Mathematical Properties of Continuous Ranked Probability Score Forecasting

Romain Pic[1], Clément Dombry[1], Philippe Naveau[2] et Maxime Taillardat[3]

[1]Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté

[2]Laboratoire des Sciences du Climat et de l'Environnement, Université de Versailles Saint-Quentin

[3]Centre National de Recherches Météorologiques, Météo France

Table of Contents

Probabilistic Forecasting

All those whose duty it is to issue regular daily forecasts know that there are times when they feel **very confident** and other times when they are **doubtful** as to coming weather. It seems to me that the condition of confidence or otherwise forms a **very important part of the prediction**.

Ernest Cook (MWR, 1906)

- Various approaches :
  - Ensemble prediction
  - Quantile regression
  - Expectile regression
  - **Distributional regression** : cumulative distribution function, density, quantile function, copula...

- Various approaches :
  - Ensemble prediction
  - Quantile regression
  - Expectile regression
  - **Distributional regression** : cumulative distribution function, density, quantile function, copula...

- How can we compare a distribution and an observation?

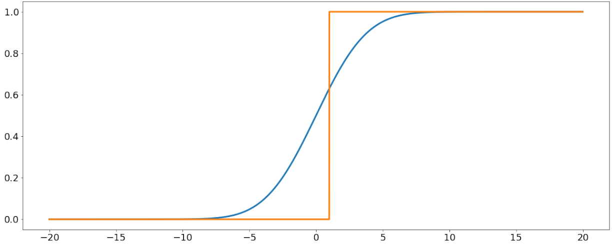
Probabilistic Forecasting Techniques

- Various approaches :
  - Ensemble prediction
  - Quantile regression
  - Expectile regression
  - **Distributional regression** : cumulative distribution function, density, quantile function, copula...
- How can we compare a distribution and an observation? $\rightarrow$ **Scoring Rules**

Continuous Ranked Probability Score

- Continuous Ranked Probability Score (CRPS) : [Matheson and Winkler, 1976]

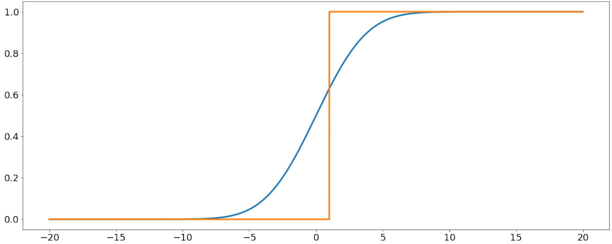
$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z$$



Continuous Ranked Probability Score

- Continuous Ranked Probability Score (CRPS) : [Matheson and Winkler, 1976]

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y \leq z})^2 \mathrm{d}z$$



- Difference of expected scores :

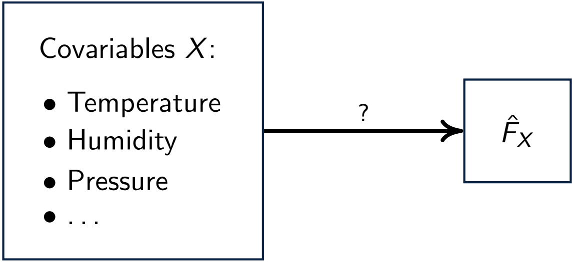
$$\overline{\mathrm{CRPS}}(F, G) - \overline{\mathrm{CRPS}}(G, G) = \int_{\mathbb{R}} (F(z) - G(z))^2 \mathrm{d}z$$

$$\overline{\mathrm{CRPS}}(F, G) \geq \overline{\mathrm{CRPS}}(G, G) \text{ (strictly proper)}$$

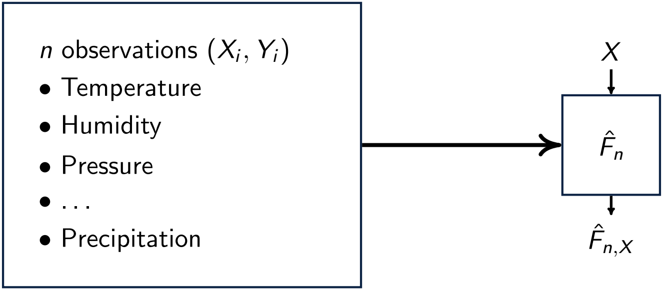
Table of Contents

Theoretical framework

- $Y \in \mathbb{R}$ variable of interest, $X \in \mathbb{R}^d$ covariables with $(X, Y) \sim P$.
- Goal : estimate the conditional distribution of Y given X, noted F_X^*.



- In practice : estimate the conditional distribution of Y given X based on $\boldsymbol{n}$ **observations** $D_n = \{(X_i, Y_i), i \in [\![1; n]\!]\}$ where (X_i, Y_i) are assumed i.i.d. following P.



- Verification with the CRPS

Applications of this framework

Methods concerned by the framework :

Applications of this framework

Methods concerned by the framework :

- Not only methods based on the minimization of the CRPS, also methods using the CRPS for verification.

Applications of this framework

Methods concerned by the framework :

- Not only methods based on the minimization of the CRPS, also methods using the CRPS for verification.
- Predict a parametric or nonparametric distribution : Censored-Shifted Gamma, Censored-GEV, EGPD, QRF, Bernstein polynomials...

Applications of this framework

Methods concerned by the framework :

- Not only methods based on the minimization of the CRPS, also methods using the CRPS for verification.
- Predict a parametric or nonparametric distribution : Censored-Shifted Gamma, Censored-GEV, EGPD, QRF, Bernstein polynomials...
- Predicted distribution represented as an ensemble of values : Random/Quantile Ensembles, Generators...

Convergence and rate of convergence

$$R_P(\hat{F}_n) = \overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*)$$

Convergence and rate of convergence

$$R_P(\hat{F}_n) = \overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*)$$

$$R_P(F^*) = \overline{\mathrm{CRPS}}(F_X^*, F_X^*)$$

Convergence and rate of convergence

$$R_P(\hat{F}_n) = \overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*)$$

$$R_P(F^*) = \overline{\mathrm{CRPS}}(F_X^*, F_X^*)$$

- Rate of convergence for a given class of distributions ?

Convergence and rate of convergence

$$R_P(\hat{F}_n) = \overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*)$$

$$R_P(F^*) = \overline{\mathrm{CRPS}}(F_X^*, F_X^*)$$

- Rate of convergence for a given class of distributions ?
- Minimization of the maximal error on a class of distributions. (minimax error)

Convergence and rate of convergence

$$R_P(\hat{F}_n) = \overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*)$$

$$R_P(F^*) = \overline{\mathrm{CRPS}}(F_X^*, F_X^*)$$

- Rate of convergence for a given class of distributions ?
- Minimization of the maximal error on a class of distributions. (minimax error)

Definition

A sequence of positive numbers $(\boldsymbol{a_n})$ is called an **optimal minimax rate of convergence** on the class $\mathcal{D}$ if

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} > 0 \qquad \text{(L)}$$

and

$$\limsup_{n\to\infty} \inf_{\hat{F}_n} \sup_{P\in\mathcal{D}} \frac{\mathbb{E}_{D_n\sim P^n}[R_P(\hat{F}_n)] - R_P(F^*)}{a_n} < \infty, \qquad \text{(U)}$$

where the infimum is taken over all distributional regression models $\hat{F}_n$ trained on D_n.

Consider the following classes :

Definition

For $h \in (0, 1]$, $C > 0$ and $M > 0$, let $\mathcal{D}^{(h,C,M)}$ be the class of distributions P such that $F^*_x(y) = P(Y \leq y | X = x)$ satisfies :

i) $X \in [0, 1]^d$ P_X-a.s.;

ii) For all $x \in [0, 1]^d$, $\int_{\mathbb{R}} F^*_x(z)(1 - F^*_x(z))\mathrm{d}z \leq M$;

iii) $\|F^*_{x'} - F^*_x\|_{L^2} \leq C\|x' - x\|^h$ for all $x, x' \in [0, 1]^d$.

Remark : Conditions similar to point regression [Györfi et al., 2002].

Class of distributions

i) $\boldsymbol{X \in [0,1]^d}$ $\boldsymbol{P_X}$**-a.s.;**
$\rightarrow$ More generally a compact.

Class of distributions

i) $\boldsymbol{X \in [0,1]^d}$ **$\boldsymbol{P_X}$-a.s.;**
$\rightarrow$ More generally a compact.

ii) For all $\boldsymbol{x \in [0,1]^d}$, $\boldsymbol{\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z \leq M}$;
$\rightarrow$ The dispersion of $Y|X = x$ remains bounded for all $x \in [0,1]^d$.

i) $\boldsymbol{X} \in [\mathbf{0}, \mathbf{1}]^{\boldsymbol{d}}$ $\boldsymbol{P_X}$**-a.s.;**
$\rightarrow$ More generally a compact.

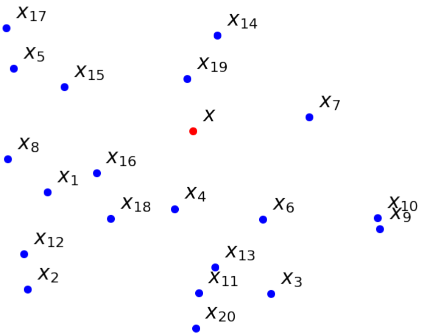
ii) **For all** $\boldsymbol{x} \in [\mathbf{0}, \mathbf{1}]^{\boldsymbol{d}}$, $\int_{\mathbb{R}} \boldsymbol{F_x^*(z)(1 - F_x^*(z))}\mathrm{d}\boldsymbol{z} \leq \boldsymbol{M}$**;**
$\rightarrow$ The dispersion of $Y|X = x$ remains bounded for all $x \in [0, 1]^d$.

iii) $\boldsymbol{\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^h}$ **for all** $\boldsymbol{x, x'} \in [\mathbf{0}, \mathbf{1}]^{\boldsymbol{d}}$.

$$\overline{\mathrm{CRPS}}(\hat{F}_{n,X}, F_X^*) - \overline{\mathrm{CRPS}}(F_X^*, F_X^*) = \int_{\mathbb{R}} (\hat{F}_{n,X}(z) - F_X^*(z))^2 \mathrm{d}z = \|\hat{F}_{n,X} - F_X^*\|_{L^2}$$

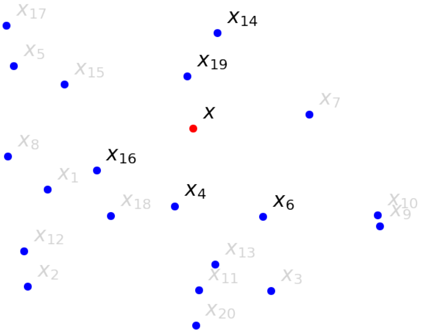
Using knowledge from previous observations at $X = x_i$ to extrapolate the value at $X = x$ $\rightarrow$ Need regularity of F^*.

Table of Contents



$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{y_{i:n}(x) \leq z}, \quad {}_{i:n}(x) \text{ index of the } i\text{-th nearest neighbor of } x.$$

k-Nearest Neighbors



$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{y_{i:n}(x) \leq z}, \quad {}_{i:n}(x) \text{ index of the } i\text{-th nearest neighbor of } x.$$

k-Nearest Neighbors

- (L) Use a subclass with a binary response to obtain a **lower minimax rate of convergence** : $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$.

• (L) Use a subclass with a binary response to obtain a **lower minimax rate of convergence** : $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$.

Proposition

Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the k-NN model. Then,

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \begin{cases} 8^h C^2 \left(\dfrac{k_n}{n}\right)^h + \dfrac{M}{k_n} & \text{if } d = 1, \\ {c_d}^h C^2 \left(\dfrac{k_n}{n}\right)^{2h/d} + \dfrac{M}{k_n} & \text{if } d \geq 2, \end{cases}$$

where $c_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$ and V_d is the volume of the unit ball in $\mathbb{R}^d$.

Theorem

For $d \geq 2$, the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$ is $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$. Moreover, the k-NN algorithm reaches the optimal rate of convergence for

$$k_n = \left(\frac{Md}{2hC^2c_d^h}\right)^{\frac{d}{2h+d}} n^{\frac{2h}{2h+d}}.$$

k-Nearest Neighbors

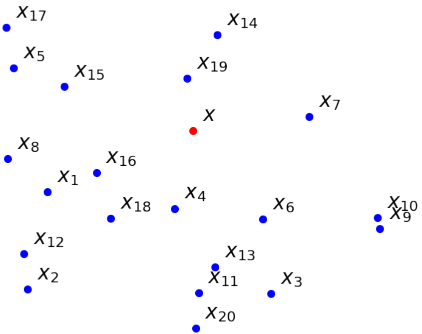
Theorem

For $d \geq 2$, the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$ is $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$. Moreover, the k-NN algorithm reaches the optimal rate of convergence for

$$k_n = \left(\frac{Md}{2hC^2c_d^h}\right)^{\frac{d}{2h+d}} n^{\frac{2h}{2h+d}}.$$

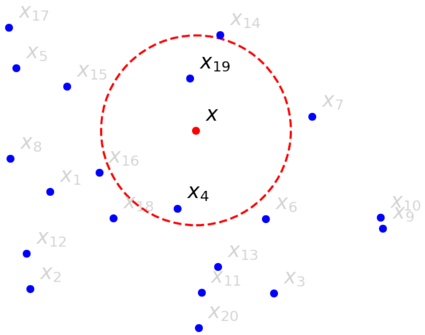
- What happens in $d = 1$?
- Interesting result but k-NN not used in practice.

Kernel Method



$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h_n})\mathbb{1}_{y_i \leq z}}{\sum_{i=1}^{n} K(\frac{x-x_i}{h_n})}, \text{ with } K(z) = \mathbb{1}_{\{\|z\| \leq 1\}}.$$

Kernel Method



$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)\mathbb{1}_{y_i \leq z}}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_n}\right)}, \text{ with } K(z) = \mathbb{1}_{\{\|z\| \leq 1\}}.$$

Kernel Method

- (L) Same **lower minimax rate of convergence** as previously : $a_n = n^{-\frac{2h}{2h+d}}$.

Kernel Method

- (L) Same **lower minimax rate of convergence** as previously : $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$.

Proposition

Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the naive kernel model. Then,

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \tilde{c}_d \frac{2M + Cd^{h/2} + \frac{M}{n}}{nh_n^d} + C^2 h_n^{2h}$$

where $\tilde{c}_d$ only depends on $d \geq 1$.

- (L) Same **lower minimax rate of convergence** as previously : $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$.

Proposition

Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the naive kernel model. Then,

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \tilde{c}_d \frac{2M + Cd^{h/2} + \frac{M}{n}}{nh_n^d} + C^2 h_n^{2h}$$

where $\tilde{c}_d$ only depends on $d \geq 1$.

Theorem

For any d, the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$ is $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$. Moreover, the naive kernel algorithm reaches the optimal rate of convergence for $h_n = \left(\frac{\tilde{c}_d d(M + Cd^{h/2} + \frac{M}{n})}{2hC^2}\right)^{\frac{1}{2h+d}} n^{-\frac{1}{2h+d}}$.

- (L) Same **lower minimax rate of convergence** as previously : $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$.

Proposition

Assume $P \in \mathcal{D}^{(h,C,M)}$ and let $\hat{F}_n$ be the naive kernel model. Then,

$$\mathbb{E}_{D_n \sim P^n}[R_P(\hat{F}_n)] - R_P(F^*) \leq \tilde{c}_d \frac{2M + Cd^{h/2} + \frac{M}{n}}{nh_n^d} + C^2 h_n^{2h}$$

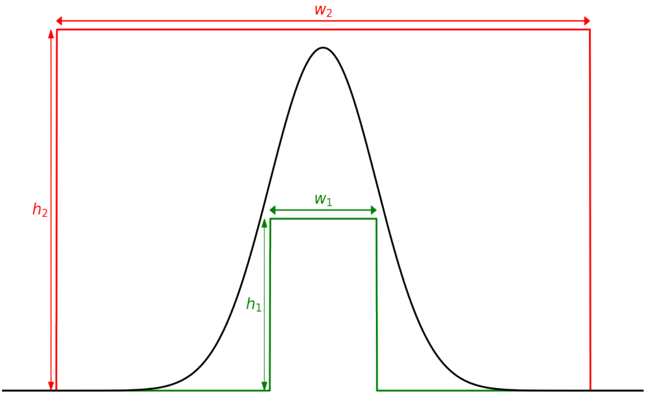
where $\tilde{c}_d$ only depends on $d \geq 1$.

Theorem

For any d, the optimal minimax rate of convergence on the class $\mathcal{D}^{(h,C,M)}$ is $\boldsymbol{a_n = n^{-\frac{2h}{2h+d}}}$. Moreover, the naive kernel algorithm reaches the optimal rate of convergence for $h_n = \left(\frac{\tilde{c}_d d(M + Cd^{h/2} + \frac{M}{n})}{2hC^2}\right)^{\frac{1}{2h+d}} n^{-\frac{1}{2h+d}}$.

- Optimal minimax rate of convergence for any d.
- Used in practice ?

Boxed Kernel



Conclusion

- Optimal minimax rate of convergence for distributional regression.
- Upper bound on the convergence rate for k-NN and kernel methods at fixed n.
- Extension to usual weighted CRPSs.
- Perspectives :
    - Study other algorithms : Random Forests (e.g. QRF).
    - Study other definitions of convergence : other distances.
    - Adapt other classical results to the distributional regression framework.

Preprint : Mathematical Properties of Continuous Ranked Probability Score Forecasting, Pic et al. (https://arxiv.org/abs/2205.04360)

Bremnes, John Bjørnar (2020). "Ensemble post-processing using quantile function regression based on neural networks and Bernstein polynomials". In: *Monthly Weather Review* 148 (1). DOI: 10.1175/MWR-D-19-0227.1.

Gneiting, Tilmann and Matthiass Katzfuss (2014). "Probabilistic Forecasting". In: *Annual Review of Statistics and its Applications*. DOI: 10.1146/annurev-statistics-062713-085831.

Györfi, Lászlò et al. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer.

Matheson, James E. and Robert L. Winkler (1976). "Scoring Rules for Continuous Probability Distributions". In: *Management Science* 22 (10). DOI: 10.2307/2629907.

Naveau, Philippe et al. (2016). "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection". In: *Water Resources Research4* 52 (4). DOI: 10.1002/2015wr018552.

Scheuerer, Michael and Thomas M. Hamill (2015). "Statistical Post-Processing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions". In: *Monthly Weather Review* 143 (11). DOI: 10.1175/MWR-D-15-0061.1.

References II

Schulz, Benedikt and Sebastian Lerch (2021). *Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison*. arXiv:2106.09512. eprint: 2106.09512 (stat.ML).

Taillardat, Maxime et al. (2016). "Calibrated Ensemble Forecasts using Quantile Regression Forests and Ensemble Model Output Statistics.". In: *Monthly Weather Review* 144 (6). DOI: 10.1175/MWR-D-15-0260.1.

Zamo, Michaël and Philippe Naveau (2018). "Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts". In: *Mathematical Geosciences* 50.2, pp. 209–234. DOI: 10.1007/s11004-017-9709-7.

Simulations

- $X_1, X_2 \sim \mathcal{U}([0,1])$
- $Y = X_1 + X_2 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$
- $Y|X \sim \mathcal{N}(X_1 + X_2, \sigma^2)$

Simulations

- $X_1, X_2 \sim \mathcal{U}([0,1])$
- $Y = X_1 + X_2 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$
- $Y|X \sim \mathcal{N}(X_1 + X_2, \sigma^2)$
- Checking the conditions :

Simulations

- $X_1, X_2 \sim \mathcal{U}([0,1])$
- $Y = X_1 + X_2 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$
- $Y|X \sim \mathcal{N}(X_1 + X_2, \sigma^2)$
- Checking the conditions :

  i) $\boldsymbol{X \in [0,1]^d}$ $\boldsymbol{P_X}$**-a.s.;** ✓

  ii) For all $\boldsymbol{x \in [0,1]^d}$**,** $\boldsymbol{\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z \leq M}$**;**
  $\rightarrow M = \frac{\sigma}{\sqrt{\pi}}$ ✓

  iii) $\boldsymbol{\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^h}$ **for all** $\boldsymbol{x, x' \in [0,1]^d}$.
  $\rightarrow$ Hard to get optimal values for C and h but $h = 1$ works. ✓

Simulations

- $X_1, X_2 \sim \mathcal{U}([0,1])$
- $Y = X_1 + X_2 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$
- $Y|X \sim \mathcal{N}(X_1 + X_2, \sigma^2)$
- Checking the conditions :
  i) $\boldsymbol{X \in [0,1]^d}$ **P_X-a.s.;** ✓
  ii) **For all** $\boldsymbol{x \in [0,1]^d}$, $\boldsymbol{\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z \leq M}$**;**
  $\rightarrow M = \frac{\sigma}{\sqrt{\pi}}$ ✓
  iii) $\boldsymbol{\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^h}$ **for all** $\boldsymbol{x, x' \in [0,1]^d}$.
  $\rightarrow$ Hard to get optimal values for C and h but $h = 1$ works. ✓
- k-NN :

$$\hat{F}_{n,x}(z) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{1}_{y_{i:n}(x) \leq z}$$

Simulations

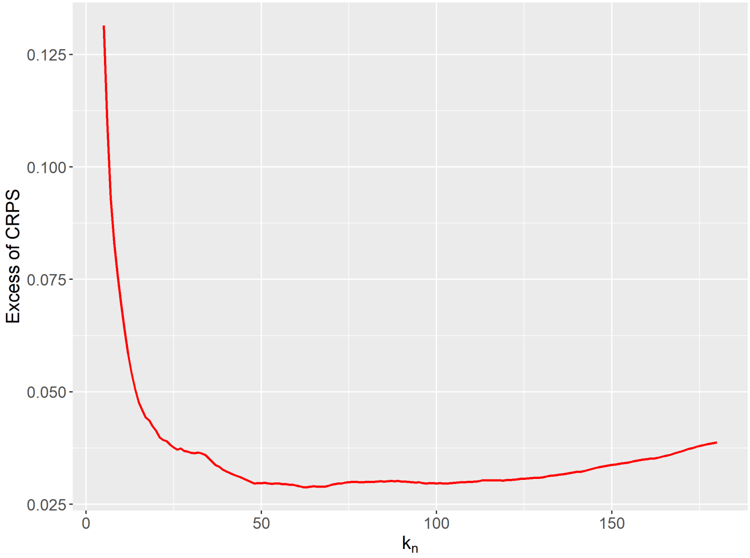
- $X_1, X_2 \sim \mathcal{U}([0,1])$
- $Y = X_1 + X_2 + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0,1)$
- $Y|X \sim \mathcal{N}(X_1 + X_2, \sigma^2)$
- Checking the conditions :
  i) $\boldsymbol{X \in [0,1]^d}$ $\boldsymbol{P_X}$**-a.s.;** ✓
  ii) **For all** $\boldsymbol{x \in [0,1]^d}$**,** $\boldsymbol{\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z \leq M}$**;**
  $\rightarrow M = \frac{\sigma}{\sqrt{\pi}}$ ✓
  iii) $\boldsymbol{\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^h}$ **for all** $\boldsymbol{x, x' \in [0,1]^d}$.
  $\rightarrow$ Hard to get optimal values for C and h but $h = 1$ works. ✓
- k-NN :

$$\hat{F}_{n,x}(z) = \frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}_{y_{i:n}(x) \leq z}$$

$$\overline{\mathrm{CRPS}}(F_{n,x}, F_x^*) - \overline{\mathrm{CRPS}}(F_x^*, F_x^*) = \int_{\mathbb{R}} \left(\frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{1}_{y_{i:n}(x) \leq z} - \Phi\left(\frac{z - (x_1 + x_2)}{\sigma}\right)\right)^2 \mathrm{d}z$$

CRPS vs. k_n

Parameters : $\sigma = 1$ and $n = 200$.

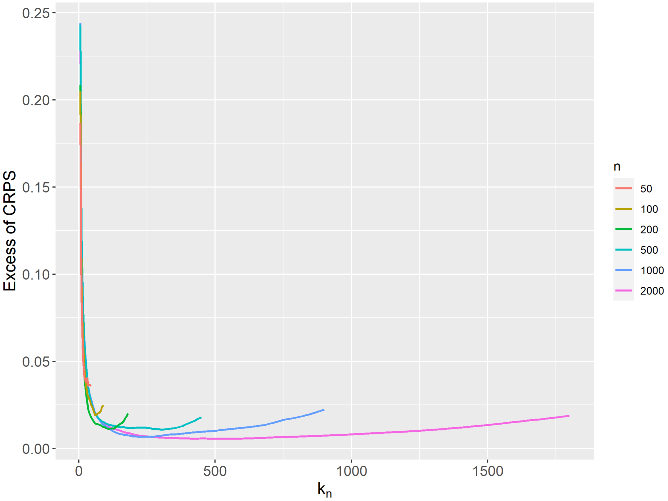


Scaling of k_n with n, $\sigma = 2$

$$k_n \propto n^{\frac{2h}{2h+d}}$$

Scaling of k_n with n, $\sigma = 2$

$$k_n \propto n^{\frac{2h}{2h+d}}$$



Scaling of k_n with n, $\sigma = 2$

$$k_n \propto n^{\frac{2h}{2h+d}}$$

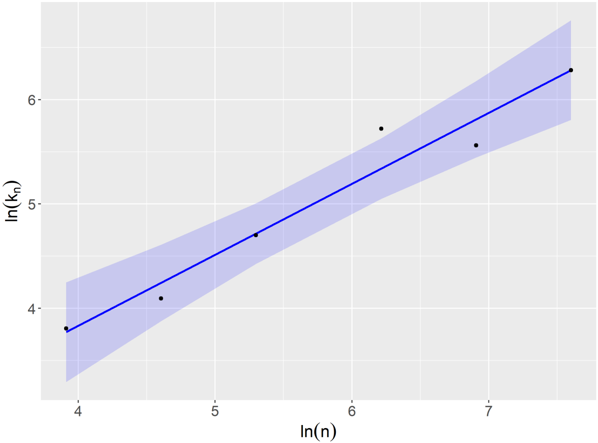


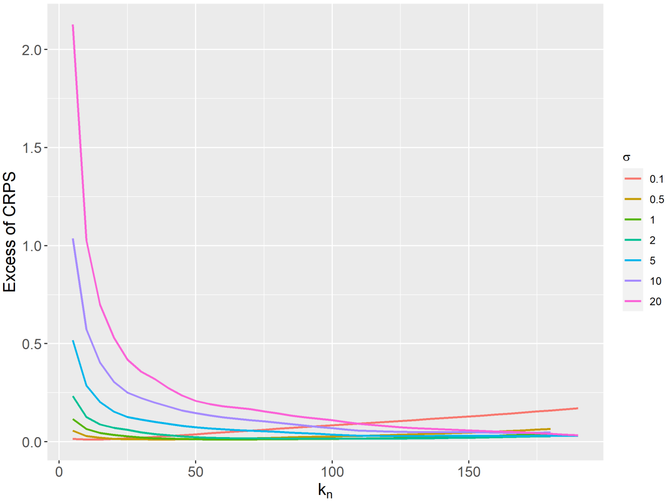
Figure: Equation : $y = 1.1 + 0.68x, \; R^2 = 0.952$

Scaling of k_n with σ, $n = 200$

$$k_n \propto M^{\frac{d}{2h+d}} \propto \sigma^{\frac{d}{2h+d}}$$

Scaling of k_n with σ, $n = 200$

$$k_n \propto M^{\frac{d}{2h+d}} \propto \sigma^{\frac{d}{2h+d}}$$



Scaling of k_n with σ, $n = 200$

$$k_n \propto M^{\frac{d}{2h+d}} \propto \sigma^{\frac{d}{2h+d}}$$

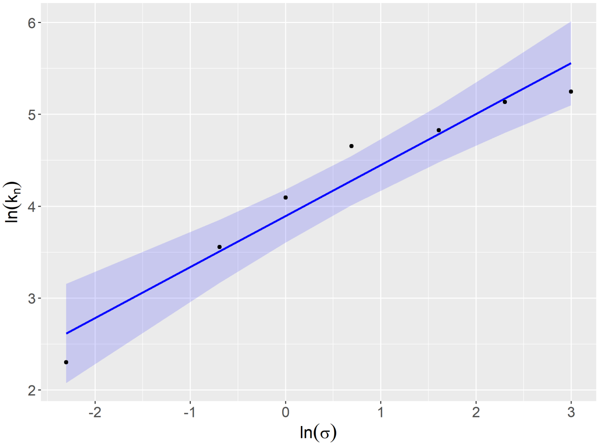


Figure: Equation : $y = 3.9 + 0.56x$, $R^2 = 0.942$