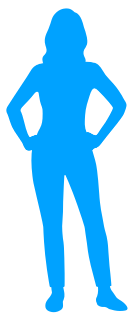
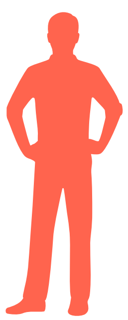
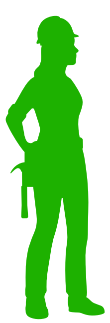
Nearest Neighbor and Kernel Survival Analysis

Nonasymptotic Error Bounds and Strong Consistency Rates

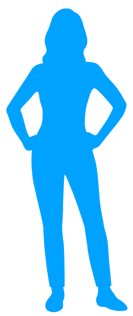
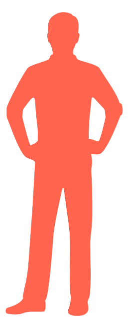
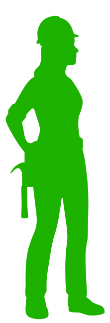
George H. Chen
Assistant Professor of Information Systems
Carnegie Mellon University

June 11, 2019

Survival Analysis

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

Survival Analysis

	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

Survival Analysis

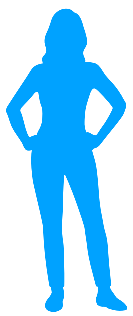
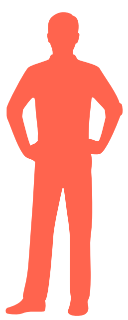
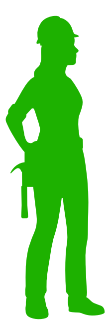
	Gluten allergy	Immuno-suppressant	Low resting heart rate	Irregular heart beat	High BMI	Time of death
	✗	✗	✓	✗	✗	Day 2
	✓	✓	✗	✗	✓	Day 10
	✗	✗	✗	✓	✗	Day ≥ 6

Feature vector X

Observed time Y

When we stop collecting training data, not everyone has died!

Survival Analysis

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

Feature vector X

Observed time Y

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

Problem Setup

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

Problem Setup

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

1. Sample feature vector $X \sim \mathbb{P}_X$

Problem Setup

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

1. Sample feature vector $X \sim \mathbb{P}_X$
2. Sample time of death $T \sim \mathbb{P}_{T|X}$

Problem Setup

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

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 $C \sim \mathbb{P}_{C|X}$

Problem Setup

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

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 happens before censoring ($T \leq C$):

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

x

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

$x \to \boxed{\text{find } k \text{ training points closest to } x} \to k$ data points

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

x → [find k training points closest to x] → k data points → [Kaplan-Meier estimator] → $\widehat{S}(t \mid x)$

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

$x \rightarrow$ $\boxed{\text{find } k \text{ training points closest to } x}$ $\rightarrow$ k data points $\rightarrow$ $\boxed{\text{Kaplan-Meier estimator}}$ $\rightarrow \widehat{S}(t \mid x)$

Kernel variant is similar

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

x → find k training points closest to x → k data points → Kaplan-Meier estimator → $\widehat{S}(t \mid x)$

Kernel variant is similar

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

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

$x \rightarrow$ $\boxed{\text{find } k \text{ training points closest to } x}$ $\rightarrow$ k data points $\rightarrow$ $\boxed{\text{Kaplan-Meier estimator}}$ $\rightarrow$ $\widehat{S}(t \mid x)$

Kernel variant is similar

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

$\boxed{\text{Enough of the } n \text{ training data have } Y \text{ values} > \tau}$

Problem Setup

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

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 happens before censoring ($T \le C$): Set $Y = T, \delta = 1$

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

Estimator (Beran 1981):

Feature space is separable metric space (intrinsic dimension d)

x → find k training points closest to x → k data points → Kaplan-Meier estimator → $\widehat{S}(t \mid x)$

Kernel variant is similar

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

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

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Continuous r.v. in **time** & smooth w.r.t. **feature space** (Hölder index α)

Estimator (Beran 1981):

Feature space is separable metric space (intrinsic dimension d)

$x \rightarrow$ find k training points closest to x $\rightarrow$ k data points $\rightarrow$ Kaplan-Meier estimator $\rightarrow$ $\widehat{S}(t \mid x)$

Kernel variant is similar

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

Enough of the n training data have Y values > τ

Problem Setup

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

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 happens before censoring ($T \leq C$): Set $Y = T, \delta = 1$

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

Borel prob. measure

Continuous r.v. in **time** & smooth w.r.t. **feature space** (Hölder index α)

Estimator (Beran 1981):

Feature space is separable metric space (intrinsic dimension d)

$x \rightarrow$ [find k training points closest to x] $\rightarrow$ k data points $\rightarrow$ [Kaplan-Meier estimator] $\rightarrow \widehat{S}(t \mid x)$

Kernel variant is similar

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

Enough of the n training data have Y values > τ

Theory (Informal)

Theory (Informal)

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

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

Theory (Informal)

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

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

If no censoring, problem reduces to conditional CDF estimation

Theory (Informal)

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

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

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

Theory (Informal)

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

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

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:

Theory (Informal)

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

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

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

Theory (Informal)

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

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

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 ($-\log S(t \mid x)$)

Theory (Informal)

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

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

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 ($-\log S(t \mid x)$)
- Generalization bound for automatic *k* using validation data

Theory (Informal)

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

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

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 ($-\log S(t \mid x)$)
- Generalization bound for automatic *k* using validation data

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

Theory (Informal)

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

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

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

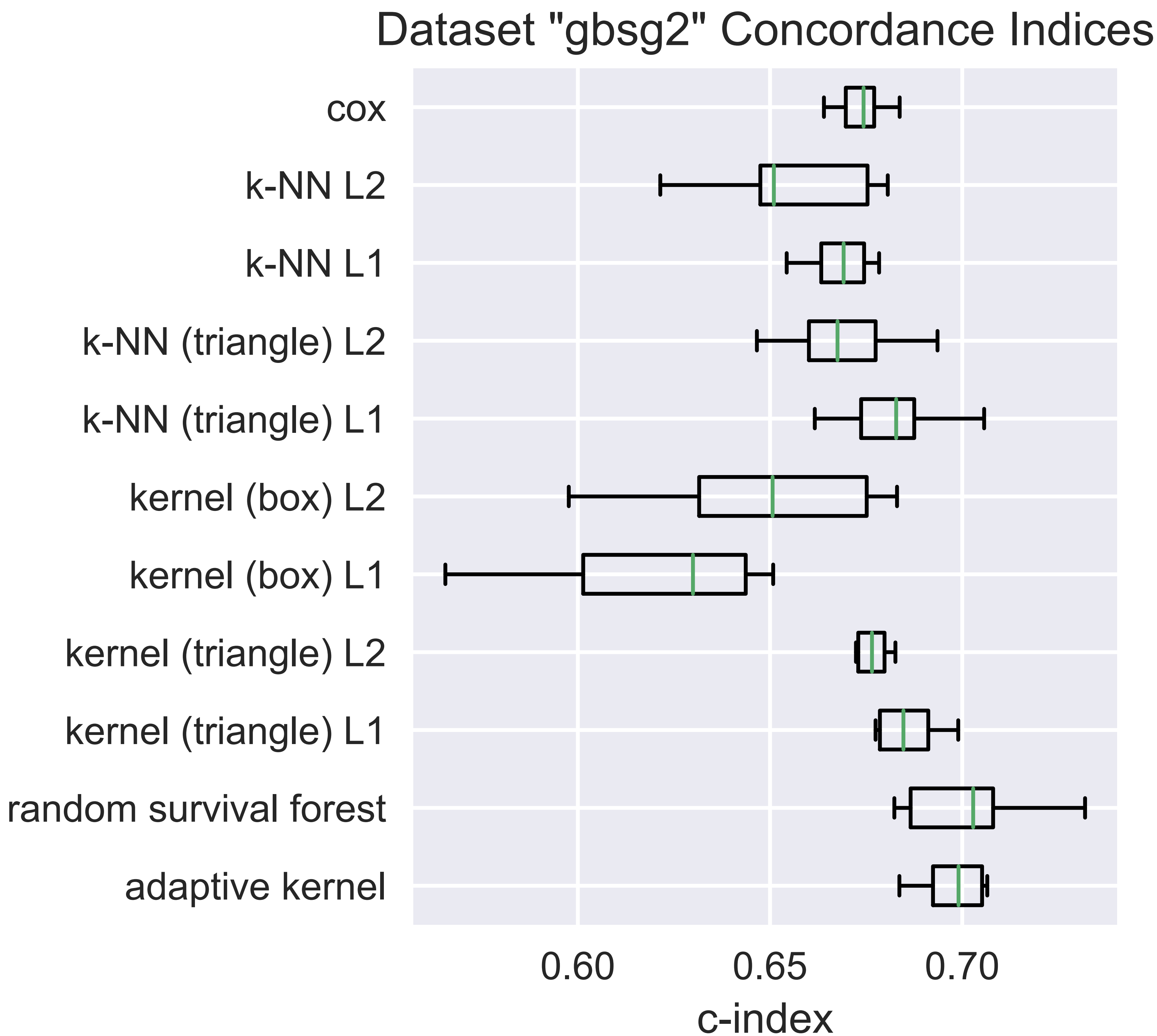
Proof ideas also give finite sample rates for:

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

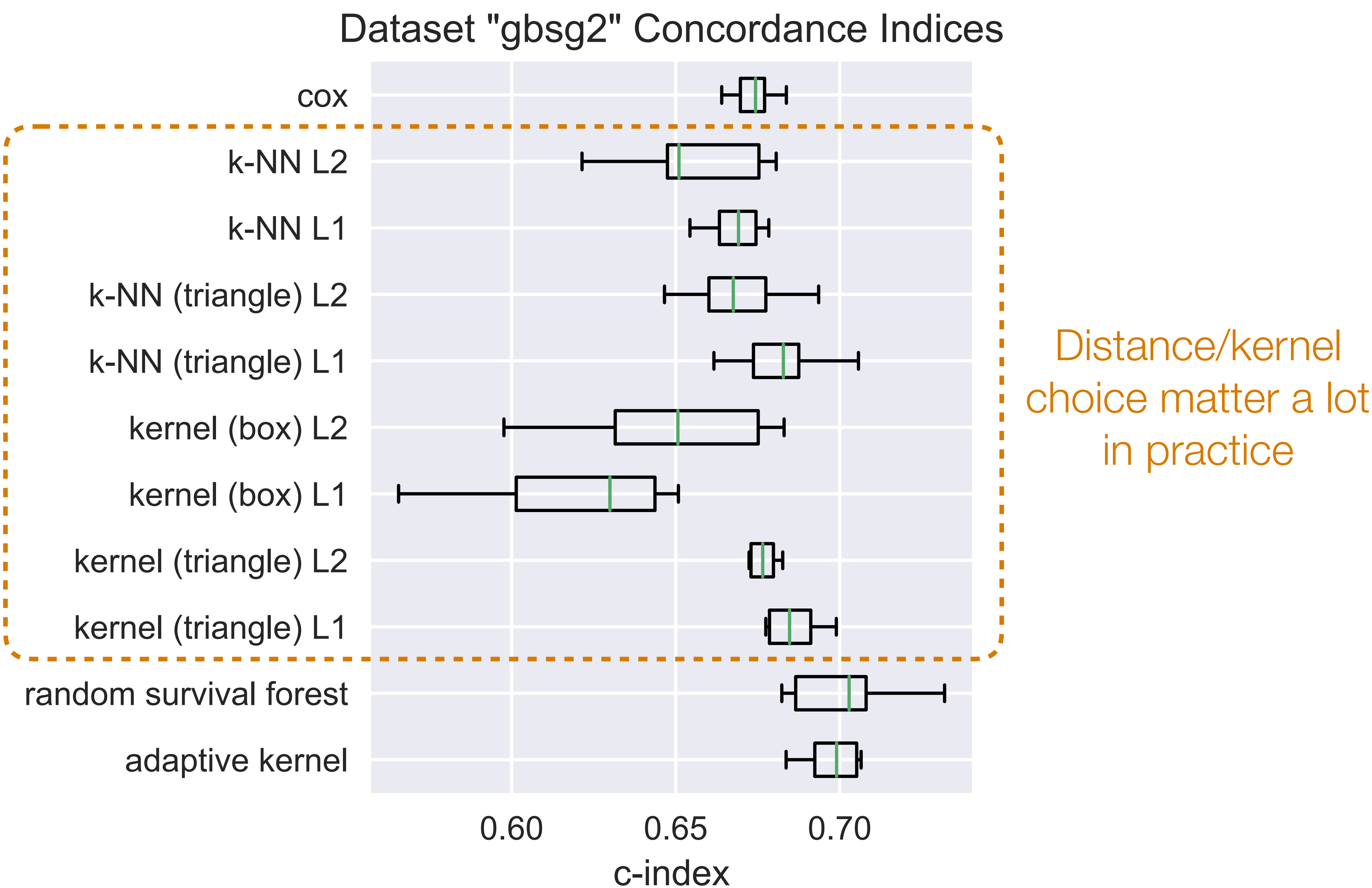
Most general finite sample theory for *k*-NN and kernel survival estimators

Existing kernel results only for Euclidean space
(Dabrowska 1989, Van Keilegom & Veraverbeke 1996, Van Keilegom 1998)

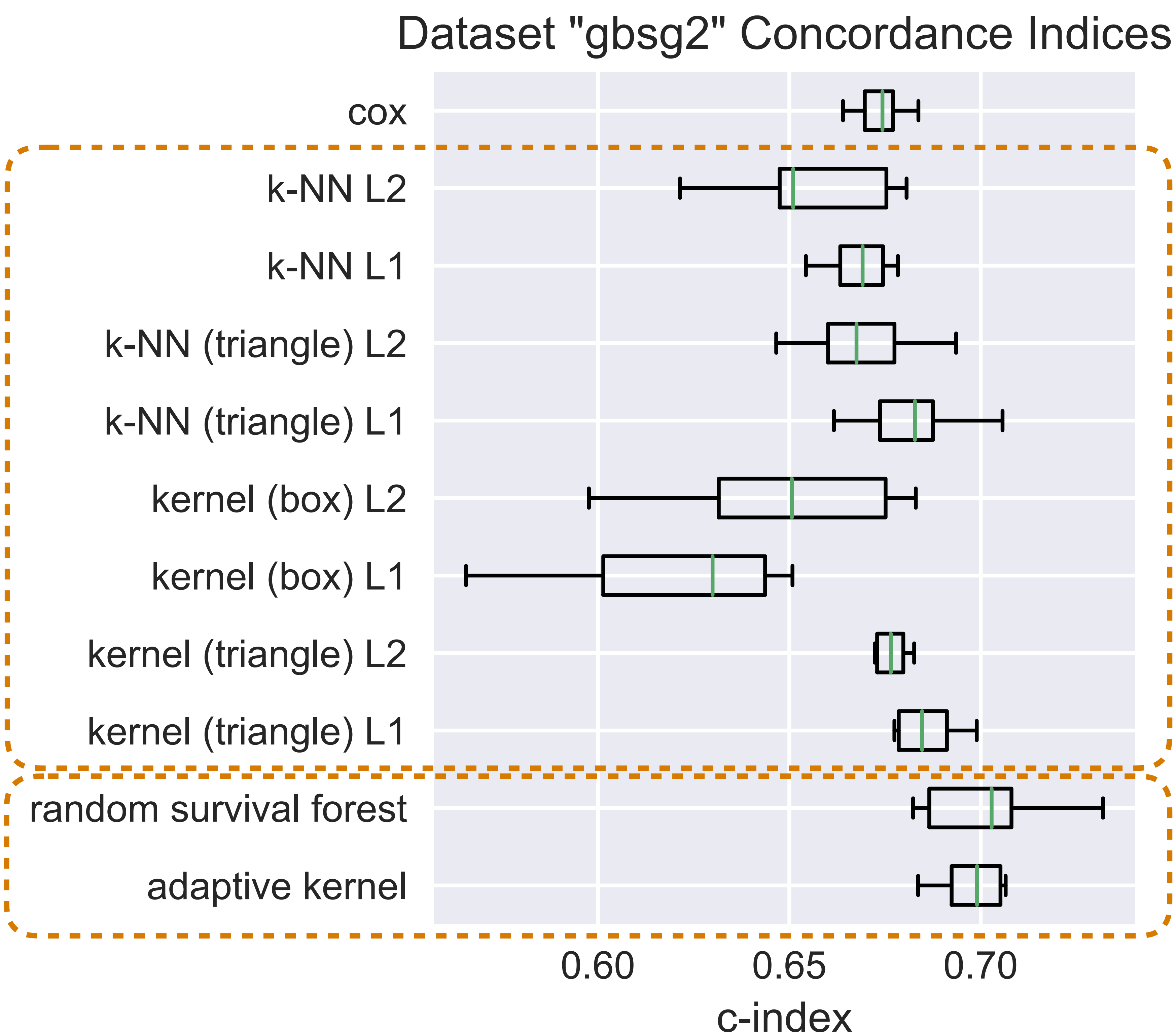
Experiments



Experiments



Experiments



Distance/kernel choice matter a lot in practice

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