3	51	257	265	96.8
		3	252	255	98.5	21	316	301	98.0	53	262	270	96.9

Table 3

Simulation results when $L = 2$ (Bias: Empirical bias $\times 10^3$; SE: empirical standard deviations $\times 10^3$; ASE_s : average of estimated standard deviation $\times 10^3$ based on smooth bootstrap; CP_s : empirical coverage probabilities of 95% confidence intervals $\times 10^2$ based on smooth bootstrap).

σ	n	$\hat{\vartheta}_\kappa$				$\hat{\vartheta}_\tau$				$\hat{\vartheta}_{CC}$			
		Bias	SE	$ASE_{\hat{\vartheta}_\kappa}$	CP_κ	Bias	SE	$ASE_{\hat{\vartheta}_\tau}$	CP_τ	Bias	SE	$ASE_{\hat{\vartheta}_{CC}}$	CP_{CC}
0.6	50	4	201	210	93.7	5	219	225	93.0	49	205	215	94.0
	100	1	161	171	94.0	2	180	191	94.8	38	164	175	93.9
	200	4	129	137	93.9	3	146	152	92.8	32	133	139	92.6
1.0	50	9	346	368	93.4	1	411	433	92.8	81	358	394	94.2
	100	12	282	292	92.5	20	345	353	92.2	72	301	306	92.2
	200	5	211	230	94.6	5	261	282	93.4	42	220	237	93.8

Table 4

Simulation results when $L = 3$ (Bias: Empirical bias $\times 10^3$; SE: empirical standard deviations $\times 10^3$; ASE_s : average of estimated standard deviation based on smooth bootstrap $\times 10^3$; CP_s : empirical coverage probabilities of 95% confidence intervals $\times 10^2$ based on smooth bootstrap).

σ	n	ϑ_{κ}				ϑ_{τ}				ϑ_{CC}			
		Bias	SE	ASE_s	CP_s	Bias	SE	ASE_s	CP_s	Bias	SE	ASE_s	CP_s
0.6	50	-22	227	242	94.3	-83	232	252	93.4	-49	231	244	94.2
		22	235	238	92.5	75	240	249	92.5	-47	235	241	92.1
	100	-29	187	197	93.4	-82	202	213	91.3	-51	184	198	91.8
		21	196	199	92.6	72	211	213	91.0	-47	196	199	92.3
	200	-38	152	159	92.6	-82	167	173	90.0	-34	153	159	92.6
		42.2	148	159	94.0	85	159	173	92.1	-27	148	159	94.3
1.0	50	31	408	416	92.0	-41	474	511	92.5	-89	433	445	91.2
		-20	415	419	90.6	64	487	509	93.3	-94	435	449	92.8
	100	8	320	335	92.7	-37	408	430	93.3	-77	339	351	92.2
		-7	340	336	94.3	28	394	434	94.3	-76	321	353	93.4
	200	9	250	263	93.1	-10	326	358	94.5	-43	258	276	94.0
		8	252	264	93.8	31	321	357	95.1	-40	259	275	93.7

Table 5

Summary of multiple solutions in simulations (ANS: average number of solutions; PMS: percent of multiple solutions).

σ	n	$\hat{d}_{\kappa}$		$\hat{d}_{\tau}$		$\hat{d}_{CC}$	
		ANS	PMS	ANS	PMS	ANS	PMS
$L=2$							
0.6	50	1.052	4.0%	1.023	2.3%	1.374	30.2%
	100	1.048	3.7%	1.014	1.4%	1.524	38.7%
	200	1.024	2.9%	1.003	1.1%	1.731	39.7%
1.0	50	1.056	3.8%	1.009	0.9%	1.656	41.6%
	100	1.030	2.6%	1.002	0.4%	1.751	45.1%
	200	1.045	2.1%	1.003	1.8%	1.749	46.1%
$L=3$							
0.6	50	1.004	0.4%	1.015	1.5%	1.651	42.5%
	100	1.000	0.0%	1.005	0.5%	2.216	58.8%
	200	1.000	0.0%	1.001	0.1%	2.602	66.1%
1.0	50	1.004	0.3%	1.002	0.2%	2.491	64.1%
	100	1.000	0.0%	1.000	0.0%	2.900	68.1%
	200	1.000	1.0%	1.000	0.0%	3.397	71.4%