Consistency and Robustness Properties of Support Vector Machines for Heavy-Tailed Distributions

Arnout Van Messem Andreas Christmann

Vrije Universiteit Brussel

UNIVERSITÄT BAYREUTH

PhD Research Day VUB, May 28, 2010

Notation

Assumptions:

- $\mathcal{X} \subseteq \mathbb{R}^d$ closed, $\mathcal{Y} \subseteq \mathbb{R}$ closed, $\mathcal{X} \neq \varnothing$, $\mathcal{Y} \neq \varnothing$
- $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, $1 \leq i \leq n$
- (X_i, Y_i) i.i.d. $\sim \mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$, P (totally) unknown
  $\hookrightarrow \mathrm{P}_X$ on $\mathcal{X}$, $\mathrm{P}(y|x)$ on $\mathcal{Y}$

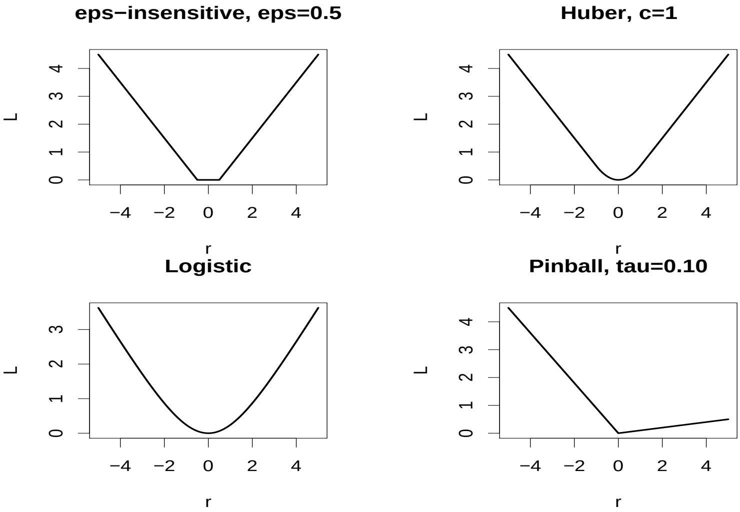
Aim:

- $f(x) =$ quantity of interest
  e.g., conditional median for robust regression

Assumption:

- **Loss function** $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, $L(x, y, f(x))$

Loss functions for regression



Support Vector Machines (SVMs)

Definition

$$f_{L,\mathrm{P},\lambda} := \arg\inf_{f\in\mathcal{H}} \mathbb{E}_{\mathrm{P}} L(X,Y,f(X)) + \lambda \|f\|_{\mathcal{H}}^2$$

- $Y_i|x_i$ depends on an *unknown* function $f:\mathcal{X}\to\mathbb{R}$
- **RKHS** $\mathcal{H} \rightleftarrows$ **kernel** $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, k **measurable**
- $\lambda > 0$ regularization parameter
- $f_{L,\mathrm{D},\lambda} := \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L\big(Y_i, f(X_i)\big) + \lambda \|f\|_{\mathcal{H}}^2$,

  where D is empirical distribution for data set D

Support Vector Machines

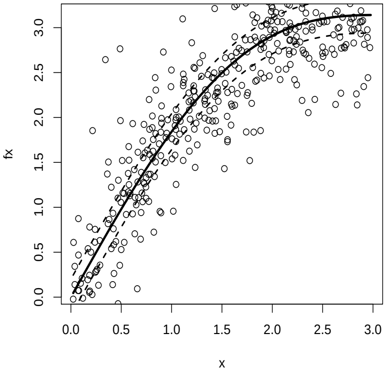
Notions

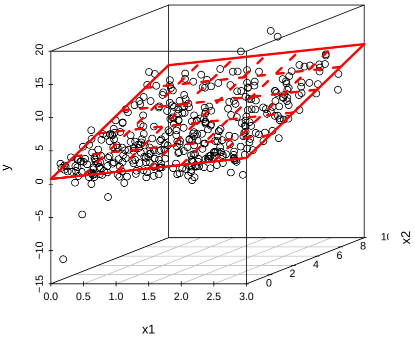
- L is called **convex**, **continuous**, **Lipschitz continuous**, **differentiable**, if L has this property w.r.t. 3^{rd} argument
- k is called **bounded**, if $||k||_\infty := \sqrt{\sup_{x\in\mathcal{X}} k(x,x)} < \infty$

  e.g. **Gaussian RBF:** $k(x,x') = e^{-\gamma||x-x'||_2^2}$, $\gamma > 0$
- $\Phi : \mathcal{X} \to \mathcal{H}$, $\Phi(x) := k(\cdot, x)$, is called **canonical feature map**
- Reproducing property:

$$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}} \quad \forall\, f \in \mathcal{H}, \forall\, x \in \mathcal{X}.$$

Example for feature map $\Phi(\mathbf{x}) = \mathbf{k}(\mathbf{x}, \cdot)$





Risk

Definitions

Risk	$\mathcal{R}_{L,\mathrm{P}}(f)$	$\mathbb{E}_{\mathrm{P}} L(X,Y,f(X))$
Bayes risk	$\mathcal{R}^*_{L,\mathrm{P}}$	$\inf_{f:\mathcal{X}\to\mathbb{R}\text{ measurable}} \mathcal{R}_{L,\mathrm{P}}(f)$
Bayes function	$f^*_{L,\mathrm{P}}$	$\arg\inf_{f:\mathcal{X}\to\mathbb{R}\text{ measurable}} \mathcal{R}_{L,\mathrm{P}}(f)$

Questions

Under which conditions on $\mathcal{X}$, $\mathcal{Y}$, L, $\mathcal{H}$, and k do we have:

1. $f_{L,\mathrm{P},\lambda}$: existence, uniqueness
2. Universal consistency to Bayes risk/function, i.e., $\forall\ \mathrm{P}$

$$\mathcal{R}_{L,\mathrm{P}}(f_{L,\mathrm{D},\lambda}) \xrightarrow{\mathrm{P}} \mathcal{R}^*_{L,\mathrm{P}} \text{ for } |D| = n \to \infty$$

$$f_{L,\mathrm{D},\lambda} \xrightarrow{\mathrm{P}} f^*_{L,\mathrm{P}} \text{ for } |D| = n \to \infty$$

3. Robustness of $f_{L,\mathrm{P},\lambda}$?

Known

Support Vector Machines are **consistent** and **robust**, if based on Lipschitz continuous loss and bounded kernel.

Christmann & Van Messem '08
Steinwart & Christmann '08
Christmann & Steinwart '07

Question

Can the assumptions $f \in L_1(\mathrm{P}_X)$ and $\int |Y| \, d\mathrm{P} < \infty$ be weakened?

(both for regression and classification problems)

$f \in L_1(\mathrm{P}_X)$ if $\int_{\mathcal{X}} |f(x)| \, d\mathrm{P}_X(x) < \infty$

Shifted loss function

Loss function $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ measurable

Definition

$L^\star : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ with

$$L^\star(x, y, t) := L(x, y, t) - L(x, y, 0).$$

Huber, 1967

$L^\star$ can be negative!

Properties

- L (strictly) convex, then $L^\star$ (strictly) convex.
- L Lipschitz continuous, then $L^\star$ Lipschitz continuous.

Shifted loss function

Conditions for finite risk

For L Lipschitz continuous

- $\mathcal{R}_{L,\mathrm{P}}(f) < \infty$ if $f \in L_1(\mathrm{P}_X)$ and $\mathbb{E}_{\mathrm{P}}|Y| < \infty$.
- $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ if $f \in L_1(\mathrm{P}_X)$.

Equality of SVMs

If $f_{L,\mathrm{P},\lambda}$ exists, then $f_{L^\star,\mathrm{P},\lambda} = f_{L,\mathrm{P},\lambda}$.

Existence and Uniqueness of SVM solution

Uniqueness

- L convex and $\mathcal{R}_{L^\star,\mathrm{P}}(f) < \infty$ for some $f \in \mathcal{H}$ and $\mathcal{R}_{L^\star,\mathrm{P}}(f) > -\infty$ for all $f \in \mathcal{H}$
  OR
- L is convex, Lipschitz continuous and $f \in L_1(\mathrm{P}_X)$.

Then, for all $\lambda > 0$, there **exists at most one** SVM $f_{L^\star,\mathrm{P},\lambda}$.

Existence

- L convex, Lipschitz continuous,
- $\mathcal{H}$ RKHS of a bounded measurable kernel k.

Then, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and for all $\lambda > 0$, there **exists** an SVM solution $f_{L^\star,\mathrm{P},\lambda}$.

Consistency

Theorem

- L convex, Lipschitz continuous loss function,
- $\mathcal{H}$ RKHS of a bounded, measurable kernel k,
- (λ_n) sequence of strictly positive numbers with $\lambda_n \to 0$.

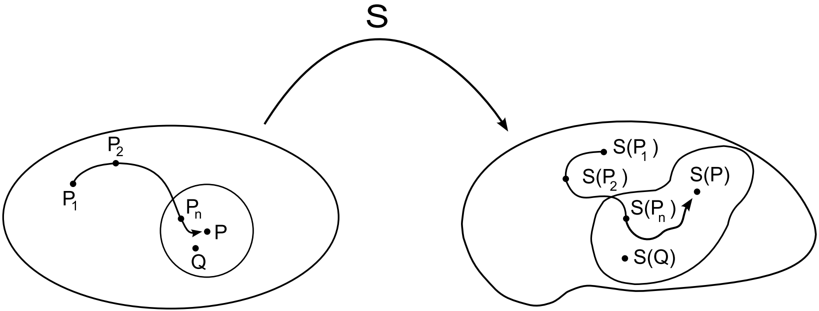
Then, for all $\mathrm{P} \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and all D with $|D| = n$,

1. if $\lambda_n^2 n \to \infty$, then $\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \xrightarrow{\mathrm{P}} \mathcal{R}^*_{L^\star,\mathrm{P}}$.
2. if $\lambda_n^{2+\delta} n \to \infty$ for some $\delta \in (0,\infty)$, then
$$\mathcal{R}_{L^\star,\mathrm{P}}(f_{L^\star,\mathrm{D},\lambda_n}) \xrightarrow{\mathrm{a.s.}} \mathcal{R}^*_{L^\star,\mathrm{P}} .$$
3. if $L = L_\tau$ pinball loss: $d(f_{L^\star,\mathrm{D},\lambda_n}, f^*_{L_\tau,\mathrm{P}}) \to 0.$

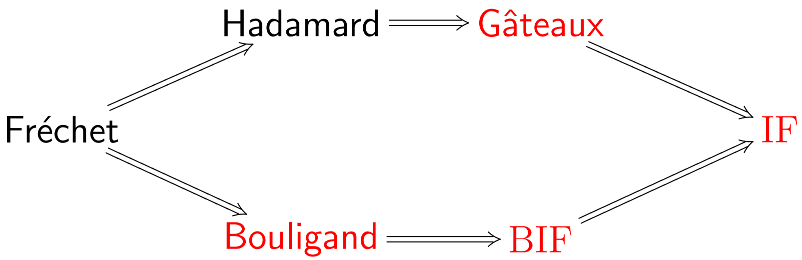
d is a metric describing convergence in probability.

Robustness

1. What if (X_i, Y_i) i.i.d. $\sim \mathrm{P}$, $\mathrm{P} \in \mathcal{M}_1$ unknown is invalid?
2. What is the impact on $S : \mathrm{P} \mapsto f_{L^\star,\mathrm{P},\lambda}$?



Derivatives and Influence Functions



Christmann & Van Messem (2008)

Notation: ∇^F, ∇^G, ∇^B, ∇^B_3, etc.

Property: $\nabla^F_3 L^\star = \nabla^F_3 L$, $\nabla^B_3 L^\star = \nabla^B_3 L$

Bouligand differentiability

Bouligand-derivative

$f : U \to Z$ is **Bouligand-differentiable** at $x_0 \in U$, if $\exists$ a positive homogeneous function $\nabla^B f(x_0) : U \to Z$ such that

$$f(x_0 + h) = f(x_0) + \nabla^B f(x_0)(h) + o(h)\,,$$

i.e.

$$\lim_{h \downarrow 0} \frac{\left\| f(x_0 + h) - f(x_0) - \nabla^B f(x_0)(h) \right\|_Z}{\|h\|_U} = 0.$$

$g : E \to F$ positive homogeneous if

$$g(\alpha x) = \alpha g(x) \qquad \forall\, \alpha \geq 0\,,\, \forall x \in E$$

Influence Function

Definition (Hampel, '68, Hampel et al. '86)

The **influence function** (IF) of a function $S : \mathcal{M}_1 \to \mathcal{H}$ for a distribution P is given by

$$\mathrm{IF}(z; S, \mathrm{P}) := \lim_{\varepsilon \downarrow 0} \frac{S\big((1-\varepsilon)\mathrm{P} + \varepsilon\delta_z\big) - S(\mathrm{P})}{\varepsilon},$$

in those $z := (x, y) \in \mathcal{X} \times \mathcal{Y}$ where this limit exists.

If $\nabla^G(z; S, \mathrm{P})$ exists: $\nabla^G = \mathrm{IF}$ and IF is linear and continuous

Goal: Bounded IF

Problem: Loss function L often not Fréchet-differentiable

Bouligand Influence Function

Definition (C&VM '08)

The **Bouligand influence function** (BIF) of a function $S : \mathcal{M}_1 \to \mathcal{H}$ for a distribution P in the direction of a distribution $\mathrm{Q} \neq \mathrm{P}$ is the special Bouligand-derivative

$$\lim_{\varepsilon \downarrow 0} \frac{\left\| S\big((1-\varepsilon)\mathrm{P} + \varepsilon \mathrm{Q}\big) - S(\mathrm{P}) - \mathrm{BIF}(\mathrm{Q}; S, \mathrm{P}) \right\|_{\mathcal{H}}}{\varepsilon} = 0$$

(if it exists).

If BIF exists and $\mathrm{Q} = \delta_z$: IF exists and $\mathrm{BIF} = \mathrm{IF}$

Goal: Bounded BIF

Result for BIF

Assumptions

- $\mathcal{H}$ is RKHS with **bounded**, continuous kernel k
- L **convex** and **Lipschitz continuous** with $|L|_1 \in (0,\infty)$
- $\nabla_3^B L(x,y,\cdot)$ and $\nabla_{3,3}^B L(x,y,\cdot)$ measurable with $\kappa_1 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_3^B L(x,y,\cdot)\right\|_\infty \in (0,\infty)$, $\kappa_2 := \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\|\nabla_{3,3}^B L(x,y,\cdot)\right\|_\infty < \infty$
- $\lambda > \frac{1}{2}\kappa_2 \|k\|_\infty^{\ 3}$ ($\kappa_2 = 0$ for eps-insensitive and pinball)
- $\mathrm{P} \neq \mathrm{Q}$, probability measures on $\mathcal{X} \times \mathcal{Y}$

Theorem BIF

Then $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ with $S(\mathrm{P}) := f_{L^\star,\mathrm{P},\lambda}$ and $\mathrm{Q} \neq \mathrm{P} \in \mathcal{M}_1$

1. exists,
2. equals

$$T^{-1}\Big(\mathbb{E}_{\mathrm{P}} \nabla_3^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X) - \mathbb{E}_{\mathrm{Q}} \nabla_3^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X))\Phi(X)\Big),$$

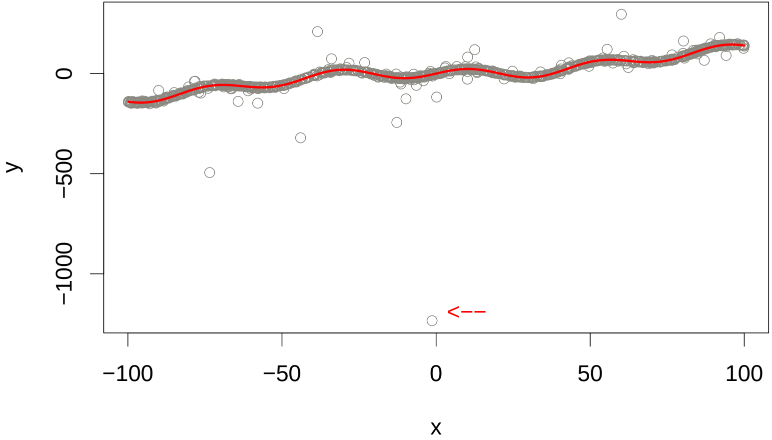
   where $T : \mathcal{H} \to \mathcal{H}$ with $T(\cdot) := 2\lambda\,\mathrm{id}_{\mathcal{H}}(\cdot) + \mathbb{E}_{\mathrm{P}} \nabla_{3,3}^B L^\star(X, Y, f_{L^\star,\mathrm{P},\lambda}(X))\langle \Phi(X), \cdot \rangle_{\mathcal{H}} \Phi(X)$,
3. is bounded.

Simulated data

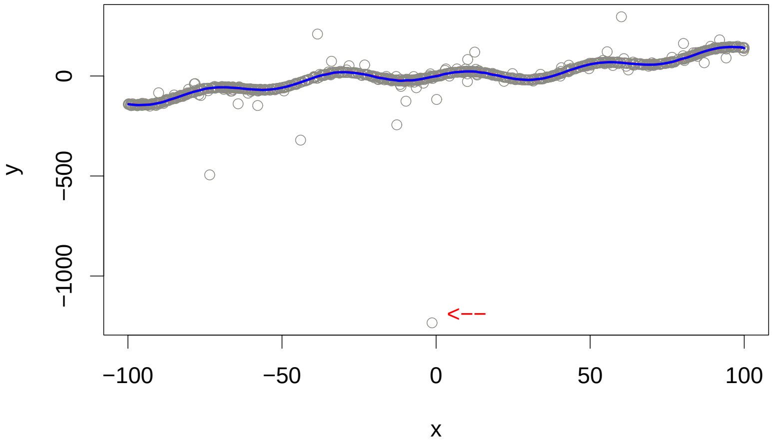
- Predict $f(x) = 50\sin(x/20)\cos(x/10) + x$
- $n = 1000$ data points $x_i \sim \mathcal{U}(-100, 100)$
- Output $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim$ Cauchy distribution
- ϵ-insensitive loss and Gaussian RBF kernel
- hyperparameters $(\lambda, \epsilon, \gamma)$ determined by minimizing $L^\star$-risk via grid search over $17 \times 12 \times 17 = 3468$ knots
  - λ regularization parameter of SVM
  - ϵ parameter of ϵ-insensitive loss
  - γ parameter of Gaussian RBF kernel

  Result $(\lambda, \epsilon, \gamma) = \left(2^{-12}, 2^{-8}, 2^{-4}\right)$

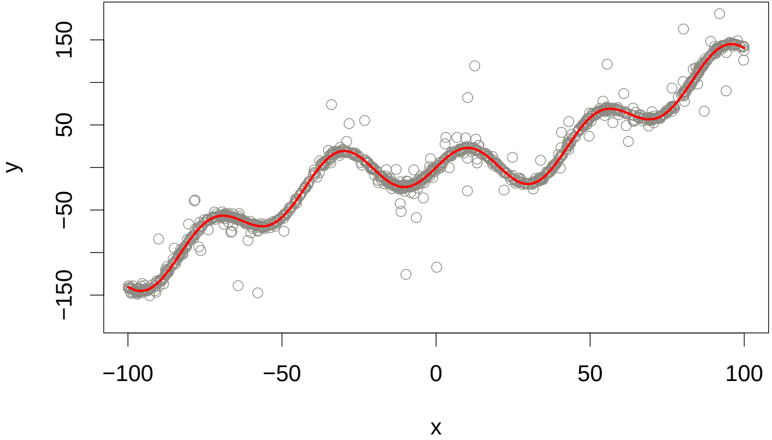
Simulated data



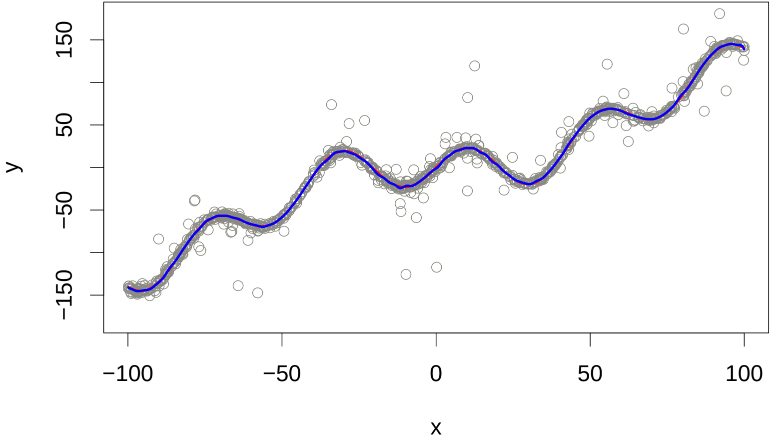
Simulated data



Simulated data



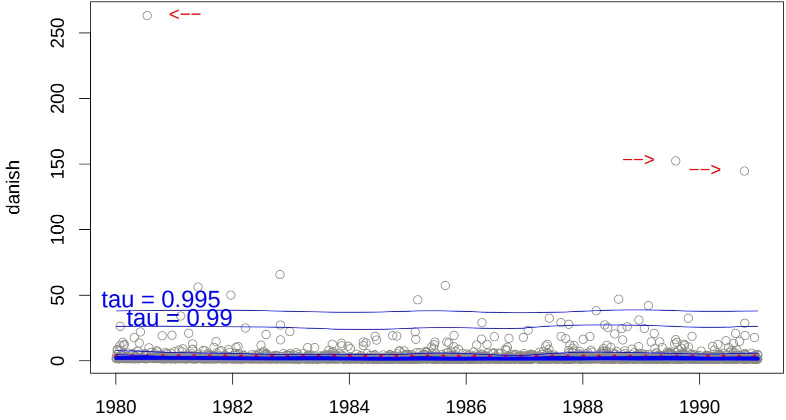
Simulated data



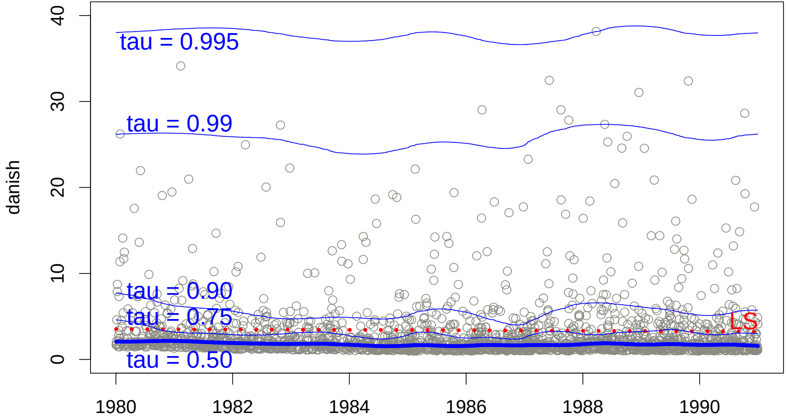
Danish data

- 2167 fire insurance claims over 1 million DKK (1980 – 1990)
- Regression with time as explanatory variable
    - Classical least squares regression
    - Conditional quantile regression using SVMs
        - Pinball loss for $\tau \in \{0.50, 0.75, 0.90, 0.99, 0.995\}$
        - Gaussian RBF kernel
- Extreme value distribution

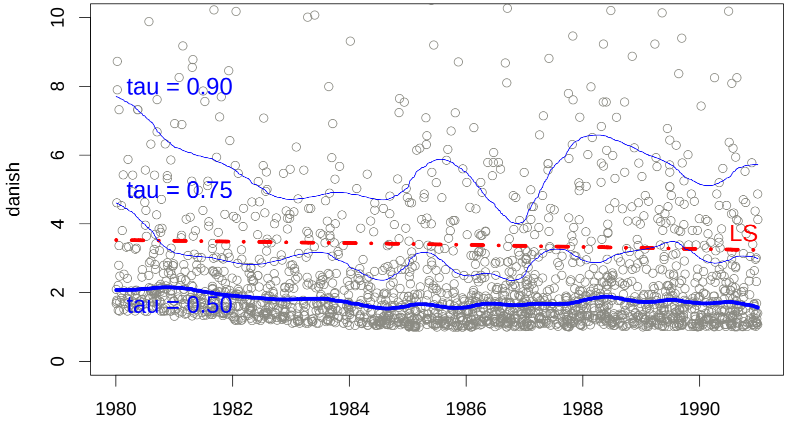
Danish data


danish
0
50
100
150
200
250
1980
1982
1984
1986
1988
1990
tau = 0.995
tau = 0.99
<--
-->
-->

Danish data


danish
0
10
20
30
40
1980
1982
1984
1986
1988
1990
tau = 0.995
tau = 0.99
tau = 0.90
tau = 0.75
tau = 0.50
LS

Danish data



tau = 0.90
tau = 0.75
tau = 0.50
LS
danish
0
2
4
6
8
10
1980
1982
1984
1986
1988
1990

Conclusions

SVMs based on $L^\star(x,y,t) := L(x,y,t) - L(x,y,0)$

1. Weaker assumption on P: only $f \in L_1(\mathrm{P}_X)$ is needed
   e.g. f bounded and $\mathcal{X} \subset \mathbb{R}^d$ bounded
2. Existence and uniqueness of $f_{L^\star,\mathrm{P},\lambda}$
3. Consistency of risk and SVM solution
4. Robustness
   - Existence of BIF
   - $\mathrm{BIF}(\mathrm{Q}; S, \mathrm{P})$ bounded if $\nabla_3^B L$, $\nabla_{3,3}^B L$ measurable and bounded as well as k continuous and bounded

References

- Van Messem & Christmann (2010). *Advances in Data Analysis and Classification*, accepted.
- Christmann, Van Messem & Steinwart (2009). *Statistics and Its Interface*, **2**, 311-327.
- Christmann & Van Messem (2008). *Journal of Machine Learning Research*, **9**, 915-936.
- Steinwart & Christmann (2008). *Support Vector Machines.* Springer, New York.
- Christmann & Steinwart (2007). *Bernoulli*, **13**, 799-819.
- Hampel (1974). *J. Amer. Statist. Assoc.*, **69**, 383-393.
- Huber (1967). *Proceedings of the* 5^{th} *Berkeley Symposium*.
- Koenker (2005). *Quantile regression.* Cambridge University Press.
- Schölkopf & Smola (2002). *Learning with kernels.* MIT Press.
- Vapnik (1998). *Statistical learning theory.* Wiley.

Reason

Conditions for finite risk

For L Lipschitz continuous

- $\mathbb{E}_{\mathrm{P}} L(X, Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$ and $Y \in L_1(\mathrm{P}_{Y|x})$.

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq |L|_1 \Big(\int_{\mathcal{X}} |f(x)| d\mathrm{P}_X(x) + \int_{\mathcal{X}} \int_{\mathcal{Y}} |y| d\mathrm{P}(y|x)\, d\mathrm{P}_X(x) \Big)$$

- $\mathbb{E}_{\mathrm{P}} L^{\star}(X, Y, f(X)) < \infty$ if $f \in L_1(\mathrm{P}_X)$.

$$\mathcal{R}_{L,\mathrm{P}}(f) \leq |L|_1 \int_{\mathcal{X}} |f(x)|\, d\mathrm{P}_X(x)$$