Robustness versus consistency
in ill-posed classification and regression problems

Robert Hable, Andreas Christmann

Robert Hable
Department of Mathematics
University of Bayreuth

Parametric Statistical Problem:

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Parametric Model:

$$P_0 \in \mathcal{P} = \left\{ P_\theta \,\middle|\, \theta \in \Theta \right\}$$

Goal: Estimation of the true $\theta_0 \in \Theta$

Parametric Statistical Problem:

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Parametric Model:

$$P_0 \in \mathcal{P} = \left\{ P_\theta \,\middle|\, \theta \in \Theta \right\}$$

Goal: Estimation of the true $\theta_0 \in \Theta$

Functional Formalization:

$$T : \mathcal{P} \to \mathbb{R}^k, \qquad P_\theta \mapsto \theta$$

Parametric Statistical Problem:

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Parametric Model:

$$P_0 \in \mathcal{P} = \left\{ P_\theta \,\middle|\, \theta \in \Theta \right\}$$

Goal: Estimation of the true $\theta_0 \in \Theta$

Functional Formalization:

$$T : \mathcal{P} \to \mathbb{R}^k, \qquad P_\theta \mapsto \theta$$

Example: $P_\theta = \mathcal{N}(\theta, 1)$, $\theta = T(P_\theta) = \int z \, P_\theta(dz)$

Non-Parametric Statistical Problem

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Non-Parametric Model:

$$P_0 \in \mathcal{P} = \text{a large set of probability measures}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathbb{R}^k, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$

Non-Parametric Statistical Problem

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Non-Parametric Model:

$$P_0 \in \mathcal{P} = \text{a large set of probability measures}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathbb{R}^k, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$

Example: $T(P) = \int z \, P(dz)$

$$\mathcal{P} = \left\{ P \;\middle|\; \int |z| \, P(dz) < \infty \right\}$$

Non-Parametric Statistical Problem

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Non-Parametric Model:

$$P_0 \in \mathcal{P} = \text{a large set of probability measures}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathbb{R}^k, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$

Example: $T(P) = \int z \, P(dz)$

$$\mathcal{P} = \left\{ P \;\middle|\; \int |z| \, P(dz) < \infty \right\}$$

Non-Parametric Statistical Problem

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Non-Parametric Model:

$$P_0 \in \mathcal{P} = \text{a large set of probability measures}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$

Non-Parametric Statistical Problem

$$Z_1, \ldots, Z_n \sim P_0 \qquad \text{i.i.d.}$$

Non-Parametric Model:

$$P_0 \in \mathcal{P} = \text{a large set of probability measures}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$

Example: $T(P) =$ the λ-density of P

$$\mathcal{P} = \left\{ P \,\middle|\, P \text{ has a } \lambda\text{-density} \right\}$$

Non-Parametric Regression

$$(X_1, Y_1), \ldots, (X_n, Y_n) \;\sim\; P_0 \qquad \text{i.i.d.}$$

Regression:

$$y_i \;=\; f_0(x_i) + \varepsilon_i\,, \qquad i \in \{1, \ldots, n\}$$

Functional Formalization:

$$T \;:\; \mathcal{P} \;\to\; \mathcal{F}\,, \qquad P \;\mapsto\; T(P)$$

- $\mathcal{F}$ = a large set of functions $f : x \mapsto f(x)$
- $T(P) = f \;:\; x \mapsto \int y\, P(dy|x)$

Non-Parametric Classification

$$(X_1, Y_1), \ldots, (X_n, Y_n) \quad \sim \quad P_0 \qquad \text{i.i.d.}$$

Classification:

$$Y_i \in \{0, 1\}, \qquad i \in \{1, \ldots, n\}$$

Functional Formalization:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

- $\mathcal{F}$ = a large set of functions $f : x \mapsto f(x)$
- $T(P) = f : x \mapsto P(Y = 1 \,|\, X = x)$

Good Estimators

Observations: $Z_1, \ldots, Z_n \sim P_0$ i.i.d.

Statistical functional:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$ (the true P_0 is unknown)

Good Estimators

Observations: $Z_1, \ldots, Z_n \sim P_0$ i.i.d.

Statistical functional:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$ (the true P_0 is unknown)

Desirable properties of an estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

are

Good Estimators

Observations: $Z_1, \ldots, Z_n \sim P_0$ i.i.d.

Statistical functional:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$ (the true P_0 is unknown)

Desirable properties of an estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

are

- Consistency: $S_n \xrightarrow{P_0} T(P_0)$ for $n \to \infty$

Good Estimators

Observations: $Z_1, \ldots, Z_n \sim P_0$ i.i.d.

Statistical functional:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$ (the true P_0 is unknown)

Desirable properties of an estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

are

- Consistency: $S_n \xrightarrow{P_0} T(P_0)$ for $n \to \infty$
- Robustness

Qualitative Robustness

Small errors in the data should not change the results too much.

Qualitative Robustness

Small errors in the data should not change the results too much.

- "Small errors in the data"
  - Small errors in many of the data points (rounding etc.)
  - Large errors in a few data points (gross errors, outliers)

Qualitative Robustness

Small errors in the data should not change the results too much.

- "Small errors in the data"
  - Small errors in many of the data points (rounding etc.)
  - Large errors in a few data points (gross errors, outliers)
- "should not change the results too much"
  i.e.: the distribution of the estimator is hardly affected

(distribution of the estimator = performance of the estimator)

Qualitative Robustness

Small errors in the data should not change the results too much.

- "Small errors in the data"
  - Small errors in many of the data points (rounding etc.)
  - Large errors in a few data points (gross errors, outliers)
- "should not change the results too much"
  i.e.: the distribution of the estimator is hardly affected

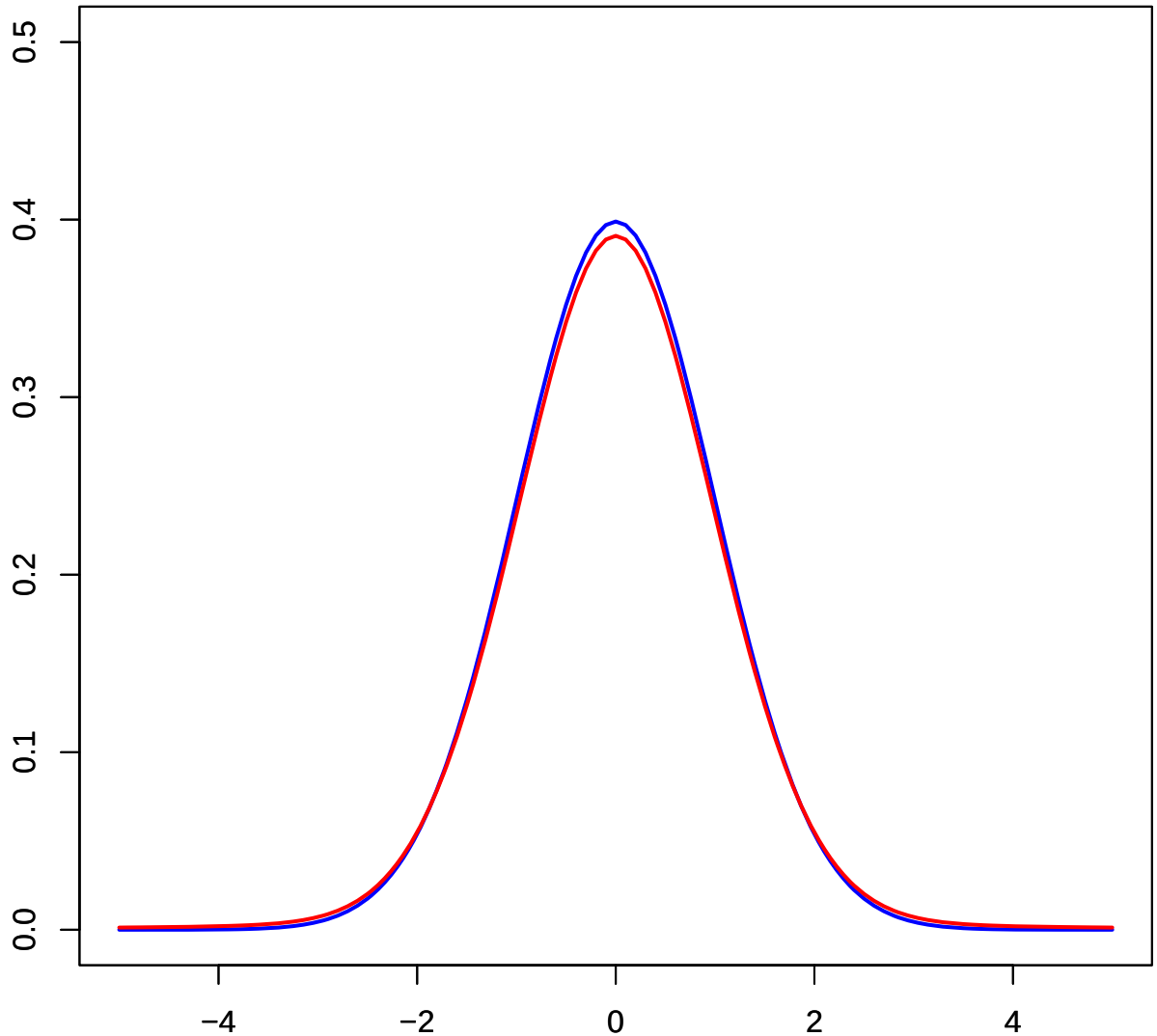
(distribution of the estimator $=$ performance of the estimator)

Qualitative Robustness: (Hampel ,1971)
A sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called qualitatively robust if

$$\forall\, P \;\; \forall\, \epsilon > 0 \;\; \exists\, \delta > 0 \;\text{ such that }\; \forall\, Q \;\text{ with }\; d_{\text{Pro}}(Q, P) < \delta$$

$$\sup_{n\in\mathbb{N}} d_{\text{Pro}}\big(S_n(Q^n), S_n(P^n)\big) < \varepsilon$$

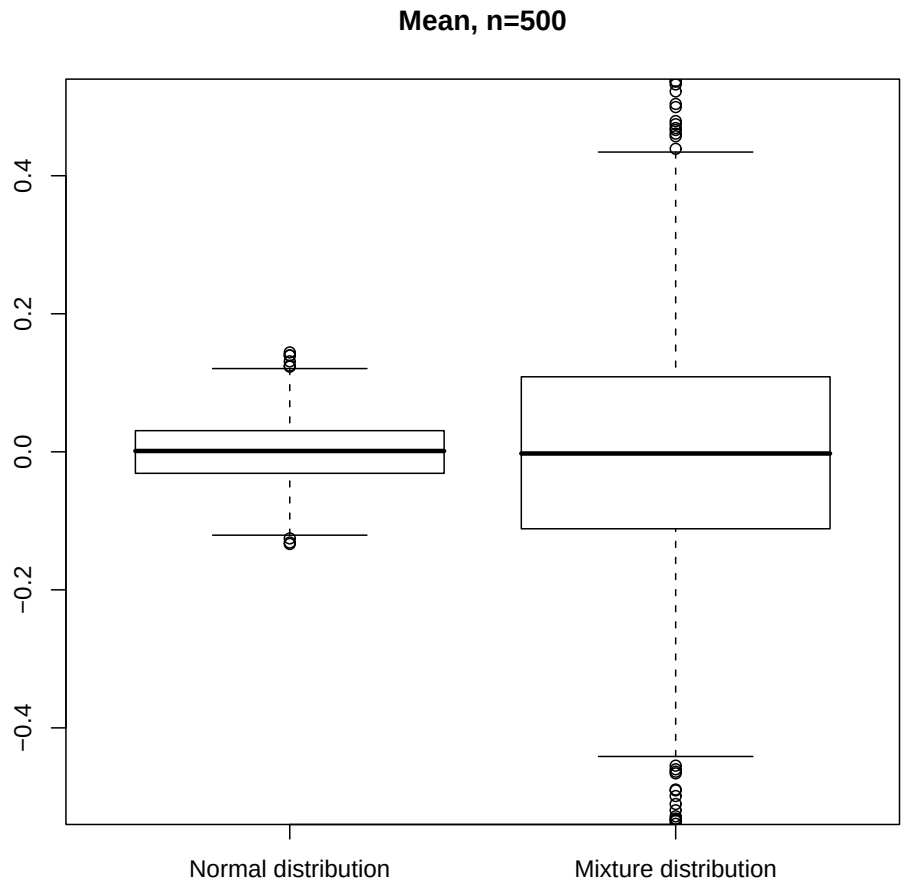
Qualitative Robustness – Parametric Example



Qualitative Robustness – Parametric Example

"mean" applied in 1000 runs

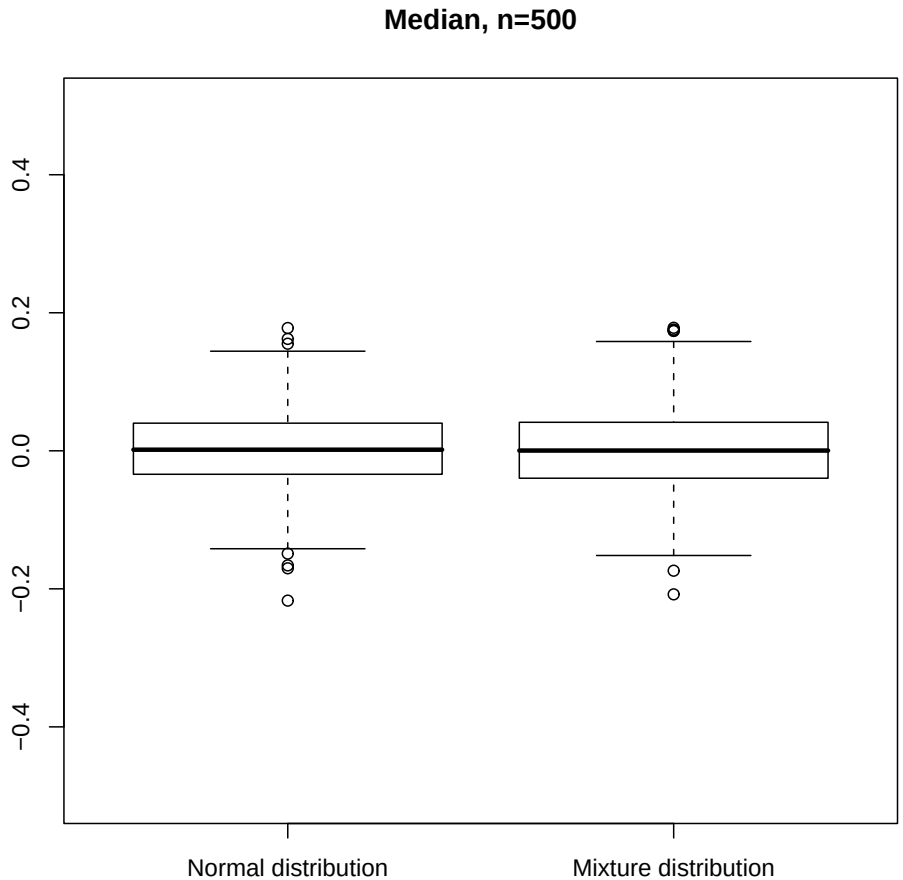
each run consists of a sample with 500 data points



Qualitative Robustness – Parametric Example

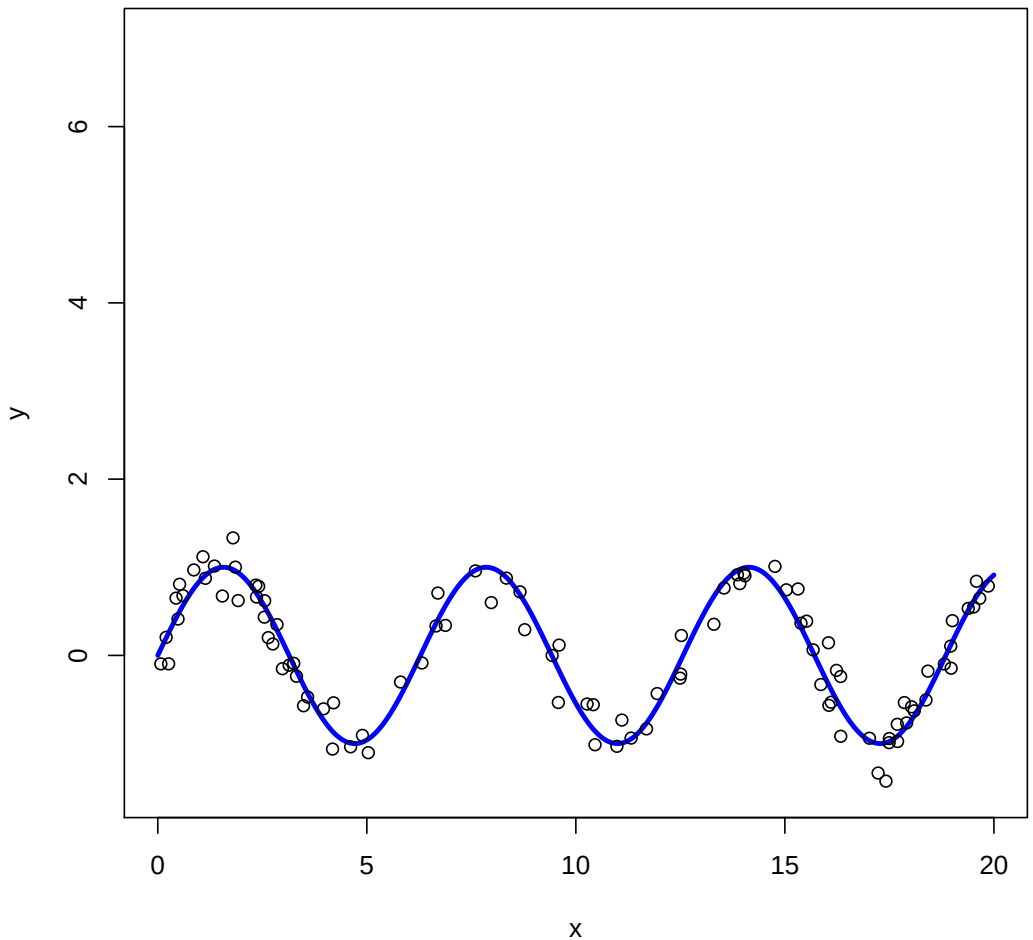
"median" applied in 1000 runs

each run consists of a sample with 500 data points



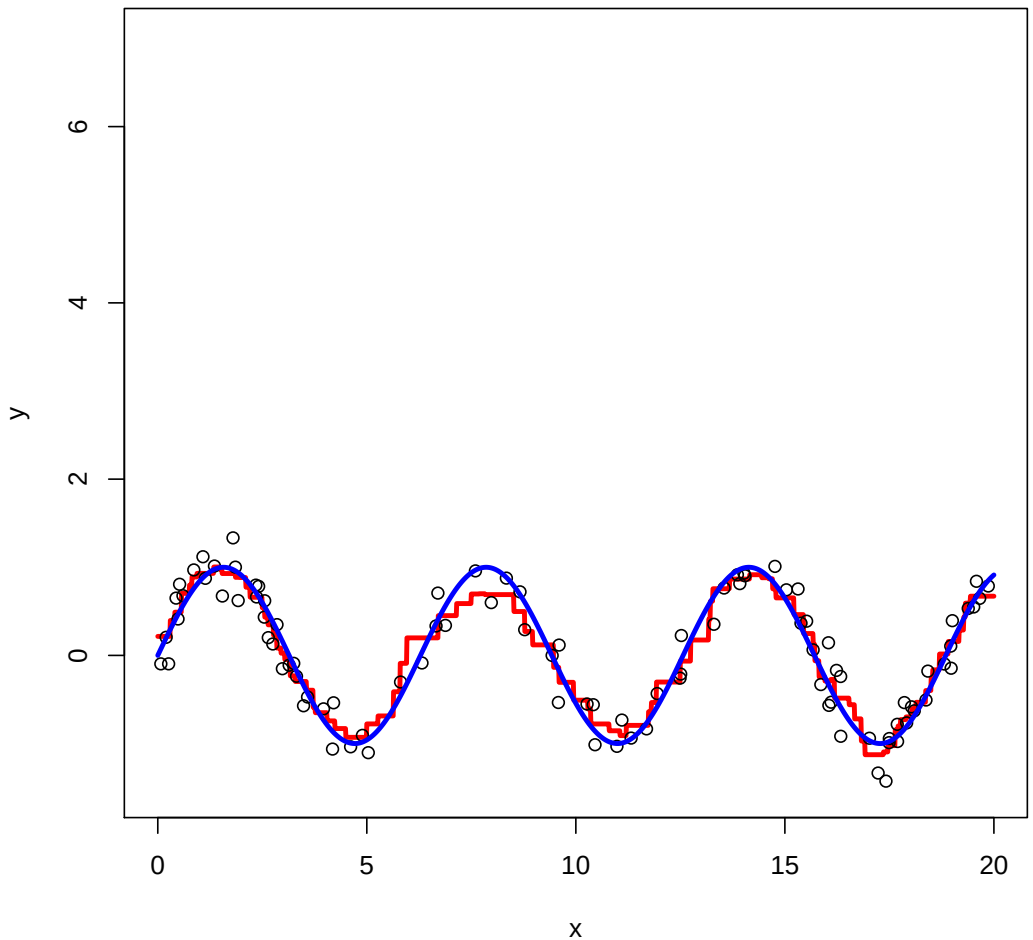
Qualitative Robustness – Non-Parametric Example

Regression:



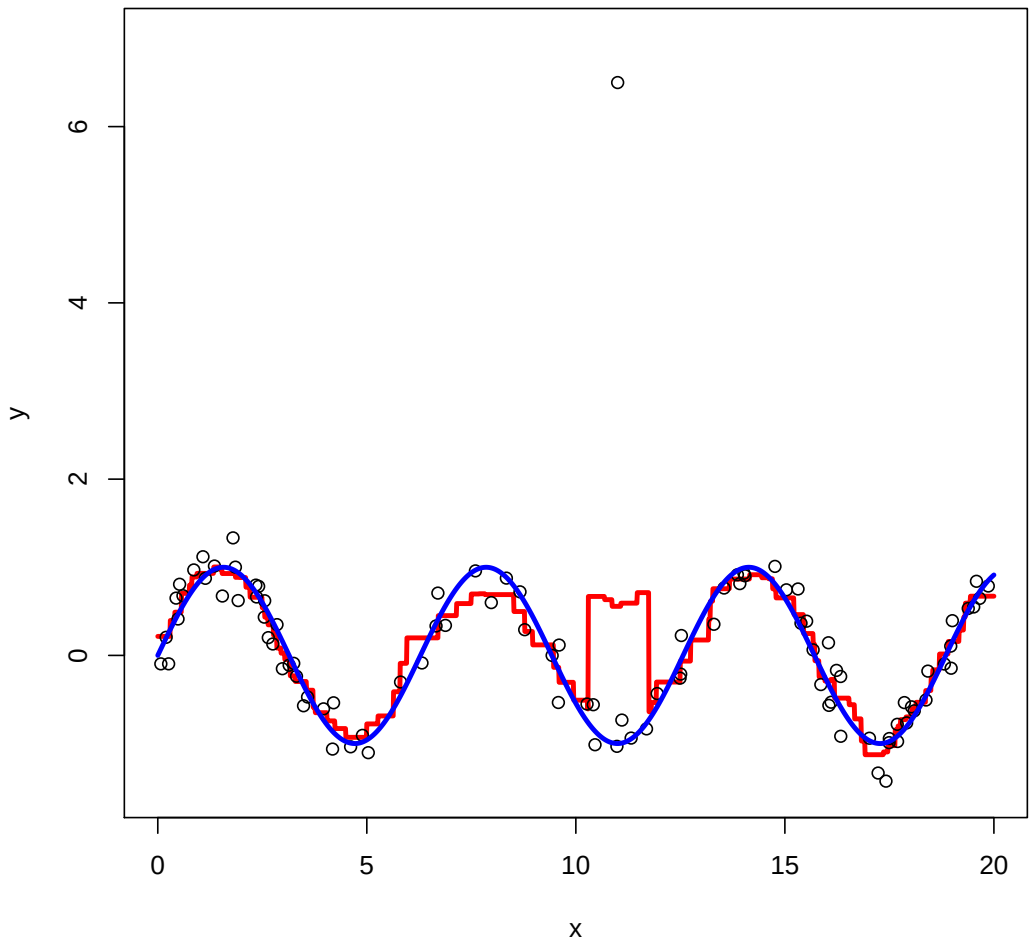
Qualitative Robustness – Non-Parametric Example

Regression: k-nearest neighbor



Qualitative Robustness – Non-Parametric Example

Regression: k-nearest neighbor



Good Estimators

Observations: $Z_1, \ldots, Z_n \sim P_0$ i.i.d.

Statistical functional:

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

Goal: Estimation of $T(P_0)$ (the true P_0 is unknown)

Desirable properties of an estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

are

- Consistency: $S_n \xrightarrow{P_0} T(P_0)$ for $n \to \infty$
- Robustness

Ill-Posed Statistical Problems

$\mathcal{P}$ a set of probability measures
$\mathcal{F}$ a metric space

Dey & Ruymgaart (1999):

- The statistical problem

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

is **well-posed** if T is continuous. That is:

$$\text{if} \quad P_n \overset{w}{\Longrightarrow} P_0 \qquad \text{then} \quad \lim_{n\to} T(P_n) = T(P_0)$$

- The statistical problem is **ill-posed** if T is <u>not</u> continuous.

Ill-Posed Statistical Problems

$\mathcal{P}$ a set of probability measures
$\mathcal{F}$ a metric space

Dey & Ruymgaart (1999):

- The statistical problem

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

is **well-posed** if T is continuous. That is:

$$\text{if} \quad P_n \overset{w}{\Longrightarrow} P_0 \qquad \text{then} \quad \lim_{n\to} T(P_n) = T(P_0)$$

- The statistical problem is **ill-posed** if T is <u>not</u> continuous.

Parametric models : T is usually well-posed
Non-parametric models : T is often ill-posed

Ill-Posed Statistical Problems

$\mathcal{P}$ a set of probability measures
$\mathcal{F}$ a metric space

Theorem: If the statistical problem

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

is ill-posed, then no estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

can simultaneously be consistent and qualitatively robust.

Solution: Idea 1

Use weaker properties:

consistency $\rightsquigarrow$ risk-consistency

robustness $\rightsquigarrow$ risk-robustness

Regression/Classification: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_0$ i.i.d.

Risk of a predictor f : $\mathcal{R}_{P_0}(f) = \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

consistency:

$$S_n \xrightarrow{P_0} T(P_0) \qquad \text{for } n \to \infty$$

robustness:

small errors should not change the estimator too much

Solution: Idea 1

Use weaker properties:

consistency $\rightsquigarrow$ risk-consistency

robustness $\rightsquigarrow$ risk-robustness

Regression/Classification: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_0$ i.i.d.

Risk of a predictor f : $\mathcal{R}_{P_0}(f) = \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

Risk-consistency:

$$\mathcal{R}_{P_0}\big(S_n\big) \xrightarrow{P_0} \mathcal{R}_{P_0}\big(T(P_0)\big) \qquad \text{for } n \to \infty$$

robustness:

small errors should not change the estimator too much

Solution: Idea 1

Use weaker properties:

consistency $\rightsquigarrow$ risk-consistency

robustness $\rightsquigarrow$ risk-robustness

Regression/Classification: $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_0$ i.i.d.

Risk of a predictor f : $\mathcal{R}_{P_0}(f) = \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

Risk-consistency:

$$\mathcal{R}_{P_0}\big(S_n\big) \xrightarrow{P_0} \mathcal{R}_{P_0}\big(T(P_0)\big) \qquad \text{for } n \to \infty$$

Risk-robustness:

small errors should not change the risk of the estimator too much

Ill-Posed Statistical Problems

Theorem: If the statistical problem

$$T : \mathcal{P} \to \mathcal{F}, \qquad P \mapsto T(P)$$

is ill-posed, then no estimator

$$S_n : \mathcal{Z}^n \to \mathcal{F}, \qquad (z_1, \ldots, z_n) \mapsto S_n(z_1, \ldots, z_n)$$

can simultaneously be consistent and qualitatively robust.

Ill-Posed Statistical Problems

Theorem: If the statistical problem

$$T \;:\; \mathcal{P} \;\to\; \mathcal{F}\,, \qquad P \;\mapsto\; T(P)$$

is ill-posed, then no estimator

$$S_n \;:\; \mathcal{Z}^n \;\to\; \mathcal{F}\,, \qquad (z_1, \ldots, z_n) \;\mapsto\; S_n(z_1, \ldots, z_n)$$

can simultaneously be consistent and qualitatively robust.

Theorem (Regression): If the statistical regression problem

$$T \;:\; \mathcal{P} \;\to\; \mathcal{F}\,, \qquad P \;\mapsto\; T(P)$$

is ill-posed, then no estimator

$$S_n \;:\; \big((x_1, y_1), \ldots, (x_n, y_n)\big) \;\mapsto\; S_n\big((x_1, y_1), \ldots, (x_n, y_n)\big)$$

can simultaneously be risk-consistent and qualitatively risk-robust.

Solution: Idea 2

Qualitative Robustness: (Hampel ,1971)

A sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called qualitatively robust if

$$\forall\, P \;\; \forall\, \epsilon > 0 \;\; \exists\, \delta > 0 \;\text{ such that }\; \forall\, Q \;\text{ with }\; d_{\text{Pro}}(Q, P) < \delta$$

$$\sup_{n\in\mathbb{N}} d_{\text{Pro}}\big(S_n(Q^n), S_n(P^n)\big) < \varepsilon$$

Solution: Idea 2

Qualitative Robustness: (Hampel ,1971)

A sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called qualitatively robust if

$$\forall\, P \;\; \forall\, \epsilon > 0 \;\; \exists\, \delta > 0 \;\text{ such that }\; \forall\, Q \;\text{ with }\; d_{\text{Pro}}(Q,P) < \delta$$

$$\sup_{n\in\mathbb{N}} d_{\text{Pro}}\big(S_n(Q^n), S_n(P^n)\big) < \varepsilon$$

Finite Sample Qualitative Robustness:

A sequence of estimators $(S_n)_{n\in\mathbb{N}}$ is called qualitatively robust if

$$\forall\, P \;\; \forall\, \epsilon > 0 \;\; \forall\, n \in \mathbb{N} \;\; \exists\, \delta_n > 0 \;\text{ such that }\; \forall\, Q \;\text{ with }\; d_{\text{Pro}}(Q,P) < \delta_n$$

$$d_{\text{Pro}}\big(S_n(Q^n), S_n(P^n)\big) < \varepsilon$$

Example: Regression/Classification by SVMs

Support vector machine (SVM)

$$S_n\big((x_1,y_1),\ldots,(x_n,y_n)\big) \;=\; \arg\inf_{f\in H} \; \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda_n \|f\|_H^2$$

SVMs can be (depending on $\lambda_n \in (0,\infty)$, $n \in \mathbb{N}$)

either (risk-)consistent **or** qualitatively (risk-)robust

Example: Regression/Classification by SVMs

Support vector machine (SVM)

$$S_n\big((x_1,y_1),\ldots,(x_n,y_n)\big) \;=\; \arg\inf_{f\in H} \; \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda_n \|f\|_H^2$$

SVMs can be (depending on $\lambda_n \in (0,\infty)$, $n \in \mathbb{N}$)

either (risk-)consistent **or** qualitatively (risk-)robust

But: SVMs are always finite sample qualitatively robust
(under some mild assumptions on L and H)

- SVMs can simultaneously be (risk-)consistent and finite sample qualitatively robust.

See Hable & Christmann (2009)

References

- **A.K. Dey and F.H. Ruymgaart (1999)**: Direct density estimation as an ill-posed inverse estimation problem. *Statistica Neerlandica*, 53(3):309–326.
- **R. Hable and A. Christmann (2009)**: Qualitative robustness of support vector machines. *Submitted.* `arXiv:0912.0874v1`
- **F.R. Hampel (1971)**: A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896.

The handout to this talk is also available on my homepage

http://www.staff.uni-bayreuth.de/~btms04/index.html