Nearest Neighbor and Kernel Survival Analysis: Nonasymptotic Error Bounds and Strong Consistency Rates

Carnegie Mellon University Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

George H. Chen

Introduction

Survival analysis Reason about time durations until a critical event happens, such as in:

- Health care: time until death, or time until a disease relapses
- Manufacturing: time until a device fails
- Criminology: time until a convicted criminal reoffends

Critical event need not be death but we use this as the running example

Illustrative Example Data

	Gluten allergy	Immuno-suppressant	Low resting heart rate	Irregular heart beat	High BMI	Time of death
	✗	✗	✓	✗	✗	Day 2
	✓	✓	✗	✗	✓	Day 10
	✗	✗	✗	✓	✗	Day ≥ 6

Feature vector X — Observed time Y

Not everyone in training data died

Goal: Estimate $S(t|x) = \mathbb{P}(\text{survive beyond time } t \mid \text{feature vector } x)$

This paper analyzes *k*-NN and kernel Kaplan-Meier estimators (Beran 1981) **for** $S(t|x)$

Contributions

- Most general finite sample error upper bounds for *k*-NN and kernel survival estimators (nearly optimal)
- Experimental evidence that learning a kernel with random survival forests (Ishwaran et al. 2008) often works well in practice (theory for this adaptive kernel variant remains an open problem)

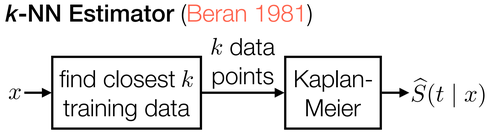
Problem Setup

Model Generate each point (X, Y, δ) i.i.d.:

1. Sample feature vector $X \sim \mathbb{P}_X$
2. Sample time of death $T \sim \mathbb{P}_{T|X}$
3. Sample time of censoring $C \sim \mathbb{P}_{C|X}$
4. If death is before censoring ($T \leq C$):
   Set $Y = T, \delta = 1$
   Otherwise:
   Set $Y = C, \delta = 0$

Goal: Estimate $S(t|x) = \mathbb{P}(T > t \mid X = x)$

***k*-NN Estimator** (Beran 1981)



- Beran also proposed kernel version, showed consistency for both (Euclidean feature space)
- Kernel version finite sample bounds (Euclidean): Dabrowska 1989, Van Keilegom 1998

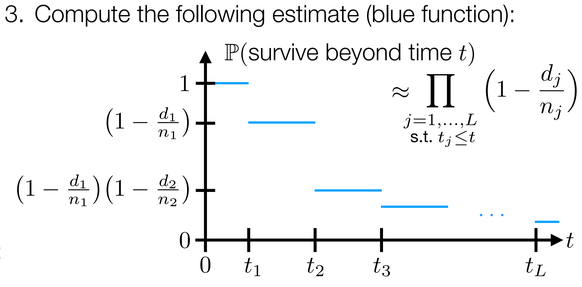
Kaplan-Meier Estimator (Kaplan & Meier 1958)

For given data points, estimate $\mathbb{P}(\text{survive beyond time } t)$:

1. Sort unique times of death: $t_1 < t_2 < \cdots < t_L$
2. Construct the following table:

	t_1	t_2	t_3	$\cdots$	t_L
# people who die	d_1	d_2	d_3	$\cdots$	d_L
# people at risk	n_1	n_2	n_3	$\cdots$	n_L

3. Compute the following estimate (blue function):



Theory

Focus on sup-norm error: $\sup_{t \in [0,\tau]} |\widehat{S}(t|x) - S(t|x)|$ for time horizon τ

Assumptions

- Probability of seeing Y values above τ is large enough: $\exists \theta \in (0, 1/2]$ **s.t.** $\mathbb{P}(Y > \tau | X = x) \geq \theta \ \forall x$
  ➔ With training set size n large enough, enough Y values exceed τ
- Conditional distributions $\mathbb{P}_{T|X}$ & $\mathbb{P}_{C|X}$ are continuous r.v.'s
  ➔ Ensures no ties in when people die and that these conditional distributions have pdf's
- Pdf's of $\mathbb{P}_{T|X}$ & $\mathbb{P}_{C|X}$ are smooth w.r.t. feature space (Hölder index α)
  ➔ Feature vectors close by have similar death/censoring time distributions
  (close enough neighbors in training data will provide useful information for prediction)
- Feature space is separable metric space, $\mathbb{P}_X$ is Borel probability measure
  ➔ Ensures probability of balls well-defined, probability of feature vector landing in support of $\mathbb{P}_X$ is 1

Theorem (informal): *k*-NN estimator with $k = \widetilde{\Theta}(n^{2\alpha/(2\alpha+d)})$ has strong consistency rate:

$$\sup_{t \in [0,\tau]} |\widehat{S}(t|x) - S(t|x)| \leq \widetilde{O}(n^{-\alpha/(2\alpha+d)})$$

d: feature space intrinsic dimension

- The kernel variant also satisfies this bound (different constants)
- If no censoring: upper bounds match lower bound by Chagny & Roche 2014 up to log factor

Main observation: $|\log \widehat{S}(t|x) - \log S(t|x)| \leq$ error in *k*-NN CDF estimate of $\mathbb{P}_{Y|X}$ + error in *k*-NN regression problem assuming $\mathbb{P}_{Y|X}$ known

Key proof ideas: { CDF estimation/regression decomposition: Földes & Rejtö 1981; Controlling bias/variance for *k*-NN estimators: Chaudhuri & Dasgupta 2014 }

Also: new finite sample results for *k*-NN and kernel Nelson-Aalen estimators, choosing *k* via validation set

Experiments

Datasets

Dataset	Description	# subjects	# dim.
pbc	primary biliary cirrhosis	276	17
gbsg2	breast cancer	686	8
recid	recidivism	1445	14
kidney	dialysis	1044	53

Accuracy: concordance index

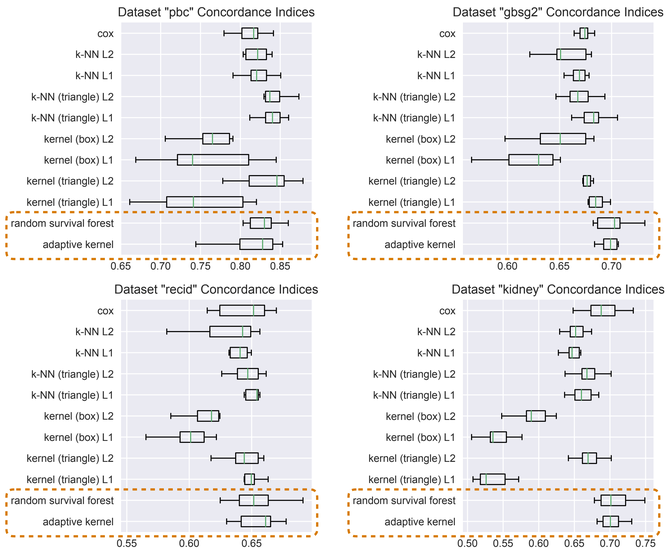
Basic experiment

1. Randomly divide data into 70%/30% train/test pieces
2. Select alg. parameter(s) via 5-fold cross val on training set
3. Evaluate on test set

Repeat basic experiment 10 times

Compare to:

- Cox proportional hazards
- Random survival forest (RSF)
- Kernel survival estimator with kernel learned using RSF ("adaptive kernel")



Distance/kernel choice matters a lot in practice
➔ Adaptively learning these tends to work best