Robustness versus consistency
in ill-posed statistical problems

Robert Hable
Department of Statistics
LMU Munich

Partially joint work with **Andreas Christmann**

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:

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

Goal: Estimation of $T(P_0)$

Non-Parametric Statistical Problem

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

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

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

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

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

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

$$\mathcal{P} = \Big\{ P \;\Big|\; P \text{ has a } \lambda\text{-density} \Big\}$$

Non-Parametric Regression

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

Regression:

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

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

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

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: $Z_1, \dots, Z_n \sim P_0$ 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

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

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:

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

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

Desirable properties of an estimator

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

are

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

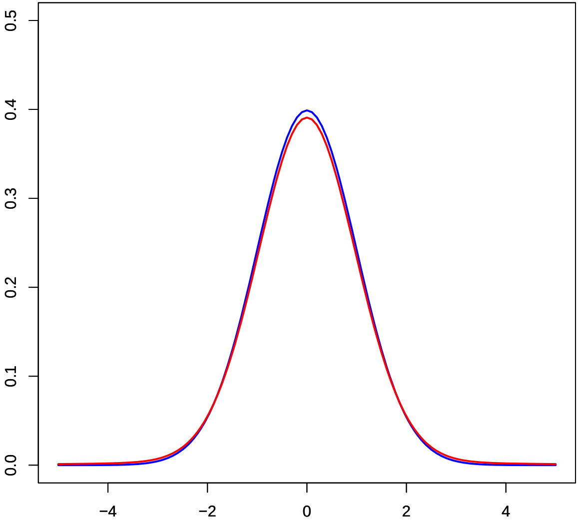
(distribution of the estimator $=$ performance of the estimator)

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

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

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

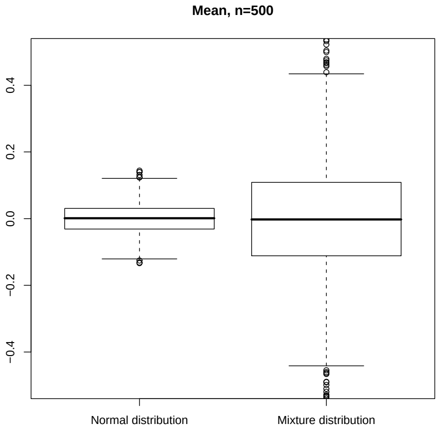
Qualitative Robustness – Parametric Example



Qualitative Robustness – Parametric Example

"mean" applied in 1000 runs

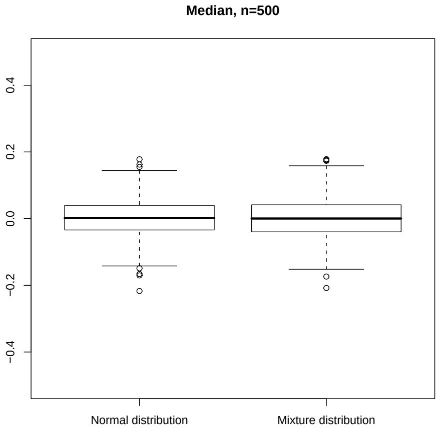
each run consists of a sample with 500 data points



Qualitative Robustness – Parametric Example

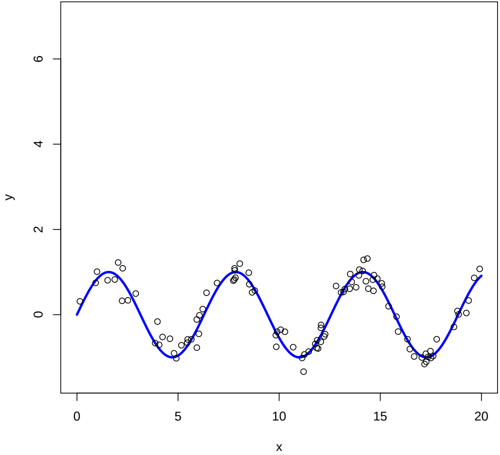
"median" applied in 1000 runs

each run consists of a sample with 500 data points



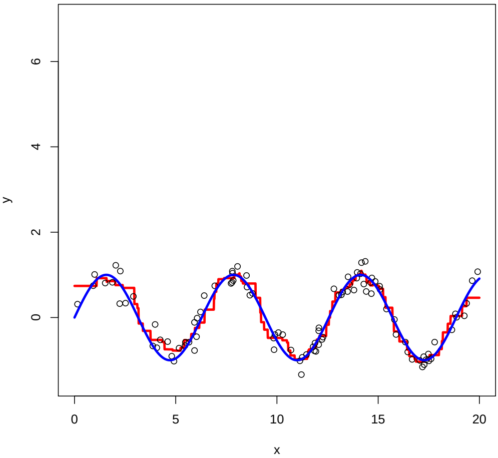
Qualitative Robustness – Non-Parametric Example

Regression:



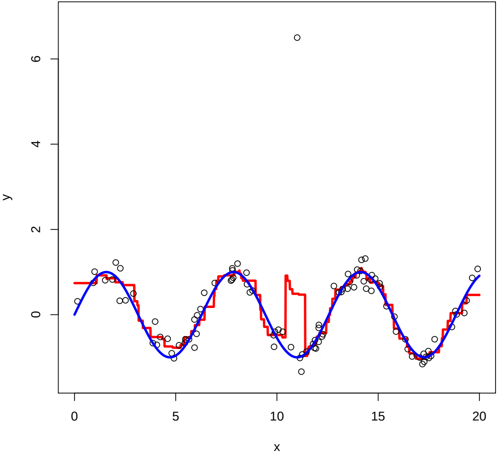
Qualitative Robustness – Non-Parametric Example

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:

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

- 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:

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

- The statistical problem is **ill-posed** if T is not 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

Reformulation of Cueva's generalization of Hampel's theorem:

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.

Example: Density Estimation

$\mathcal{P}$: the set of all probability measures P on $(\mathbb{R}^k, \mathbb{B}^k)$ with Lebesgue-density, denoted by

$$f_P \ : \ \ \mathbb{R}^k \ \rightarrow \ [0, \infty) \ .$$

Example: Density Estimation

$\mathcal{P}$: the set of all probability measures P on $(\mathbb{R}^k, \mathbb{B}^k)$ with Lebesgue-density, denoted by

$$f_P \;:\; \mathbb{R}^k \;\rightarrow\; [0, \infty)\;.$$

Theorem: (Cuevas) The statistical functional

$$T \;:\; \mathcal{P} \;\rightarrow\; L_1(\mathbb{R}^k), \quad P \;\mapsto\; f_P$$

is discontinuous at every $P \in \mathcal{P}$.

Example: Density Estimation

$\mathcal{P}$: the set of all probability measures P on $(\mathbb{R}^k, \mathbb{B}^k)$
with Lebesgue-density, denoted by

$$f_P : \mathbb{R}^k \to [0, \infty) .$$

Theorem: (Cuevas) The statistical functional

$$T : \mathcal{P} \to L_1(\mathbb{R}^k), \quad P \mapsto f_P$$

is discontinuous at every $P \in \mathcal{P}$.

Corollary: Let

$$X_1, \ldots, X_n \sim P \quad \text{i.i.d.}$$

and let S_n, $n \in \mathbb{N}$, be a sequence of density-estimators which is (weakly) consistent for every $P \in \mathcal{P}$. Then, at every $P \in \mathcal{P}$, the estimator S_n, $n \in \mathbb{N}$, is not qualitatively robust.

What can be done: 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 : $\quad \mathcal{R}_{P_0}(f) \;=\; \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

consistency:

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

robustness:

small errors should not change the estimator too much

What can be done: 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 : $\quad \mathcal{R}_{P_0}(f) = \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

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

What can be done: 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 : $\quad \mathcal{R}_{P_0}(f) \;=\; \int L\big(y, f(x)\big)\, P_0\big(d(x,y)\big)$

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.

What can be done: Idea 2

Qualitative Robustness: (Hampel ,1971)

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

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

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

What can be done: Idea 2

Qualitative Robustness: (Hampel ,1971)

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

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

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

Finite Sample Qualitative Robustness:

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

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

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

Example: Nonparametric Regression

For example,

$$Y = f_0(X) + g(X)\varepsilon$$

with

- Y : output variable
- X : input variable
- f_0 : regression function (totally unknown)
- ε : error term
- g : heteroscedasticity (unknown)

Goal: Estimation of the unknown regression function f_0

Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i)+g(X_i)\varepsilon_i\,, \qquad (X_i,Y_i)\;\sim\;P \quad \text{i.i.d.}, \qquad i\in\{1,\dots,n\}$$

Goal: Estimation of $f_0:\mathcal{X}\to\mathcal{Y}\subset\mathbb{R}$

Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i) + g(X_i)\varepsilon_i\,, \qquad (X_i, Y_i) \;\sim\; P \quad \text{i.i.d.}, \qquad i \in \{1, \ldots, n\}$$

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0, \infty)$$

$L(y, t)$: loss caused by estimation $t = \hat{f}_n(x)$ if y is true

Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i)+g(X_i)\varepsilon_i\,, \qquad (X_i,Y_i)\;\sim\;P \quad \text{i.i.d.,} \qquad i\in\{1,\ldots,n\}$$

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y}\times\mathbb{R} \;\to\; [0,\infty)$$

  $L(y,t)$: loss caused by estimation $t = \hat{f}_n(x)$ if y is true

- Risk of an estimate $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\int L\big(y,\hat{f}_n(x)\big)\,P\big(d(x,y)\big)$$

Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i) + g(X_i)\varepsilon_i\,, \qquad (X_i, Y_i) \;\sim\; P \quad \text{i.i.d.,} \qquad i \in \{1, \ldots, n\}$$

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0, \infty)$$

  $L(y,t)$: loss caused by estimation $t = \hat{f}_n(x)$ if y is true
- empirical Risk of an estimate $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, \hat{f}_n(x_i)\big)$$

Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i)+g(X_i)\varepsilon_i\,, \qquad (X_i,Y_i) \;\sim\; P \quad \text{i.i.d.}, \qquad i \in \{1,\dots,n\}$$

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0,\infty)$$

  $L(y,t)$: loss caused by estimation $t = \hat{f}_n(x)$ if y is true
- empirical Risk of an estimate $\hat{f}_n : \mathcal{X} \to \mathbb{R}$

$$\frac{1}{n}\sum_{i=1}^{n} L\big(y_i, \hat{f}_n(x_i)\big)$$

- RKHS H (certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$)

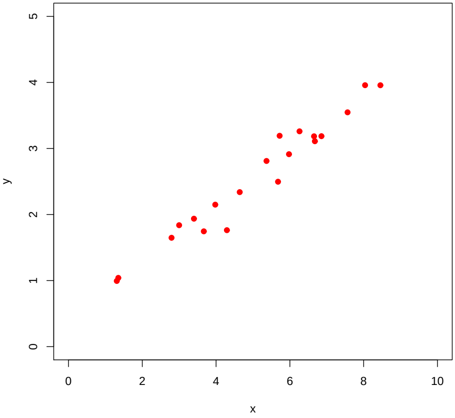
Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i)+g(X_i)\varepsilon_i\,, \qquad (X_i,Y_i)\;\sim\;P \quad \text{i.i.d.}, \qquad i\in\{1,\dots,n\}$$

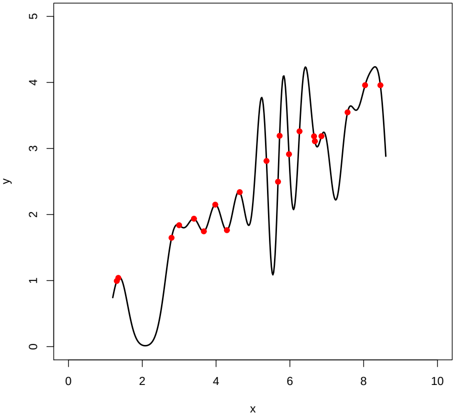
Goal: Estimation of $f_0:\mathcal{X}\to\mathcal{Y}\subset\mathbb{R}$

- Loss function
$$L\;:\;\mathcal{Y}\times\mathbb{R}\;\to\;[0,\infty)$$
$L(y,t)$: loss caused by estimation $t=\hat{f}_n(x)$ if y is true
- empirical Risk of an estimate $\hat{f}_n:\mathcal{X}\to\mathbb{R}$
$$\frac{1}{n}\sum_{i=1}^{n}L\big(y_i,\hat{f}_n(x_i)\big)$$
- RKHS H (certain Hilbert space of functions $f:\mathcal{X}\to\mathbb{R}$)
- Estimator
$$S_n\big((x_1,y_1),\dots,(x_n,y_n)\big)\;=\;\arg\inf_{f\in H}\;\frac{1}{n}\sum_{i=1}^{n}L\big(y_i,f(x_i)\big)$$

Overfitting



Overfitting



Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i)+g(X_i)\varepsilon_i\,, \qquad (X_i,Y_i)\;\sim\;P \quad \text{i.i.d.}, \qquad i\in\{1,\dots,n\}$$

Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function

  $$L \;:\; \mathcal{Y}\times\mathbb{R} \;\to\; [0,\infty)$$

  $L(y,t)$: loss caused by prediction t if y is the true value
- empirical Risk of an estimate $f : \mathcal{X} \to \mathbb{R}$

  $$\frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$

- RKHS H (certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$)
- Estimator

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

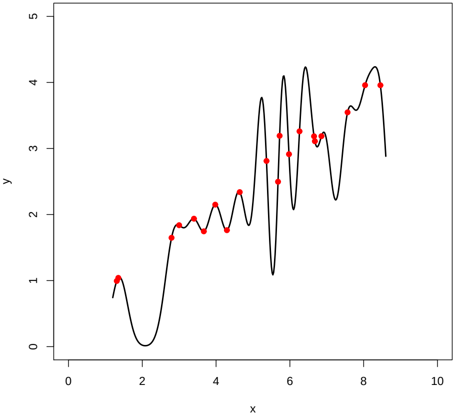
Regularized Kernel Methods

$$Y_i \;=\; f_0(X_i) + g(X_i)\varepsilon_i\,, \qquad (X_i, Y_i) \;\sim\; P \quad \text{i.i.d.}, \qquad i \in \{1, \ldots, n\}$$

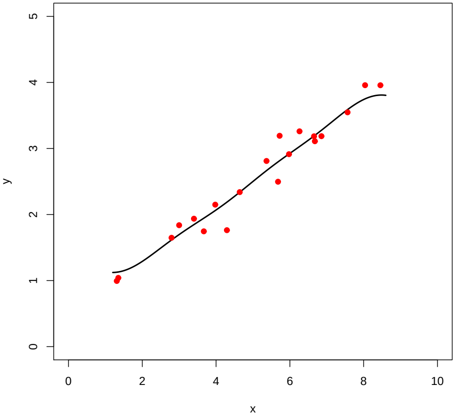
Goal: Estimation of $f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$

- Loss function
$$L \;:\; \mathcal{Y} \times \mathbb{R} \;\to\; [0, \infty)$$
$L(y, t)$: loss caused by prediction t if y is the true value
- empirical Risk of an estimate $f : \mathcal{X} \to \mathbb{R}$
$$\frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big)$$
- RKHS H (certain Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$)
- Regularized kernel methods
$$S_n\big((x_1, y_1), \ldots, (x_n, y_n)\big) \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) + \lambda \|f\|_H^2$$

Overfitting



Overfitting



Reproducing Kernel Hilbert Space (RKHS)

Regularized kernel methods

$$\begin{aligned} S_n \;:\quad (\mathcal{X} \times \mathcal{Y})^n &\quad\to\quad H\,, \\ \big((x_1, y_1), \dots, (x_n, y_n)\big) &\quad\mapsto\quad \arg\inf_{f \in H}\ \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \,+\, \lambda \|f\|_H^2 \end{aligned}$$

with H a reproducing kernel Hilbert space (RKHS)

Reproducing Kernel Hilbert Space (RKHS)

Regularized kernel methods

$$\begin{aligned} S_n \;:\quad (\mathcal{X} \times \mathcal{Y})^n \quad &\to \quad H\,, \\ \big((x_1, y_1), \ldots, (x_n, y_n)\big) \quad &\mapsto \quad \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda \|f\|_H^2 \end{aligned}$$

with H a reproducing kernel Hilbert space (RKHS)

Reproducing kernel Hilbert space H

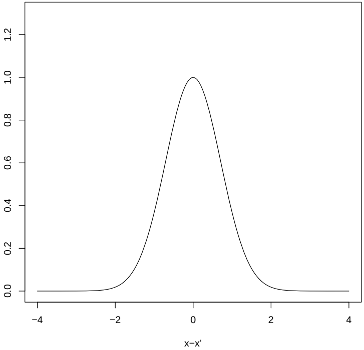
- a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$
- generated by a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- reproducing property

$$\big\langle f\,,\, k(x, \cdot) \big\rangle_H \;=\; f(x) \qquad \forall\, x \in \mathcal{X}, \quad \forall\, f \in H$$

Example: Gaussian Kernel

Gaussian Kernel $\mathcal{X} = \mathbb{R}$

$$k \; : \; \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad (x, x') \mapsto \exp\left(-\frac{1}{\gamma^2}|x - x'|^2\right)$$

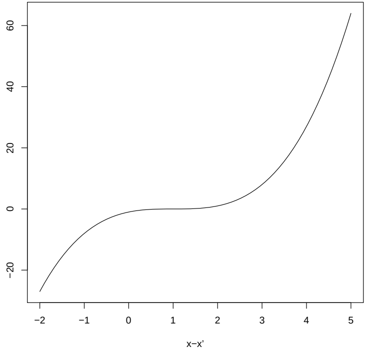


$H \subset L_p(P)$ dense

Example: Polynomial Kernel

Polynomial Kernel $\mathcal{X} = \mathbb{R}$

$$k \;:\; \mathbb{R} \times \mathbb{R} \to \mathbb{R}\,, \qquad (x, x') \mapsto \big(x \cdot x' + c\big)^m$$



$$H = \Big\{ f : \mathbb{R} \to \mathbb{R} \,\Big|\, f \text{ a polynomial with degree } \le m \Big\} \cong \mathbb{R}^{m+1}$$

Representer Theorem

How to calculate the estimator?

$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$

Estimator

$$f_{D_n,\lambda} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \,+\, \lambda \|f\|_H^2$$

Representer Theorem

How to calculate the estimator?

$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$

Estimator

$$f_{D_n,\lambda} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda \|f\|_H^2$$

Representer Theorem

There are $\alpha_{D_n,1}, \ldots, \alpha_{D_n,n} \in \mathbb{R}$ such that

$$f_{D_n,\lambda} \;=\; \sum_{i=1}^{n} \alpha_{D_n,i} k(x_i, \cdot) \; .$$

Representer Theorem

How to calculate the estimator?

$D_n = \big((x_1, y_1), \ldots, (x_n, y_n)\big)$

Estimator

$$f_{D_n,\lambda} \;=\; \arg\inf_{f \in H} \; \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \;+\; \lambda \|f\|_H^2$$

Representer Theorem

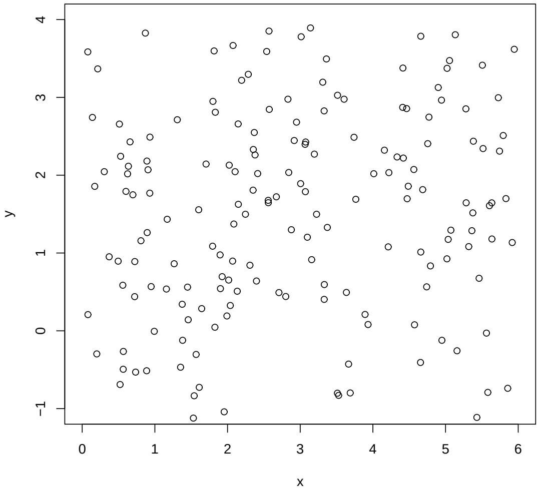
There are $\alpha_{D_n,1}, \ldots, \alpha_{D_n,n} \in \mathbb{R}$ such that

$$f_{D_n,\lambda} \;=\; \sum_{i=1}^{n} \alpha_{D_n,i} k(x_i, \cdot) \; .$$

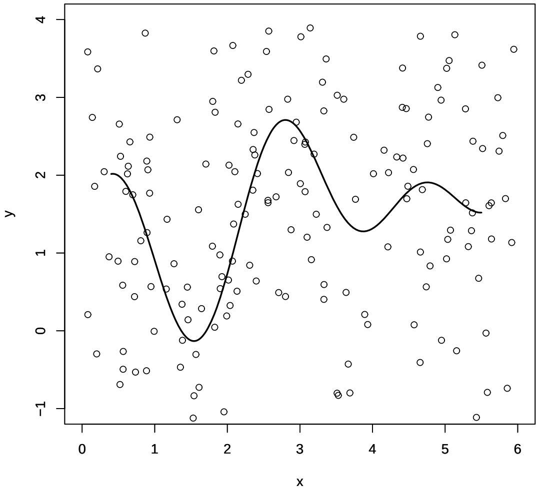
$\longrightarrow$ **just solve a finite convex optimization problem**

. . . and this really works?

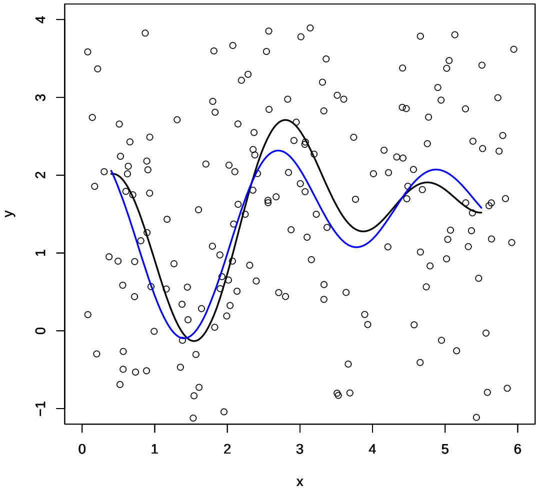
. . . and this really works? Yes, quite good.



. . . and this really works? Yes, quite good.



. . . and this really works? Yes, quite good.



Risk-Consistency

Risk of a predictor $f : \mathcal{X} \to \mathbb{R}$

$$\mathcal{R}_P(f) \;=\; \int L\big(y, f(x)\big)\, P\big(d(x,y)\big) \qquad \hat{=} \quad \text{Quality of } f$$

$$\mathbf{D}_n \;=\; \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

Estimator:

$$f_{\mathbf{D}_n,\lambda_n} \;=\; \arg\inf_{f\in H} \; \frac{1}{n}\sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) \,+\, \lambda_n \|f\|_H^2$$

Risk-Consistency

Risk of a predictor $f : \mathcal{X} \to \mathbb{R}$

$$\mathcal{R}_P(f) \;=\; \int L\big(y, f(x)\big)\, P\big(d(x,y)\big) \qquad \hat{=} \quad \text{Quality of } f$$

$$\mathbf{D}_n \;=\; \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

Estimator:

$$f_{\mathbf{D}_n,\lambda_n} \;=\; \arg\inf_{f \in H} \; \frac{1}{n}\sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) \,+\, \lambda_n \|f\|_H^2$$

Risk-consistency

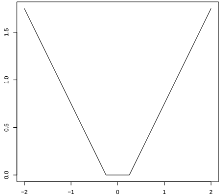
$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \;\xrightarrow[n\to\infty]{}\; \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

essentially if

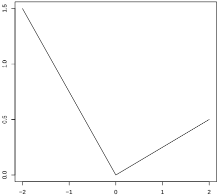
- $H \;\subset\; L_p(P)$ dense (e.g. Gaussian kernel)
- $\lambda_n \to 0$ not too fast (!)

Robustness

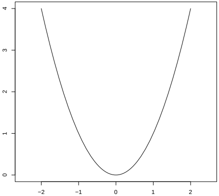
Loss function L



ε-insensitive



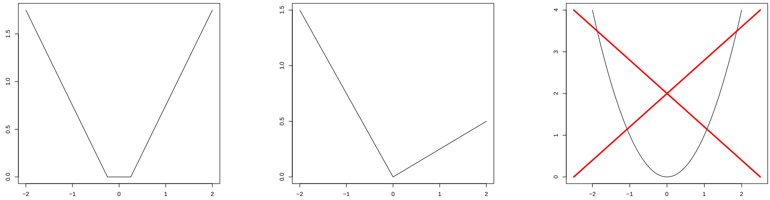
pinball



least squares

Robustness

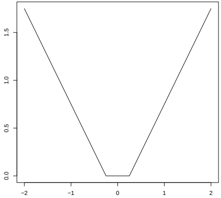
Loss function L should be Lipschitz continuous



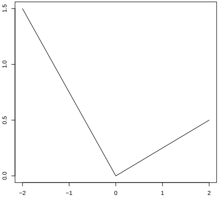
ε-insensitive   pinball   least squares

Robustness

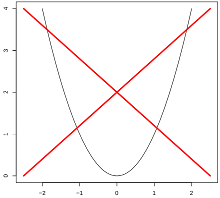
Loss function L should be Lipschitz continuous



ε-insensitive



pinball



least squares

Then: Regularized jernel methods are

- either risk-consistent

$$\text{for } \lambda_n \searrow 0$$

- or qualitatively robust

$$\text{for } \lambda_n \searrow \lambda_0 > 0$$

But: always finite sample qualitatively robust

Hable & Christmann (2011)

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \; = \; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \; = \; \text{min!} \qquad f \in H \, .$$

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems are ill-posed:

- either qualitatively robust or consistent
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems are ill-posed:

- either qualitatively robust or consistent
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)
- statistical inference is impossible

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

for every P?

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

for every P? $\longrightarrow$ **No!** (no-free-lunch theorem)

Rates of Convergence

Risk-consistency

$$\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) \xrightarrow[n\to\infty]{} \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f) \qquad \text{in probability}$$

How fast is this convergence?

Is there a uniform rate r_n such that

$$r_n\Big(\mathcal{R}_P(f_{\mathbf{D}_n,\lambda_n}) - \inf_{f:\mathcal{X}\to\mathbb{R}} \mathcal{R}_P(f)\Big) \xrightarrow[n\to\infty]{} 0 \quad \text{in probability}$$

for every P? $\longrightarrow$ **No!** (no-free-lunch theorem)

Instead,

rates r_n of convergence under assumptions on P

e.g. Steinwart and Scovel (2007), Caponnetto and De Vito (2007), Blanchard et al. (2008), Steinwart et al. (2009), Mendelson and Neeman (2010)

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems are ill-posed:

- either qualitatively robust or consistent
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)
- statistical inference is impossible

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems are ill-posed:

- either qualitatively robust or consistent
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)
- statistical inference is impossible

Idea 3: *Do not try to solve ill-posed problems; pose them well!*

What can be done: Idea 3

Goal: estimate a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

or

$$\inf_{f \in H} \mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,.$$

However, these optimization problems are ill-posed:

- either qualitatively robust or consistent
- there is no uniform rate of convergence to the solution (without substantial assumptions on P)
- statistical inference is impossible

Idea 3: *Do not try to solve ill-posed problems; pose them well!*

So, consider the regularized problem

$$\mathcal{R}_P(f) + \lambda_0 \|f\|_H^2 \;=\; \text{min!} \qquad f \in H\,.$$

Smooth Approximation of the Regression Function

- Instead of estimating a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

we may estimate the solution f_{P,λ_0} of the regularized problem

$$\mathcal{R}_P(f) + \lambda_0 \|f\|_H^2 \;=\; \text{min!} \qquad f \in H\,.$$

f_{P,λ_0} serves as a "smoother approximation" of f^*.

Smooth Approximation of the Regression Function

- Instead of estimating a solution $f^* : \mathcal{X} \to \mathbb{R}$ of

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f : \mathcal{X} \to \mathbb{R}$$

we may estimate the solution f_{P,λ_0} of the regularized problem

$$\mathcal{R}_P(f) + \lambda_0 \|f\|_H^2 \;=\; \text{min!} \qquad f \in H\,.$$

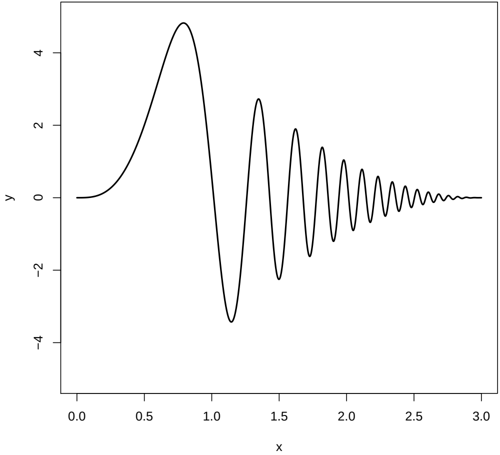
f_{P,λ_0} serves as a "smoother approximation" of f^*.

- The regularized problem is equivalent to

$$\mathcal{R}_P(f) \;=\; \text{min!} \qquad f \in H\,, \quad \|f\|_H \leq r_0\,.$$

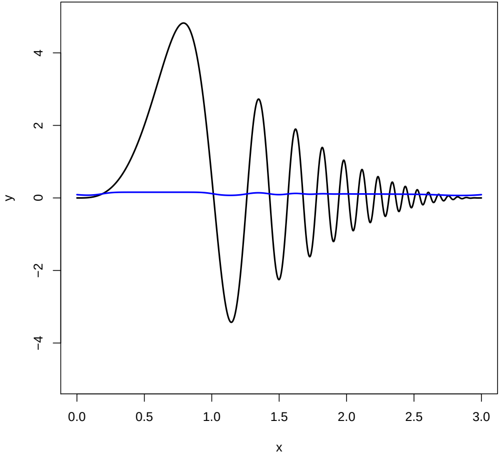
r_0: bound on complexity of "smoother approximation"

Example



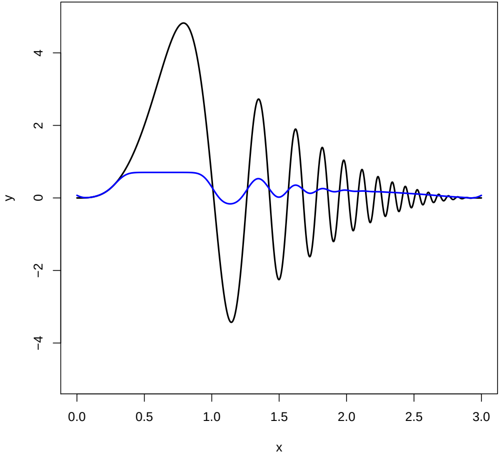
Example

$\lambda = 1$



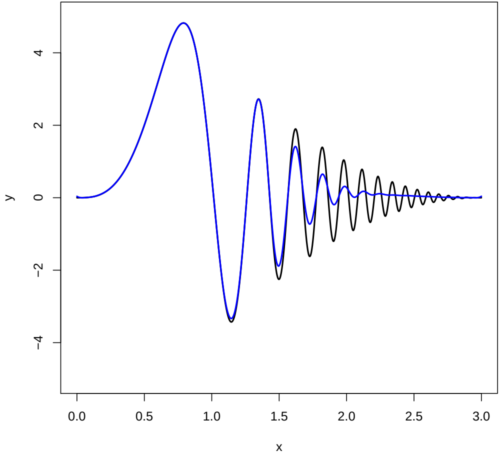
Example

$\lambda = 0.1$



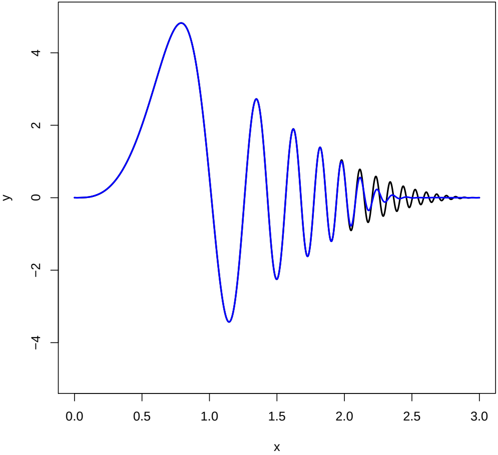
Example

$\lambda = 0.01$



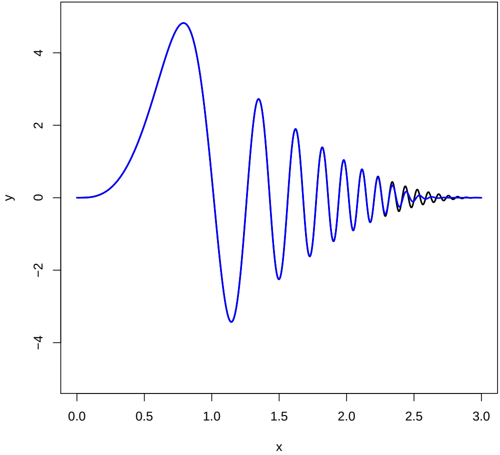
Example

$\lambda = 0.001$



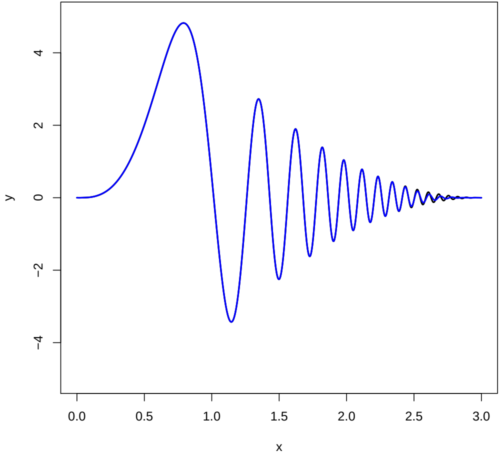
Example

$\lambda = 0.0001$



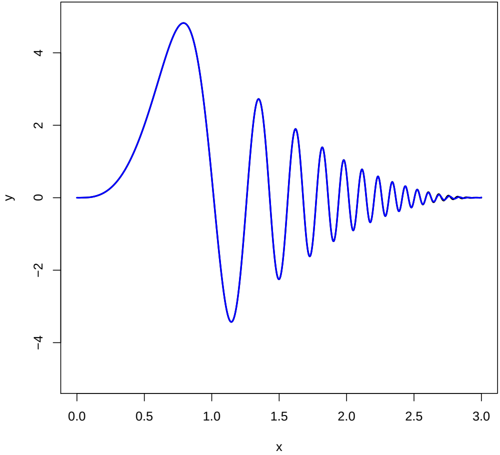
Example

$\lambda = 0.00001$



Example

$\lambda = 0.000001$



Asymptotic Normality of Regularized Problem

Under some

- assumptions on $\mathcal{X}$, L, k ($\leftrightarrow$ H), and $\lambda_{\mathbf{D}_n} \xrightarrow[n\to\infty]{} \lambda_0$
- but (essentially) no assumptions on P,

we have

$$\sqrt{n}\Big(\mathcal{R}(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}}) - \mathcal{R}(f_{P,\lambda_0})\Big) \;\rightsquigarrow\; \mathcal{N}(0,\sigma^2)$$

and, even more,

$$\sqrt{n}\big(f_{\mathbf{D}_n,\lambda_{\mathbf{D}_n}} - f_{P,\lambda_0}\big) \;\rightsquigarrow\; \text{Gaussian process in } H$$

References

- **A. Cuevas (1988)**: Qualitative robustness in abstract inference. *Journal of Statistical Planning and Inference*, 18:277–289.
- **A.K. Dey and F.H. Ruymgaart (1999)**: Direct density estimation as an ill-posed inverse estimation problem. *Statistica Neerlandica*, 53(3): 309–326.
- **Hable, R., Christmann, A. (2011)**: On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102:993-1007, 2011.
- **Hable, R. (2012)**: Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92-117.
- **Hable, R. (2012)**: Asymptotic confidence sets for support vector machine variants and other regularized kernel methods. *Submitted*.
- **F.R. Hampel (1971)**: A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896.