Navigating challenges in classification and outlier detection: a remedy based on semi-parametric density ratio models

Yukun Liu (East China Normal University)

Joint work with: Siyan Liu (East China Normal University)
Pengfei Li (University of Waterloo)
Jing Qin (National Institutes of Health)

BIRS - IASM Workshop ⋆ Hang Zhou ⋆ Dec 14, 2023

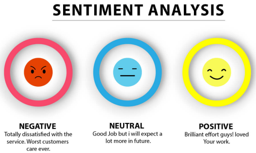
Outline

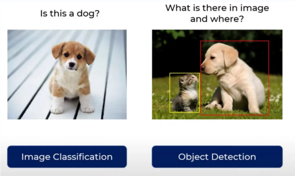
1. Introduction and motivation
2. Model, identifiability and parameter estimation
3. Semi-parametric label prediction
4. A simulation study
5. Real applications

Classification

- **Goal**: to assign categorical labels to unlabelled test data based on patterns and relationships learned from a labeled training dataset.

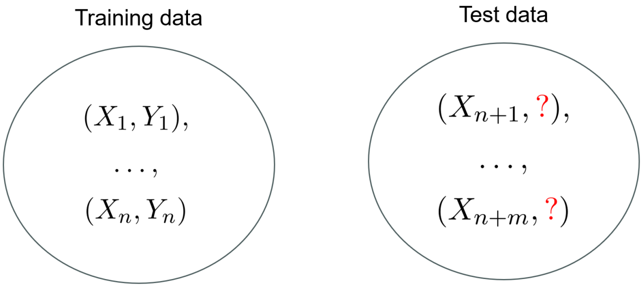
- Classification has **diverse applications**, including
  - email spam filtering (Delany et al., 2012; Fan et al., 2016),
  - sentiment analysis (Medhat et al., 2014; Wang et al., 2016),
  - image recognition (Krizhevsky et al., 2017; Pan et al., 2018).
  - ……


SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customers care ever.
NEUTRAL
Good Job but i will expect a lot more in future.
POSITIVE
Brilliant effort guys! loved Your work.



Training and test data-set paradigm

X: features/covariates/input variables $\qquad$ Y: label/response/output variable



Training data

(X_1, Y_1),

$\ldots,$

(X_n, Y_n)

Test data

$(X_{n+1}, ?)$,

$\ldots,$

$(X_{n+m}, ?)$

- Training data: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P_{\text{train}}(Y, X)$
- **Ideal** test data: $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m}) \overset{iid}{\sim} P_{\text{train}}(Y, X)$

Convention

Training and test data are often assumed to **have the same distribution**

$$P_{\text{train}}(Y, X) = P_{\text{test}}(Y, X)$$

- **Many powerful supervised learning algorithms** try to estimate the common $P(Y = y|X = x)$.
  - **Decision Trees** (Breiman, 1984; Friedl and Brodley, 1997; Kim and Loh, 2001),
  - **Random Forests** (Ho, 1995; Breiman, 2001; Ham et al., 2005; Biau, 2012),
  - **Support Vector Machines** (Cortes and Vapnik, 1995; Suykens and Vandewalle, 1999; Pavlidis et al., 2004; Cervantes et al., 2020),
  - **Neural Networks** (Dreiseitl and Ohno-Machado, 2002; Ghosh et al., 2004; Krizhevsky et al., 2017; Gurney, 2018).
- Then classify the test data using the estimated $P(Y = y|X = x)$.

Challenge

However, the conventional methods **face challenges or even underperform** when the training and test data-sets exhibits **distributions mismatches**

- **Distributions mismatch** or **distribution shift**:

$$P_{\text{train}}(Y, X) \neq P_{\text{test}}(Y, X)$$

Challenge

However, the conventional methods **face challenges or even underperform** when the training and test data-sets exhibits **distributions mismatches**

- **Distributions mismatch** or **distribution shift**:

$$P_{\text{train}}(Y, X) \neq P_{\text{test}}(Y, X)$$

- Two commonly-seen special cases
  - **Covariate shift**: $P_{\text{train}}(Y|X) = P_{\text{test}}(Y|X), \quad P_{\text{train}}(X) \neq P_{\text{test}}(X),$
  - **Label shift**: $P_{\text{train}}(X|Y) = P_{\text{test}}(X|Y), \quad P_{\text{train}}(Y) \neq P_{\text{test}}(Y),$

This talk

We focuses on the case where both **covariate shift** and **label shift** exist.

This talk

We focuses on the case where both **covariate shift** and **label shift** exist.

- The labelled training data can be reorganized as

$$\{(X_{0j}, Y_{0j} = 0)\} \bigcup \cdots \bigcup \{(X_{K-1,j}, Y_{K-1,j} = K-1)\}$$

  where $X_{kj} \sim F_k(x) = P_{\text{train}}(X \le x | Y = k), \quad k = 0, 1, \ldots, K-1.$

- In the unlabelled test data, a feature X
  - may **come from** $F_0(x), F_1(x), \ldots, F_{K-1}(x)$, or
  - **does not** come from any of them **(outliers)**

This talk

- Let $F_K(x)$ denote the cdf of the **outliers**, and categorize them into **Class** K.

This talk

- Let $F_K(x)$ denote the cdf of the **outliers**, and categorize them into **Class** K.
- In the test data, let $\pi_k = P_{\text{test}}(Y = k)$, $k = 0, 1, \ldots, K$.

This talk

- Let $F_K(x)$ denote the cdf of the **outliers**, and categorize them into **Class** K.

- In the test data, let $\pi_k = P_{\text{test}}(Y = k)$, $k = 0, 1, \ldots, K$.

- X in the test data follows a finite mixture model

$$\pi_0 F_0(x) + \cdots + \pi_{K-1} F_{K-1}(x) + \pi_K F_K(x)$$

This talk

- Let $F_K(x)$ denote the cdf of the **outliers**, and categorize them into **Class** K.
- In the test data, let $\pi_k = P_{\text{test}}(Y = k)$, $k = 0, 1, \ldots, K$.
- X in the test data follows a finite mixture model

$$\pi_0 F_0(x) + \cdots + \pi_{K-1} F_{K-1}(x) + \pi_K F_K(x)$$

The goal is to **make prediction about** Y for each X in the test data

Literature on outlier detection

- Applicable in fraud detection, network security, quality control, and more
- The problem of " **whether a data point in the test data is an outlier**" has been studied extensively recently:
  - Unconstrained least-squares importance fitting (uLSIF) method (Hido et al., 2011)
  - CNN $+$ uLSIF, (Nam and Sugiyama, 2015)
  - A robust outlier detection method incorporating k-NN algorithm (Li et al., 2022)
- **Limitations**:
  - nonparametric estimation of **density ratios**,
  - absence of a more detailed classification

Label prediction set

- The conventional classification algorithms and the outlier detection methods all provide a **prediction point** for the label of each test data point.

- An alternative is to construct a **prediction set**: the density-level set (Cadre, 2006; Lei et al., 2013; Rigollet & Vert, 2009; Sadinle et al., 2019)

$$C(x) = \{k : x \in A_k\}, \quad A_k = \{x | f_k(x) > f_{k,\alpha}\}$$

  where

  - $f_k(x)$ is the pdf corresponding to $F_k(x)$,
  - $f_{k,\alpha}$ is the α-th quantile of $f_k(X)$ for $X \sim f_k(x)$.

- $C(x)$ may contains more than one labels.

- An x with $C(x) = \varnothing$ is classified as outlier.

Label prediction set

- **Weakness of the density-level set**
  - does not utilize information comparing different classes, potentially leading to efficiency loss
- To overcome this problem, Guan and Tibshirani (2022) proposed the BCOPS (balanced and conformal optimized prediction set) to construct $C(x)$
  - Perform better because it combines information from different classes and unlabelled test samples
- The validation of BCOPS is built on the assumption that **the outliers can be perfectly separated from the observed classes** (their Assumption 6).
  - **Too strong to be satisfied by popular parametric models**, such as normal.

A challenge in BCOPS

Hereafter we assume $K = 2$ and let

$$f_{\text{test}}(x) = \pi_0 f_0(x) + \cdots \pi_1 f_1(x) + \pi_2 f_2(x).$$

- To see **their Assumption 6 is too strong**, let $\eta_l(x) = \log\{f_l(x)/f_{\text{test}}(x)\}$ and $g_{l,k}(\cdot)$ be the density of η_l in class k for $l \in \{0,1\}$ and $k \in \{0,1,2\}$.
- Define $S_l = \{z : g_{l,l} \circ \eta_l(z) \geq Q(\zeta; g_{l,l} \circ \eta_l, F_l)\}$, where $g_{l,l} \circ \eta_l(z) = g_{l,l}(\eta_l(z))$ and ζ is a user-specific positive constant, where they recommended $\zeta = 0.2$.

A challenge in BCOPS

Hereafter we assume $K = 2$ and let

$$f_{\text{test}}(x) = \pi_0 f_0(x) + \cdots \pi_1 f_1(x) + \pi_2 f_2(x).$$

- To see **their Assumption 6 is too strong**, let $\eta_l(x) = \log\{f_l(x)/f_{\text{test}}(x)\}$ and $g_{l,k}(\cdot)$ be the density of η_l in class k for $l \in \{0,1\}$ and $k \in \{0,1,2\}$.

- Define $S_l = \{z : g_{l,l} \circ \eta_l(z) \geq Q(\zeta; g_{l,l} \circ \eta_l, F_l)\}$, where $g_{l,l} \circ \eta_l(z) = g_{l,l}(\eta_l(z))$ and ζ is a user-specific positive constant, where they recommended $\zeta = 0.2$.

- **Their Assumption 6** requires

$$P_2(X \in S_l) = 0, \quad l = 0, 1,$$

where P_k takes probability when $X \sim F_k(x)$.

A challenge in BCOPS

Values of $P_2(S_0)$ and $P_2(S_1)$ when F_k is the distribution function of $N(\mu_k, I_3)$ with $\mu_0^\top = (0,0,0)$, $\pi_0 = 0.35$, and $\pi_1 = 0.3$.

$\mu_1^\top$	$\mu_2^\top$	$P_2(S_0)$	$P_2(S_1)$
$(0.25, 0.25, 0.25)$	$(1.00, -0.50, -0.50)$	0.480	0.422
$(1.00, 1.00, 0.00)$	$(1.00, -0.50, -0.50)$	0.426	0.360
$(1.00, 0.30, -0.80)$	$(-0.70, -0.20, 1.50)$	0.464	0.120
$(1.00, 0.30, -0.80)$	$(1.00, -0.50, -0.50)$	0.377	0.628

This motivates us to develop a new label prediction set.

Outline

1. Introduction and motivation
2. Model, identifiability and parameter estimation
3. Semi-parametric label prediction
4. A simulation study
5. Real applications

Identifiability

- Recall that we have data from $f_0(x)$ and $f_1(x)$, and an X in the test data follows
$$f_{\text{test}}(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x).$$

Identifiability

- Recall that we have data from $f_0(x)$ and $f_1(x)$, and an X in the test data follows

$$f_{\text{test}}(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x).$$

- **Challenge in identifiability:**
  - f_0 and f_1 are identifiable nonparametrically
  - However, there are **no direct data from f_2**, but only indirect data in the test data.

Lemma 1

For a mixture model $\lambda F(x) + (1-\lambda)G(x)$, where $\lambda \in [0,1]$ and F and G be two cdfs, if G **is known** but λ **and** F **are unknown**, then λ and F are **unidentifiable**.

$$\lambda_1 \left\{ \frac{\lambda_2}{\lambda_1} F(x) + \frac{\lambda_1 - \lambda_2}{\lambda_1} G(x) \right\} + (1-\lambda_1)G(x) = \lambda_2 F(x) + (1-\lambda_2)G(x)$$

Our model assumption

We make a semiparametric density ratio model (Anderson, 1979; DRM) assumption:

$$f_k(x) = f_0(x)\exp\{\alpha_k + \beta_k^\top \phi(x)\}, \quad k = 1,\ 2,$$

where $\phi(x)$ is a pre-specified q-variate function and usually taken as x.

Our model assumption

We make a semiparametric density ratio model (Anderson, 1979; DRM) assumption:

$$f_k(x) = f_0(x)\exp\{\alpha_k + \beta_k^\top \phi(x)\}, \quad k = 1,\ 2,$$

where $\phi(x)$ is a pre-specified q-variate function and usually taken as x.

- Satisfied by many **popular parametric distribution families**, including normal, binomial, exponential, Poisson and so on.
- Closely related to **discrimination analysis** and problems subject to **covariate shift**.

Identifiability

Under DRM, we rewrite

$$\begin{aligned} f_{\text{test}}(x) &= \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x) \\ &= f_0(x)\{\pi_0 + \pi_1 e^{\gamma_1^\top \phi_e(x)} + \pi_2 e^{\gamma_2^\top \phi_e(x)}\}. \end{aligned}$$

where $\gamma_k = (\alpha_k, \beta_k^\top)^\top$ and $\phi_e(x) = (1, \phi^\top(x))^\top$.

Assumption 1 Let $n_k = \sum_{i=1}^n I(Y_i = k)$ for $k = 0, 1$. There exist constants $c_0, c_1, c_2 \in (0, 1)$ such that $n_0/N = c_0 + o(1)$, $n_1/N = c_1 + o(1)$ and $m/N = c_2 + o(1)$ as $N \to \infty$.

Assumption 2 $\beta_1^o \neq 0$, $\beta_2^o \neq 0$, $\beta_1^o \neq \beta_2^o$, $\pi_2^o > 0$, and $\mathbb{E}_0\{\phi_e(X)\phi_e^\top(X)\}$ **is finite and positive definite**.

Lemma 2

Under Assumptions 1 and 2, $f_0(x)$ and $\theta = (\gamma_1^\top, \gamma_2^\top, \pi_0, \pi_1)$ are identifiable.

Semiparametric likelihood estimation

- Under DRM, the likelihood contribution of the training data is

$$L_0 \quad = \quad \prod_{i=1}^{n} \{ e^{Y_i \gamma_1^\top \phi_e(X_i)} dF_0(X_i) \}$$

- The likelihood contribution of the test data is

$$L_1 \quad = \quad \prod_{i=n+1}^{N} \left[\{ \pi_0 + \pi_1 e^{\gamma_1^\top \phi_e(X_i)} + \pi_2 e^{\gamma_2^\top \phi_e(X_i)} \} dF_0(X_i) \right]$$

- The likelihood based on all data is

$$L_0 \times L_1 = \prod_{I=1}^{N} \left[dF_0(X_i) \times e^{Y_i (1 - D_i) \gamma_1^\top \phi_e(X_i)} \times \{ \pi_0 + \pi_1 e^{\gamma_1^\top \phi_e(X_i)} + \pi_2 e^{\gamma_1^\top \phi_e(X_i)} \}^{D_i} \right].$$

Empirical profile likelihood function

- We use empirical likelihood to handle the baseline distribution, namely

$$F_0(x) = \sum_{i=1}^{N} p_i I(X_i \le x).$$

- Then the log-likelihood becomes

$$\tilde{\ell} = \sum_{i=1}^{N} [\log(p_i) + Y_i(1 - D_i)\gamma_1^\top \phi_e(X_i) + D_i \log\{\pi_0 + \pi_1 e^{\gamma_1^\top \phi_e(X_i)} + \pi_2 e^{\gamma_2^\top \phi_e(X_i)}\}],$$

where feasible p_i's satisfy

$$p_i \ge 0, \quad \sum_{i=1}^{N} p_i = 1, \quad \sum_{i=1}^{N} p_i\{e^{\gamma_k^\top \phi_e(X_i)} - 1\} = 0, \quad k = 1, 2. \tag{1}$$

Semiparametric profile likelihood function

- Given $\theta = (\gamma_1^\top, \gamma_2^\top, \pi_0, \pi_1)$, the log-function $\tilde{\ell}$ takes its maximum when

$$p_i = \frac{1}{N} \frac{1}{1 + \lambda_1\{e^{\gamma_1^\top \phi_e(X_i)} - 1\} + \lambda_2\{e^{\gamma_2^\top \phi_e(X_i)} - 1\}},$$

where (λ_1, λ_2) is the solution to

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\gamma_1^\top \phi_e(X_i)} - 1}{1 + \lambda_1\{e^{\gamma_1^\top \phi_e(X_i)} - 1\} + \lambda_2\{e^{\gamma_2^\top \phi_e(X_i)} - 1\}} &= 0, \\ \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\gamma_2^\top \phi_e(X_i)} - 1}{1 + \lambda_1\{e^{\gamma_1^\top \phi_e(X_i)} - 1\} + \lambda_2\{e^{\gamma_2^\top \phi_e(X_i)} - 1\}} &= 0. \end{aligned} \qquad (2)$$

- The profile log-likelihood function of θ is

$$\begin{aligned} \ell(\theta) \quad = \quad & -\sum_{k=1}^{N} \log[1 + \lambda_1\{e^{\gamma_1^\top \phi_e(X_i)} - 1\} + \lambda_2\{e^{\gamma_2^\top \phi_e(X_i)} - 1\}] \\ & + \sum_{i=1}^{N} [Y_i(1 - D_i)\gamma_1^\top \phi_e(X_i) + D_i \log\{\pi_0 + \pi_1 e^{\gamma_1^\top \phi_e(X_i)} + \pi_2 e^{\gamma_2^\top \phi_e(X_i)}\}]. \end{aligned}$$

Maximum likelihood estimation

- We propose to estimate θ by the maximum likelihood estimator (MLE)

$$\hat{\theta} := (\hat{\gamma}_1^\top, \hat{\gamma}_2^\top, \hat{\pi}_0, \hat{\pi}_1) = \arg\max_{\theta \in \Theta} \ell(\theta).$$

- Accordingly, we have the MLE $\hat{p}_i$ of p_i, and the MLEs of F_0 and F_k:

$$\begin{aligned} \hat{F}_0(x) &= \sum_{i=1}^{N} \hat{p}_i I(X_i \leq x), \\ \hat{F}_k(x) &= \sum_{i=1}^{N} \hat{p}_i e^{\hat{\gamma}_k^\top \phi_e(X_i)} I(X_i \leq x), \quad k = 1, 2. \end{aligned}$$

- These estimators provides basic elements for the construction of the proposed label prediction set.

Asymptotics

- **Assumption 3** : The function $\mathbb{E}_0[\exp\{\beta_k^\top \phi(X)\}]$ is finite for β_k in a neighborhood of β_k^o and $k = 1, 2$, and the matrix W_* is nonsingular.
- **Assumption 4**: $\Theta \subset \mathbb{R}^s$ is a closed subset, and θ^o is an interior point of Θ.

Theorem 1

Under Assumptions 1-4, as N goes to infinity,

(1) $\sqrt{N}(\hat{\theta} - \theta^o) \to N\left(0, W_*^{-1}\right)$ in distribution

(2) The stochastic process $\sqrt{N}\{\hat{F}_k(\cdot) - F_k(\cdot)\}$ converges weakly to a Gaussian process with mean zero for each $k = 0, 1, 2$.

Numerical implementation: EM algorithm

- Naturally we take the labels $\{Y_j^* : n+1 \leq j \leq n+m\}$ for the test data as natural missing data.

- Let $\mathcal{X}$ denote all the observed data. It is clear that

$$\begin{aligned} w_{jk}^{(r+1)} &= \mathbb{E}\{I(Y_j^* = k) | \mathcal{X}, \theta^{(r)}\} \\ &= \frac{\pi_k^{(r)} e^{\gamma_k^{(r)\top} \phi_e(X_j)}}{\pi_0^{(r)} + \pi_1^{(r)} e^{\gamma_1^{(r)\top} \phi_e(X_j)} + (1 - \pi_0^{(r)} - \pi_1^{(r)}) e^{\gamma_2^{(r)\top} \phi_e(X_j)}}. \end{aligned}$$

- An EM algorithm can be constructed by standard discussions. The details are omitted.

Outline

1. Introduction and motivation

2. Model, identifiability and parameter estimation

3. Semi-parametric label prediction

4. A simulation study

5. Real applications

Semi-parametric label prediction

- Following Guan and Tibshirani (2022), we consider constructing a label prediction set $C(x) \in \{\{0\}, \{1\}, \{0, 1\}, \varnothing\}$ for each $X = x$, instead of giving a label prediction point.

Semi-parametric label prediction

- Following Guan and Tibshirani (2022), we consider constructing a label prediction set $C(x) \in \{\{0\},\{1\},\{0,1\},\varnothing\}$ for each $X = x$, instead of giving a label prediction point.
- A reasonable prediction set $C(x)$ can be constructed as the minimizer of the misclassification loss averaged over the out-of-sample data

$$(\mathcal{P}) \qquad \begin{aligned} &\min \int |C(x)| f_{\text{test}}(x) dx, \\ &\text{s.t. } P_k(k \in C(X)) \geq 1 - \alpha, \quad k = 0, 1, \end{aligned}$$

  where

  - $\alpha \in (0,1)$ is a prespecified mis-coverage level,
  - $|C(x)|$ be the size of $C(x)$, and
  - the weight function $f_{\text{test}}(x)$ balances classification accuracy and power of outlier detection.

Semi-parametric label prediction

- The solution to problem $(\mathcal{P})$ is the oracle prediction set $C_*(x) = \{k : x \in A_{k*}\}$, where A_{k*} is the solution to

$$(\mathcal{P}_k) \qquad \begin{aligned} &\min \int I(x \in A_k) f_{\text{test}}(x)dx, \\ &\text{s.t. } P_k\,(x \in A_k) \geq 1 - \alpha, \quad k = 0, 1. \end{aligned}$$

Semi-parametric label prediction

- The solution to problem $(\mathcal{P})$ is the oracle prediction set $C_*(x) = \{k : x \in A_{k*}\}$, where A_{k*} is the solution to

$$(\mathcal{P}_k) \qquad \begin{aligned} &\min \int I(x \in A_k) f_{\text{test}}(x) dx, \\ &\text{s.t. } P_k\left(x \in A_k\right) \geq 1 - \alpha, \quad k = 0, 1. \end{aligned}$$

- The set A_{k*}, also called the oracle acceptance set for class k, has an explicit form in terms of density ratios $v_k(x) = f_k(x)/f_{\text{test}}(x)$, namely,

$$A_{k*} = \{x : v_k(x) \geq Q\left(\alpha; v_k, F_k\right)\}, \tag{3}$$

where $Q(\alpha; h, F)$ is the lower α percentile of a real-valued function $h(X)$ under distribution F, i.e. $Q(\alpha; h, F) = \sup\{t : \int I(h(x) \leq t) dF(x) \leq \alpha\}$.

Semi-parametric likelihood prediction Set

- **We propose a semi-parametric likelihood prediction method**, without requiring the Assumption 6 of Guan and Tibshirani (2022).

- As A_{k*} depends **only on the ordering** of $v_k(x) = f_k(x)/f_{\text{test}}(x)$, any order-preserving transformation of $v_k(x)$ is permitted when constructing A_{k*}.

Semi-parametric likelihood prediction Set

- **We propose a semi-parametric likelihood prediction method**, without requiring the Assumption 6 of Guan and Tibshirani (2022).

- As A_{k*} depends **only on the ordering** of $v_k(x) = f_k(x)/f_{\text{test}}(x)$, any order-preserving transformation of $v_k(x)$ is permitted when constructing A_{k*}.

- We take

$$
\begin{aligned}
v_0(x) &= \frac{f_0(x)}{f_0(x) + f_{\text{test}}(x)} = \frac{1}{1 + \pi_0 + \pi_1 \exp\{\gamma_1^\top \phi_e(x)\} + \pi_2 \exp\{\gamma_1^\top \phi_e(x))\}}, \\
v_1(x) &= \frac{f_1(x)}{f_1(x) + f_{\text{test}}(x)} = \frac{\exp\{\gamma_1^\top \phi_e(x)\}}{\pi_0 + (1 + \pi_1) \exp\{\gamma_1^\top \phi_e(x)\} + \pi_2 \exp\{\gamma_2^\top \phi_e(x)\}}.
\end{aligned}
$$

Semi-parametric empirical likelihood prediction Set

- Let $F_{nk}(x)$ denote the empirical distribution of $\{X_i : Y_i = k, D_i = 0\}$ for $k = 0, 1$.
- Our semi-parametric empirical likelihood prediction set (SELPS) is

$$\hat{C}(x) = \{k : x \in \hat{A}_k\},$$

where

$$\hat{A}_k = \{x : \hat{v}_k(x) \geq Q(\alpha; \hat{v}_k, F_{nk})\},$$

with

$$\begin{aligned} \hat{v}_0(x) &= \frac{1}{1 + \hat{\pi}_0 + \hat{\pi}_1 \exp\{\hat{\gamma}_1^\top \phi_e(x)\} + \hat{\pi}_2 \exp\{\hat{\gamma}_2^\top \phi_e(x)\}}, \\ \hat{v}_1(x) &= \frac{\exp\{\hat{\gamma}_1^\top \phi_e(x)\}}{\hat{\pi}_0 + (1 + \hat{\pi}_1) \exp\{\hat{\gamma}_1^\top \phi_e(x)\} + \hat{\pi}_2 \exp\{\hat{\gamma}_2^\top \phi_e(x)\}}. \end{aligned}$$

Semi-parametric empirical likelihood prediction Set

Assumption 5: The densities $f_0(x)$ and $f_1(x)$ are upper bounded by a constant. There exist constants $0 < \epsilon_1 \leq \epsilon_2$ and $\epsilon,\ \delta_0,\ \varsigma > 0$ such that for $k = 0,\ 1,$

$$\epsilon_1|\delta|^{\varsigma} \leq |P_k(v_k(X) \leq Q(t; v_k, F_k) + \delta) - t| \leq \epsilon_2|\delta|^{\varsigma}, \forall\delta \in [-\delta_0, \delta_0], t \in [\alpha - \epsilon, \alpha + \epsilon].$$

- This assumption requires that the likelihood ratio functions $v_k(x)$ are neither too steep nor too flat around the boundary of $Q(t; v_k, F_k)$ uniformly for $t \in [\alpha - \epsilon, \alpha + \epsilon]$, where $Q(\alpha; v_k, F_k)$ corresponds to the optimal decision regions A_{k*}.

Semi-parametric empirical likelihood prediction Set

Theorem 2

Suppose that Assumptions 1-5 are satisfied. Given a mis-coverage rate $\alpha > 0$, let $\hat{C}(x)$ be the proposed SELPS and $C_*(x)$ the oracle prediction set. Then

(i) there exists $M > 0$ such that

$$P_k(X \in \hat{A}_k) \geq 1 - \alpha - M \left(\frac{\log N}{N} \right)^{\frac{\min\{\varsigma, 2\}}{6}},$$

(ii) there exists a large enough constant $D > 0$ such that

$$\lim_{N \to \infty} P \left(\int (|\widehat{C}(x)| - |C_*(x)|) f_{\text{test}}(x) dx \geq D \left(\frac{\log N}{N} \right)^{\frac{\min\{\varsigma, 2\}}{6}} \right) = 0.$$

Outline

1. Introduction and motivation
2. Model, identifiability and parameter estimation
3. Semi-parametric label prediction
4. A simulation study
5. Real applications

Methods under comparison

We investigate the finite-sample performance of the proposed label prediction method SELPS at 95% coverage level.

- BCOPS(rf): the BCOPS with random forest (rf);
- BCOPS(sel): the BCOPS with the semiparametric EL estimators;
- SELPS: our proposed semi-parametric EL prediction set

Simulation scenarios

We set F_k $(k = 0, 1, 2)$ to be the distribution of $N(\mu_k, \Sigma_k)$, with

- $\mu_0 = (0, 0, \ldots, 0)^\top$, $\mu_1 = (2, 2, 0, \ldots, 0)^\top$ and $\mu_2 = (-2, -2, 0, \ldots, 0)^\top$ are three 10-dimensional vectors,
- Σ_k are 10×10 matrices with diagonal elements being 1 and general (i, j) element being $\rho_k^{|i-j|}$.
  - (ρ_0, ρ_1, ρ_2)=$(0, 0, 0)$ (homogeneous case);
  - (ρ_0, ρ_1, ρ_2)=$(0, 0.5, 0.2)$ (heterogeneous case, model mis-specification).

Simulation scenarios

We set F_k $(k = 0, 1, 2)$ to be the distribution of $N(\mu_k, \Sigma_k)$, with

- $\mu_0 = (0, 0, \ldots, 0)^\top$, $\mu_1 = (2, 2, 0, \ldots, 0)^\top$ and $\mu_2 = (-2, -2, 0, \ldots, 0)^\top$ are three 10-dimensional vectors,

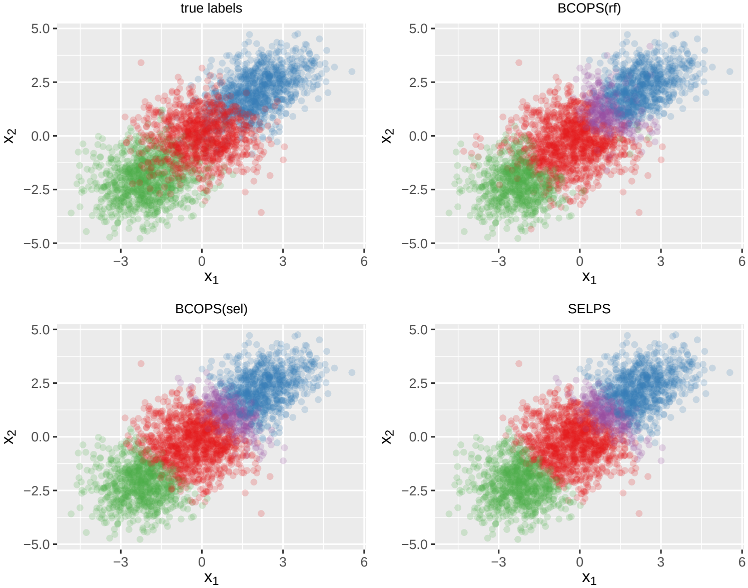
- Σ_k are 10×10 matrices with diagonal elements being 1 and general (i, j) element being $\rho_k^{|i-j|}$.

    - (ρ_0, ρ_1, ρ_2)=$(0, 0, 0)$ (homogeneous case);

    - (ρ_0, ρ_1, ρ_2)=$(0, 0.5, 0.2)$ (heterogeneous case, model mis-specification).

- In each case, for training data-set, $n_0 = 1000$, $n_1 = 2000$; for training data-set , $m = 3000$, one third of which come from F_k for $k = 0, 1, 2$.

An example with heterogeneous variances ($\rho_0 = 0, \rho_1 = 0.5, \rho_2 = 0.2$)



- Red, $\{0\}$
- Blue, $\{1\}$
- Purple, $\{0, 1\}$
- Green, $\varnothing$ (outlier)

Table: Simulation results on abstention rate R, prediction accuracy of BCOPS(rf), BCOPS(sel) and SELPS, and their coverages in terms of coverages I and II at the 95% prediction level

	R	accuracy	coverage I	coverage II
	Homogeneous case: $(\rho_0, \rho_1, \rho_2) = (0, 0, 0)$			
BCOPS(rf)	0.671	0.774	0.956	0.947
BCOPS(sel)	0.746	0.810	0.965	0.961
SELPS	0.774	0.833	0.950	0.957
	Heterogeneous case: $(\rho_0, \rho_1, \rho_2) = (0, 0.5, 0.2)$			
BCOPS(rf)	0.721	0.760	0.963	0.937
BCOPS(sel)	0.763	0.766	0.957	0.936
SELPS	0.778	0.784	0.952	0.937

- Coverage I (II) is defined by the proportion of points (x, y) with $y = 0$ ($y = 1$) in the test data whose predicted sets are either $\{0\}$ ($\{1\}$) or $\{0, 1\}$.

Simulated RMSE and bias (in paratheses) of the estimators for π_k's

n_0/n	(π_1, π_2)	$\tilde{\pi}_0$	$\tilde{\pi}_1$	$\tilde{\pi}_2$	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\pi}_2$
Homogeneous case: $(\rho_0, \rho_1, \rho_2) = (0, 0, 0)$							
0.333	$(0.333, 0.333)$	0.056 (0.054)	0.011 (-0.006)	0.05 (-0.048)	0.014 (0)	0.005 (0)	0.011 (0)
	$(0.400, 0.200)$	0.049 (0.046)	0.012 (-0.006)	0.043 (-0.041)	0.014 (-0.001)	0.006 (0)	0.011 (0)
	$(0.250, 0.500)$	0.058 (0.057)	0.010 (-0.007)	0.052 (-0.050)	0.012 (0)	0.004 (0)	0.010 (0.001)
0.500	$(0.333, 0.333)$	0.049 (0.048)	0.01 (-0.005)	0.044 (-0.043)	0.013 (0)	0.005 (0)	0.010 (0)
	$(0.400, 0.200)$	0.041 (0.039)	0.011 (-0.004)	0.038 (-0.035)	0.012 (0)	0.006 (0)	0.009 (0)
	$(0.250, 0.500)$	0.054 (0.053)	0.009 (-0.006)	0.048 (-0.047)	0.011 (-0.001)	0.004 (0)	0.009 (0.001)
Heterogeneous case: $(\rho_0, \rho_1, \rho_2) = (0, 0.5, 0.2)$							
0.333	$(0.333, 0.333)$	0.063 (0.061)	0.010 (-0.005)	0.058 (-0.056)	0.015 (0.004)	0.006 (0)	0.013 (-0.003)
	$(0.400, 0.200)$	0.053 (0.051)	0.011 (-0.005)	0.048 (-0.046)	0.016 (0.003)	0.006 (0)	0.013 (-0.003)
	$(0.250, 0.500)$	0.066 (0.065)	0.009 (-0.005)	0.061 (-0.059)	0.014 (0.003)	0.006 (0)	0.011 (-0.003)
0.500	$(0.333, 0.333)$	0.058 (0.056)	0.011 (-0.007)	0.051 (-0.050)	0.015 (0.004)	0.006 (-0.001)	0.012 (-0.003)
	$(0.400, 0.200)$	0.047 (0.045)	0.011 (-0.006)	0.042 (-0.040)	0.014 (0.004)	0.007 (-0.001)	0.010 (-0.003)
	$(0.250, 0.500)$	0.062 (0.061)	0.009 (-0.006)	0.057 (-0.055)	0.013 (0.003)	0.005 (-0.001)	0.010 (-0.002)

Outline

1 Introduction and motivation

2 Model, identifiability and parameter estimation

3 Semi-parametric label prediction

4 A simulation study

5 Real applications

Real applications

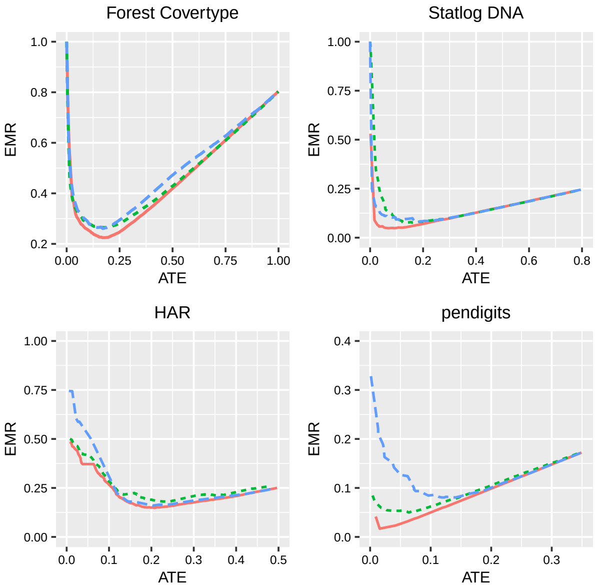
In this section we further investigate the finite-sample performance of the proposed SELPS by analyzing four real-world data-sets:

- Forest Covertype data-set,
    - contains 54 features of 9,813 trees among which 3,969 are Douglas fir (class 0), 4,505 are Krummholz (class 1), and 1,339 are Cottonwood Willow (class 2).
- Human Activity Recognition (HAR) data-set,
    - contains 561 features of three activities, walking (class 0), sitting (class 1) and standing (class 2), with sample size 1,722, 1,777, and 1,906 respectively.
- StatLog DNA data-set,
    - contains 60 features of DNA fragments, including the following three categories: donors (class 0), acceptors (class 1), and neither (class 2), with sample size being 767, 765 and 1,654, respectively.
- pendigits data-set,
    - contains 16 features of pen-based recognition of handwritten digits 0, 1 and 2, among which 779 are of digit 1, 780 are of digit 2 and 780 are of digit 0.

Table: Real data results on abstention rate R, prediction accuracy of BCOPS(rf), BCOPS(glm) and SELPS, their coverages in terms of coverages I and II at the 95% prediction level, and their proportion estimators $\hat{\pi}_1$, $\hat{\pi}_2$.

	R	accuracy	coverage I	coverage II	$\hat{\pi}_1$	$\hat{\pi}_2$
Forest Covertype: $(p, n, m, \pi_1, \pi_2) = (54, 3000, 6813, 0.441, 0.196)$						
BCOPS(rf)	0.146	0.801	0.956	0.947	0.509	0.002
BCOPS(glm)	0.001	0.811	0.957	0.952	0.463	0.048
SELPS	0.255	0.818	0.943	0.941	0.486	0.106
StatLog DNA: $(p, n, m, \pi_1, \pi_2) = (180, 800, 2386, 0.153, 0.693)$						
BCOPS(rf)	0.886	0.781	0.951	0.934	0.196	0.611
BCOPS(glm)	0.909	0.821	0.967	0.945	0.172	0.678
SELPS	0.969	0.915	0.948	0.918	0.141	0.714
HAR: $(p, n, m, \pi_1, \pi_2) = (561, 1600, 3405, 0.257, 0.501)$						
BCOPS(rf)	0.187	0.963	0.963	0.962	0.550	0.210
BCOPS(glm)	0.080	0.711	0.973	0.983	-	-
SELPS	0.249	0.967	0.980	0.954	0.268	0.488
pendigits: $(p, n, m, \pi_1, \pi_2) = (16, 800, 1539, 0.246, 0.507)$						
BCOPS(rf)	0.996	0.889	0.931	0.968	0.259	0.499
BCOPS(glm)	0.992	0.755	0.942	0.966	0.252	0.517
SELPS	1.00	0.937	0.942	0.932	0.245	0.515

Plots of actual type I error (ATE) versus empirical misclassification rate



- BCOPS(rf): green, dotted
- BCOPS(glm): blue, dashed
- SELPS: red, solid

Summary

- The unlabelled test data follow a mixture model, and it can not be identified nonparametrically.
- We propose to model the test data by a finite semiparametric mixture model under density ratio model
- We construct a semiparametric empirical likelihood prediction set (SELPS) for the labels in the test data.
  - All underlying parameters are identifiable.
  - Our method circumvents a stringent separation assumption, which is required by Guan and Tibshirani (2022) but is often violated by commonly-used distributions.
  - We establish the consistency and asymptotic normalities of our estimators, and asymptotic optimality of the proposed SELPS.

Thanks

Yukun Liu

East China Normal University

Email: ykliu@sfs.ecnu.edu.cn