

Nearest Neighbor and Kernel Survival Analysis

Nonasymptotic Error Bounds and Strong Consistency Rates

George H. Chen

Assistant Professor of Information Systems

Carnegie Mellon University

Survival Analysis

	Gluten allergy	Immuno-suppressant	Low resting heart rate	Irregular heart beat	High BMI	Time of death
						Day 2
Feature vector X						Observed time Y
						Day 10
						Day ≥ 6
When we stop collecting training data, not everyone has died!						

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

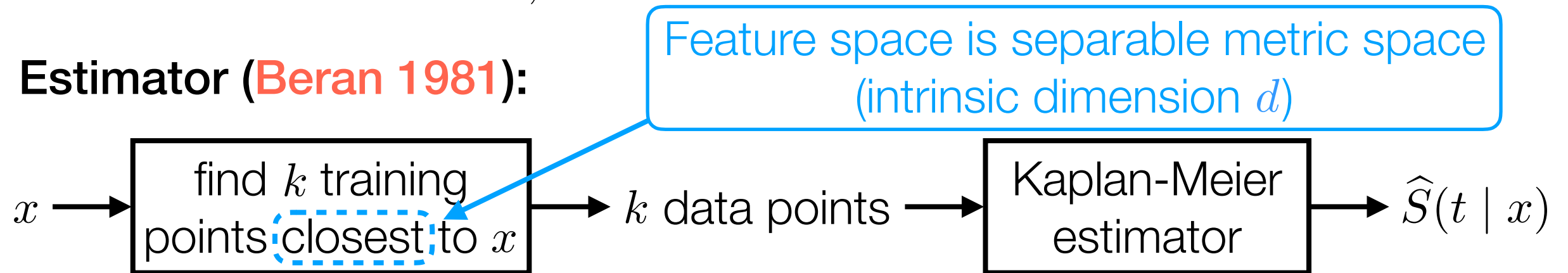
Problem Setup

Model: Generate data point (X, Y, δ) as follows:

1. Sample feature vector $X \sim \mathbb{P}_X$ ← Borel prob. measure
2. Sample time of death $T \sim \mathbb{P}_{T|X}$ ← Continuous r.v. in **time** & smooth w.r.t. **feature space** (Hölder index α)
3. Sample time of censoring $C \sim \mathbb{P}_{C|X}$ ← Continuous r.v. in **time** & smooth w.r.t. **feature space** (Hölder index α)
4. If death happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

Otherwise: Set $Y = C, \delta = 0$

Estimator (Beran 1981):



Kernel variant is similar

Error: $\sup_{t \in [0, \tau]} |\hat{S}(t|x) - S(t|x)|$ for time horizon τ

Enough of the n training data have Y values $> \tau$

Theory (Informal)

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

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

If no censoring, problem reduces to conditional CDF estimation

→ Error upper bound, up to a log factor, matches conditional CDF estimation lower bound by [Chagny & Roche 2014](#)

Proof ideas also give finite sample rates for:

- Kernel Kaplan-Meier estimators
- k -NN & kernel Nelson-Aalen *cumulative hazard* estimators ($-\log S(t | x)$)
- Generalization bound for automatic k using validation data

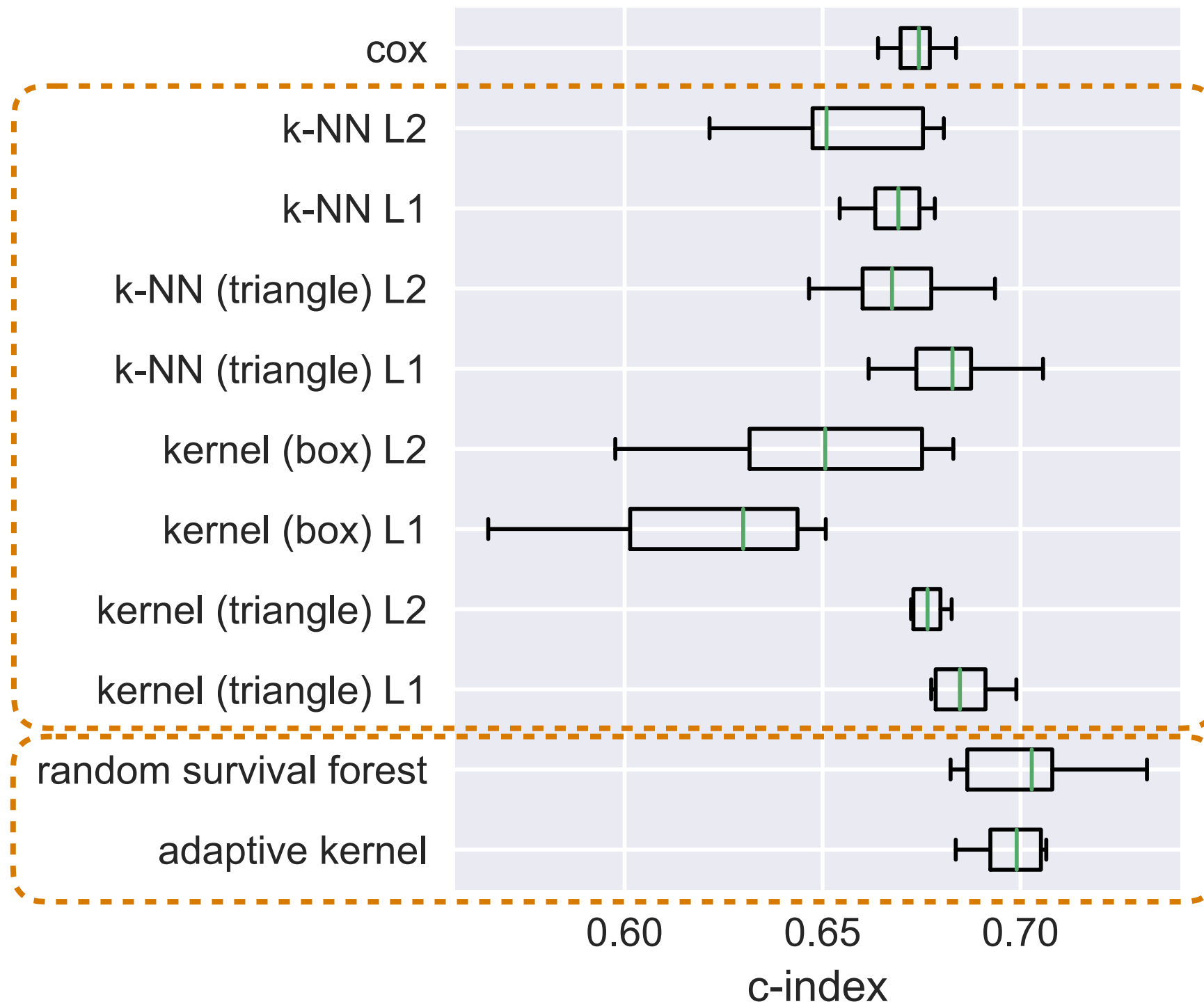
Most general finite sample theory for k -NN and kernel survival estimators

Existing kernel results only for Euclidean space

(Dabrowska 1989, Van Keilegom & Veraverbeke 1996, Van Keilegom 1998)

Experiments

Dataset "gbsg2" Concordance Indices



Distance/kernel choice matter a lot in practice

Learning the kernel typically has best performance (but no theory yet!)