On Support Vector Machines to Estimate Scale Functions

Robert Hable & Andreas Christmann

Robert Hable
Department of Mathematics
University of Bayreuth
Nonparametric Regression

\[ Y = f_0(X) + \varepsilon \]
Nonparametric Regression

\[ Y = f_0(X) + \varepsilon \]

Example:

LIDAR = Light Detection And Ranging (data set with \( n = 221 \))
Nonparametric Regression

\[ Y = f_0(X) + \varepsilon \]

Example:

LIDAR = Light Detection And Ranging (data set with \( n = 221 \))
Nonparametric Regression with Heteroscedastic Errors

\[ Y = f_0(X) + g_0(X) \varepsilon \]

Example:

LIDAR = Light Detection And Ranging (data set with \( n = 221 \))
Scale functions

Heteroscedastic model:

\[ Y = f_0(X) + g_0(X) \varepsilon \]

**Note** that

- heteroscedasticity occurs in practice
- regularized kernel methods (SVM) estimate location \( f_0 \)
  - in a non-parametric way
  - also in case of heteroscedasticity

**Question:**
How to estimate the scale function \( g_0 \)?
Estimation of Scale

$Y, Y_1, \ldots, Y_n \sim \text{i.i.d. } P_Y$

We want to estimate: $\text{Scale}(Y)$

- **Variance:** $\text{mean}((Y - \text{mean}(Y))^2)$
  - estimate $\text{mean}(Y)$ by $\hat{\mu}_n$ and put $\tilde{Y}_i := (Y_i - \hat{\mu}_n)^2$
  - estimate $\text{mean}(\tilde{Y}_i)$
Estimation of Scale

\( Y, Y_1, \ldots, Y_n \sim \text{i.i.d.} \quad P_Y \)

We want to estimate: \( \text{Scale}(Y) \)

- **Variance:** \( \text{mean}\left( (Y - \text{mean}(Y))^2 \right) \)
  - estimate \( \text{mean}(Y) \) by \( \hat{\mu}_n \) and put \( \tilde{Y}_i := (Y_i - \hat{\mu}_n)^2 \)
  - estimate \( \text{mean}(\tilde{Y}_i) \)

- **MAD:** \( \text{median}\left( |Y - \text{median}(Y)| \right) \)
  - estimate \( \text{median}(Y) \) by \( \hat{\mu}_n \) and put \( \tilde{Y}_i := |Y_i - \hat{\mu}_n| \)
  - estimate \( \text{median}(\tilde{Y}_i) \)
Estimation of Scale

\( Y, Y_1, \ldots, Y_n \sim \text{i.i.d. } P_Y \)

We want to estimate: **Scale**(\( Y \))

- **Variance:** \( \text{mean}\left((Y - \text{mean}(Y))^2\right) \)
  - estimate mean(\( Y \)) by \( \hat{\mu}_n \) and put \( \tilde{Y}_i := (Y_i - \hat{\mu}_n)^2 \)
  - estimate mean(\( \tilde{Y}_i \))

- **MAD:** \( \text{median}(|Y - \text{median}(Y)|) \)
  - estimate median(\( Y \)) by \( \hat{\mu}_n \) and put \( \tilde{Y}_i := |Y_i - \hat{\mu}_n| \)
  - estimate median(\( \tilde{Y}_i \))

- **IQR:** \( \text{upper.quartile}(Y) - \text{lower.quartile}(Y) \)
  - estimate upper and lower quartiles by \( \hat{Q}_3 \) and \( \hat{Q}_1 \)
  - calculate \( \hat{Q}_3 - \hat{Q}_1 \)
Estimation of Scale Functions

\((X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \sim \text{i.i.d. } P\)

We want to estimate: \(\text{Scale}(Y|X=\cdot)\)

- **Variance function:** \(\text{mean}((Y - \text{mean}(Y|X=\cdot))^2|X=\cdot)\)
  - estimate \(\text{mean}(Y|X=\cdot)\) by \(\hat{f}_n\) and put \(\tilde{Y}_i := (Y_i - \hat{f}_n(X_i))^2\)
  - estimate \(\text{mean}(\tilde{Y}_i|X_i=\cdot)\)

- **MAD function:** \(\text{median}(|Y - \text{median}(Y|X=\cdot)||X=\cdot)\)
  - estimate \(\text{median}(Y|X=\cdot)\) by \(\hat{f}_n\) and put \(\tilde{Y}_i := |Y_i - \hat{f}_n(X_i)|\)
  - estimate \(\text{median}(\tilde{Y}_i|X_i=\cdot)\)

- **IQR function:** \(\text{upper.quartile}(Y|X=\cdot) - \text{lower.quartile}(Y|X=\cdot)\)
  - estimate upper and lower quartile functions by \(\hat{f}_3\) and \(\hat{f}_1\)
  - calculate \(\hat{f}_3 - \hat{f}_1\)
Estimation of Scale Functions

**Summing up:**

- Estimation of **scale** can be done by estimating **mean**, **median**, or **quartiles**
- Estimation of **scale functions** can be done by estimating conditional **mean**, **median**, or **quartile** functions
Estimation of Scale Functions

Summing up:

➤ Estimation of scale can be done by estimating mean, median, or quartiles

➤ Estimation of scale functions can be done by estimating conditional mean, median, or quartile functions

and

➤ Regularized kernel-methods (SVMs) estimate conditional mean, median, or quartile functions

Hence: Regularized kernel-methods should be able to estimate scale functions
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.,} \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R} \)
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.,} \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : X \rightarrow Y \subset \mathbb{R} \)

- Loss function
  \[ L : Y \times \mathbb{R} \rightarrow [0, \infty) \]
  \[ L(y, t) : \text{loss caused by estimation } t = \hat{f}_n(x) \text{ if } y \text{ is true} \]
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i , \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R} \)

- **Loss function**
  \[
  L : \mathcal{Y} \times \mathbb{R} \to [0, \infty) \\
  L(y, t) : \text{loss caused by estimation } t = \hat{f}_n(x) \text{ if } y \text{ is true}
  \]

- **Risk of an estimate** \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[
  \int L(y, \hat{f}_n(x)) \, P(d(x, y))
  \]
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i)\varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.,} \quad i \in \{1, \ldots, n\} \]

**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** \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)

  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i)) \]
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \text{ i.i.d., } i \in \{1, \ldots, n\} \]

**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** \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i)) \]

- **RKHS** \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \to \mathbb{R} \))
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**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** \( \hat{f}_n : \mathcal{X} \to \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_n(x_i)) \]

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

- **Support vector machine**
  \[ S_n((x_1, y_1), \ldots, (x_n, y_n)) = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]
Overfitting
Overfitting
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \quad \text{i.i.d.}, \quad i \in \{1, \ldots, n\} \]

**Goal:** Estimation of \( f_0 : \mathcal{X} \rightarrow \mathcal{Y} \subset \mathbb{R} \)

- Loss function
  \[ L : \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty) \]
  \( L(y, t) \): loss caused by prediction \( t \) if \( y \) is the true value
- Empirical risk of an estimate \( f : \mathcal{X} \rightarrow \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]
- RKHS \( H \) (certain Hilbert space of functions \( f : \mathcal{X} \rightarrow \mathbb{R} \))
- Support vector machine
  \[ S_n((x_1, y_1), \ldots, (x_n, y_n)) = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]
Regularized Kernel-Methods (SVMs)

\[ Y_i = f_0(X_i) + g_0(X_i) \varepsilon_i, \quad (X_i, Y_i) \sim P \text{ i.i.d.,} \quad i \in \{1, \ldots, n\} \]

**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 \( f : \mathcal{X} \to \mathbb{R} \)
  \[ \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) \]

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

- Support vector machine
  \[ S_n((x_1, y_1), \ldots, (x_n, y_n)) = \arg \inf_{f \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \| f \|_H^2 \]
Overfitting
Overfitting
Estimation of Scale Functions

Summing up:

▶ Estimation of scale can be done by estimating mean, median, or quartiles

▶ Estimation of scale functions can be done by estimating conditional mean, median, or quartile functions

and

▶ Regularized kernel-methods (SVMs) estimate conditional mean, median, or quartile functions

Hence: Regularized kernel-methods should be able to estimate scale functions
Estimation of the Conditional Mean Function

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

**Estimation of the conditional mean function**

\[ x \mapsto \text{mean}(Y|X = x) \]
Estimation of the Conditional Mean Function

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

**Estimation of the conditional mean function**

\[ x \mapsto \text{mean}(Y|X = x) \]

by use of SVM with least-squares loss: \( L(y, t) = (y - t)^2 \)
Estimation of the Conditional Median Function

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

Estimation of the **conditional median function**

\[ x \mapsto \text{median}(Y \mid X = x) \]
Estimation of the Conditional Median Function

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

**Estimation of the conditional median function**

\[ x \mapsto \text{median}(Y|X = x) \]

by use of **SVM** with **absolute distance**: \[ L(y, t) = |y - t| \]
Estimation of Conditional Quantile Functions

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

Estimation of the conditional \( \tau \)-quantile function

\[ x \mapsto \text{quantile}_\tau(Y|X = x) \]
Estimation of Conditional Quantile Functions

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

Estimation of the conditional \( \tau \)-quantile function

\[ x \mapsto \text{quantile}_\tau(Y|X = x) \]

by use of SVM with pinball loss: \( L(y, t) = \begin{cases} (1 - \tau)(t - y) & \text{if } y < t \\ \tau(y - t) & \text{if } y \geq t \end{cases} \)
Estimation of Scale Functions

\((X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \sim \text{i.i.d. } P\)

We want to estimate: \(\text{Scale}(Y|X=\cdot)\)

- **Variance function:** \(\text{mean}((Y - \text{mean}(Y|X=\cdot))^2|X=\cdot)\)
  - estimate \(\text{mean}(Y|X=\cdot)\) by \(\hat{f}_n\) and put \(\tilde{Y}_i := (Y_i - \hat{f}_n(X_i))^2\)
  - estimate \(\text{mean}(\tilde{Y}_i|X_i=\cdot)\)

- **MAD function:** \(\text{median}(|Y - \text{median}(Y|X=\cdot)||X=\cdot)\)
  - estimate \(\text{median}(Y|X=\cdot)\) by \(\hat{f}_n\) and put \(\tilde{Y}_i := |Y_i - \hat{f}_n(X_i)|\)
  - estimate \(\text{median}(\tilde{Y}_i|X_i=\cdot)\)

- **IQR function:** upper.quartile\((Y|X=\cdot)\) \(-\) lower.quartile\((Y|X=\cdot)\)
  - estimate upper and lower quartile functions by \(\hat{f}_3\) and \(\hat{f}_1\)
  - calculate \(\hat{f}_3 - \hat{f}_1\)
MAD function

MAD function: \( g_0 = \text{median}(y - \text{median}(y|x = \cdot)|x = \cdot) \)
MAD function

**MAD function:** $g_0 = \text{median}( | Y - \text{median}( Y | X = \cdot ) | | X = \cdot )$

- estimate $\text{median}( Y | X = \cdot )$ by SVM $f_{\text{SVM}, n}$ with absolute deviation loss
**MAD function**

**MAD function:** \( g_0 = \text{median}(|Y - \text{median}(Y|X = \cdot)| | X = \cdot) \)

- estimate \( \text{median}(Y|X = \cdot) \) by SVM \( f_{\text{SVM},n} \) with absolute deviation loss
- put \( \tilde{Y}_i := |Y_i - f_{\text{SVM},n}(X_i)| \)
MAD function

**MAD function:** \( g_0 = \text{median}(|Y - \text{median}(Y|X = \cdot)| | X = \cdot) \)

- estimate \( \text{median}(Y|X = \cdot) \) by SVM \( f_{\text{SVM}, n} \) with absolute deviation loss
- put \( \tilde{Y}_i := |Y_i - f_{\text{SVM}, n}(X_i)| \)
- use non-i.i.d. data \((X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)\) and estimate \( \text{median}(\tilde{Y}_i|X_i = \cdot) \)
  by SVM \( \tilde{g}_{\text{SVM}, n} \) with absolute deviation loss
MAD function

**MAD function:** \( g_0 = \text{median}(\lvert Y - \text{median}(Y|X = \cdot)\rvert \mid X = \cdot) \)

- estimate \( \text{median}(Y|X = \cdot) \) by SVM \( f_{\text{SVM},n} \) with absolute deviation loss
- put \( \tilde{Y}_i := \lvert Y_i - f_{\text{SVM},n}(X_i)\rvert \)
- use non-i.i.d. data \((X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)\) and estimate \( \text{median}(\tilde{Y}_i|X_i = \cdot) \)
  by SVM \( \tilde{g}_{\text{SVM},n} \) with absolute deviation loss
- put \( g_{\text{SVM},n}(x) = \max\{\tilde{g}_{\text{SVM},n}(x), 0\} \)
MAD function

MAD function: \( g_0 = \text{median}(\left| Y - \text{median}(Y|X = \cdot) \right| | X = \cdot) \)

- estimate \( \text{median}(Y|X = \cdot) \) by SVM \( f_{\text{SVM},n} \) with absolute deviation loss
- put \( \tilde{Y}_i := |Y_i - f_{\text{SVM},n}(X_i)| \)
- use non-i.i.d. data \((X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)\) and estimate \( \text{median}(\tilde{Y}_i|X_i = \cdot) \)
  by SVM \( \tilde{g}_{\text{SVM},n} \) with absolute deviation loss
- put \( g_{\text{SVM},n}(x) = \max\{\tilde{g}_{\text{SVM},n}(x), 0\} \)

Risk of function \( g \):

\[
\mathcal{R}_P(g) = \int \left| |y - f_0(x)| - g(x) \right| P(d(x, y))
\]

Then, hopefully,

\[
\lim_{n \to \infty} \left| \mathcal{R}_P(g_{\text{SVM},n}) - \mathcal{R}_P(g_0) \right| = 0
\]
MAD-type Estimation

**MAD function:** $g_0 = \text{median}(\mid Y - \text{median}(Y \mid X = \cdot) \mid \mid X = \cdot)$

- estimate $\text{median}(Y \mid X = \cdot)$ by SVM $f_{\text{SVM},n}$ with absolute deviation loss
- put $\tilde{Y}_i := \mid Y_i - f_{\text{SVM},n}(X_i) \mid$
- use non-i.i.d. data $(X_1, \tilde{Y}_1), \ldots, (X_n, \tilde{Y}_n)$ and estimate $\text{median}_\varepsilon(\tilde{Y}_i \mid X_i = \cdot)$ by SVM $\tilde{g}_{\text{SVM},n}$ with $\varepsilon$-smoothed absolute deviation loss
- put $g_{\text{SVM},n}(x) = \max\{\tilde{g}_{\text{SVM},n}(x), 0\}$

**Risk of function $g$:**

$$\mathcal{R}_P(g) = \int \left| \mid y - f_0(x) \mid - g(x) \right| P(d(x, y))$$

Then, at least,

$$\limsup_{n \to \infty} \left| \mathcal{R}_P(g_{\text{SVM},n}) - \mathcal{R}_P(g_0) \right| \leq \varepsilon$$
IQR-type Estimation

IQR function: \( g_0 = f_{0.25}^* - f_{0.75}^* \)
where \( f_{\tau}^* := \text{quantile}_\tau(Y|X = \cdot) \)

- estimate \( \text{quantile}_{0.25}(Y|X = \cdot) \) by SVM \( f_{\text{SVM.1},n} \) with 0.25-pinball loss \( L_{0.25} \)
- estimate \( \text{quantile}_{0.75}(Y|X = \cdot) \) by SVM \( f_{\text{SVM.2},n} \) with 0.75-pinball loss \( L_{0.75} \)
- put \( g_{\text{SVM},n}(x) = f_{\text{SVM.2},n}(x) - f_{\text{SVM.1},n}(x) \)
IQR-type Estimation

**IQR function:** \( g_0 = f_{0.25}^* - f_{0.75}^* \)

where \( f_{\tau}^* := \text{quantile}_\tau(Y|X=\cdot) \)

- estimate \( \text{quantile}_{0.25}(Y|X=\cdot) \) by SVM \( f_{\text{SVM},1,n} \) with 0.25-pinball loss \( L_{0.25} \)
- estimate \( \text{quantile}_{0.75}(Y|X=\cdot) \) by SVM \( f_{\text{SVM},2,n} \) with 0.75-pinball loss \( L_{0.75} \)
- put \( g_{\text{SVM},n}(x) = f_{\text{SVM},2,n}(x) - f_{\text{SVM},1,n}(x) \)

**Risk** of function \( f \):

\[
R_{P,\tau}(f) = \int L_{\tau}(y, f(x)) P(d(x, y))
\]

Then,

\[
\lim_{n \to \infty} \left| \left( R_{P,0.25}(f_{\text{SVM},1,n}) + R_{P,0.75}(f_{\text{SVM},2,n}) \right) - \left( R_{P,0.25}(f_{0.25}^*) + R_{P,0.75}(f_{0.75}^*) \right) \right| = 0
\]
Example: LIDAR data set

LIDAR = Light Detection And Ranging (data set with \( n = 221 \))
Example: LIDAR data set

$\tau$-quantile estimation with SVM

$\tau \in \{5\%, 25\%, 50\%, 75\%, 95\%\}$

estimated width of interval $I$
with $\mathbb{P}(Y \in I|X = x) = 0.5$
IQR-type, $2 \times$ MAD-type
Robustness

Example

- LIDAR data set
- but now with additional 10% Cauchy errors in $y$-direction
Robustness

Example

- LIDAR data set
- but now with additional 10% Cauchy errors in \( y \)-direction
Robustness

Example: IQR-type estimation

- LIDAR data set
- but now with additional 10% Cauchy errors in $y$-direction

quantile regression for lidar data set

width of interval covering 50%
Robustness

Example: IQR-type estimation
- LIDAR data set
- but now with additional 10% Cauchy errors in $y$-direction

quantile regression for lidar data set

width of interval covering 50%
Robustness: Theoretical Results

**MAD-type:**
- uniform upper bound for the bias

**IQR-type:**
- uniform upper bound for the bias
- bounded Bouligand influence function
- qualitative robustness
References

- **Hable and Christmann (2011)**. Estimation of scale functions to model heteroscedasticity by support vector machines. *arXiv*:1111.1830.


The handout to this talk is also available on my homepage

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