Beyond Least Squares

Marco C. Campi

University of Brescia
Italy

(*joint work with Algo Carè
and Simone Garatti*)

Paul’s day – Eindhoven, 19 April, 2024

⟶ *LS is our key-method to construct models*

⟶ *LS is our key-method to construct models*

Pros:

- *returns a single model (handy for design, e.g. to contruct a controller)*

- *compromizes among various situations and returns a "central" model*

⟶ *LS is our key-method to construct models*

Pros:

- *returns a single model (handy for design, e.g. to contruct a controller)*

- *compromizes among various situations and returns a "central" model*

(compromizing is good, …

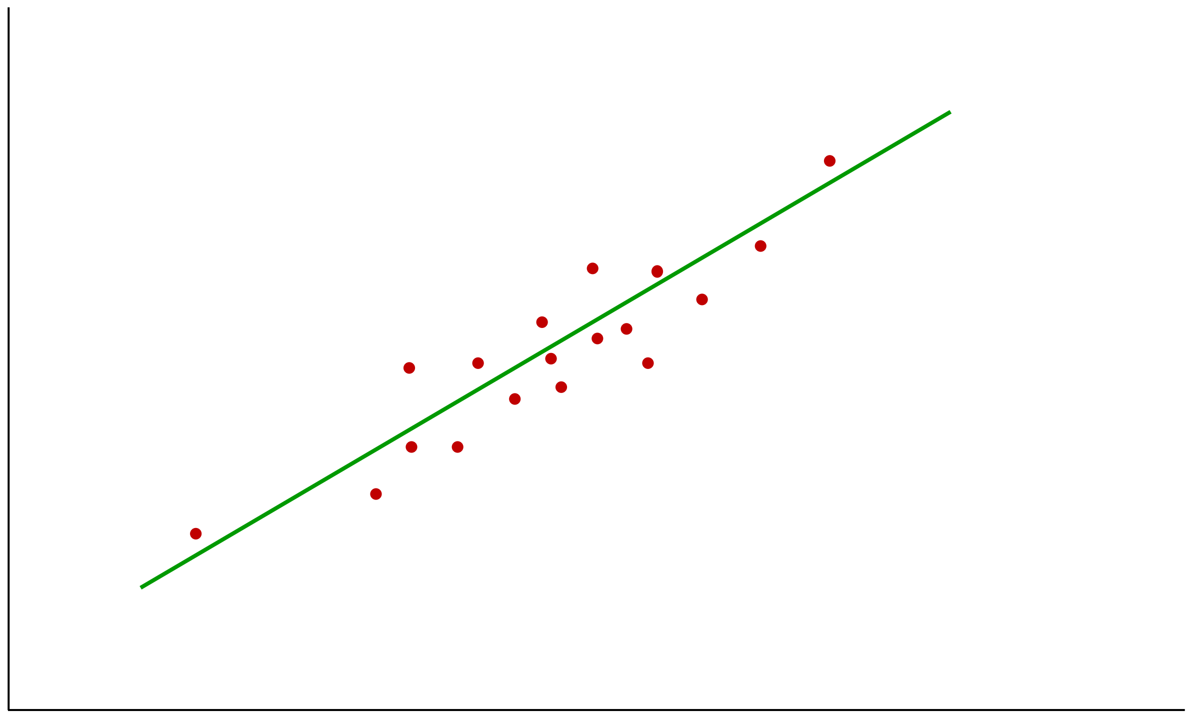
⟶ *LS is our key-method to construct models*

Pros:

- *returns a single model (handy for design, e.g. to contruct a controller)*
- *compromizes among various situations and returns a "central" model*

(compromizing is good, … especially true for an Italian man)

Least Squares



→ *not all models are built by an “averaging approach”*

⟶ *not all models are built by an “averaging approach”*

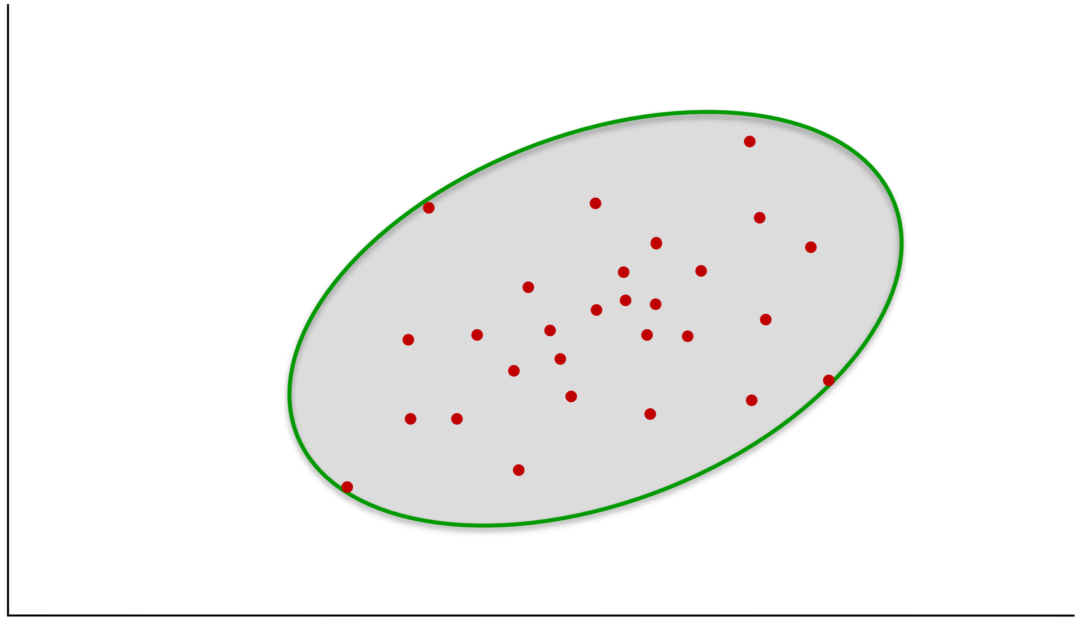
there was a flourishing statistical literature in the 1950s on so-called 'tolerance regions'

(Fraser, Guttman, Kemperman, Tukey, Wilks,…)

⟶ *not all models are built by an “averaging approach”*

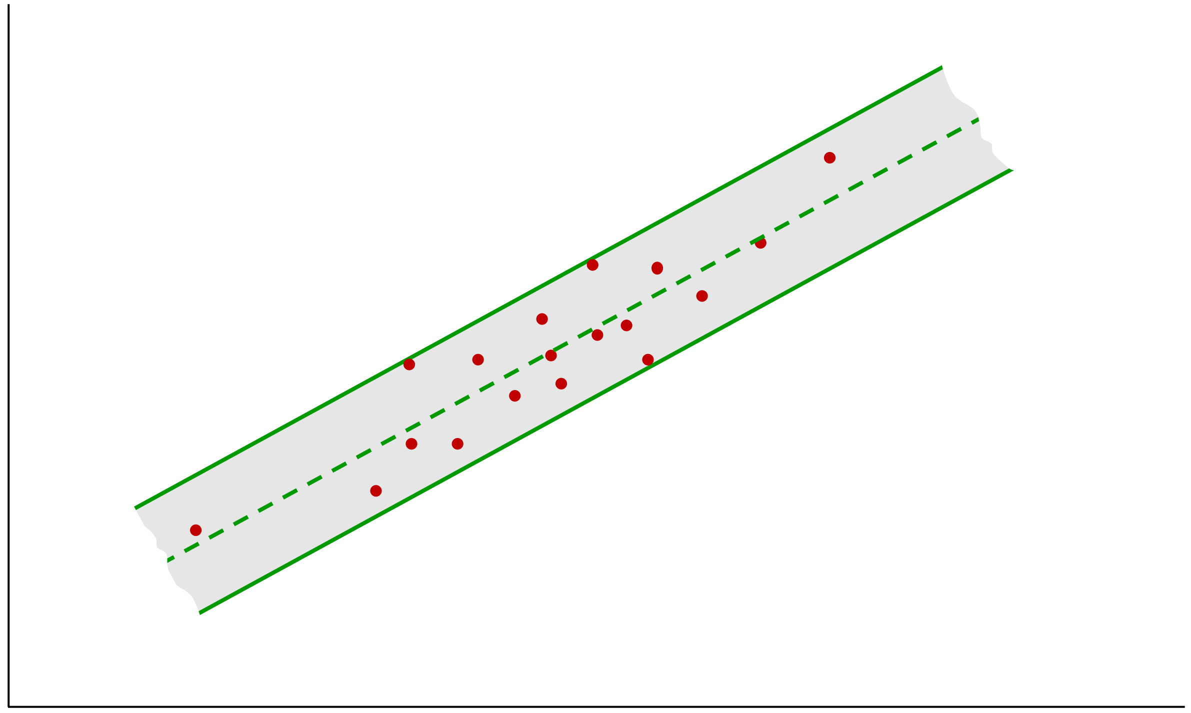
there was a flourishing statistical literature in the 1950s on so-called 'tolerance regions'

(Fraser, Guttman, Kemperman, Tukey, Wilks,…)

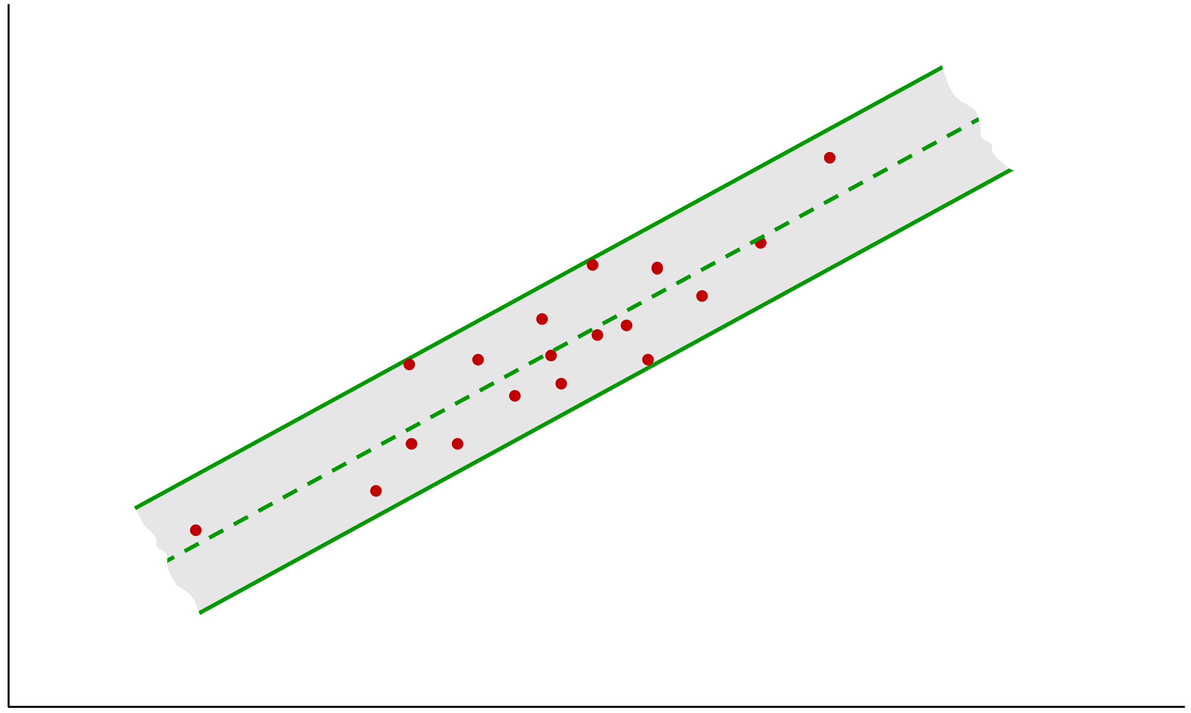


here, the idea is that of “coverage”

… back to example



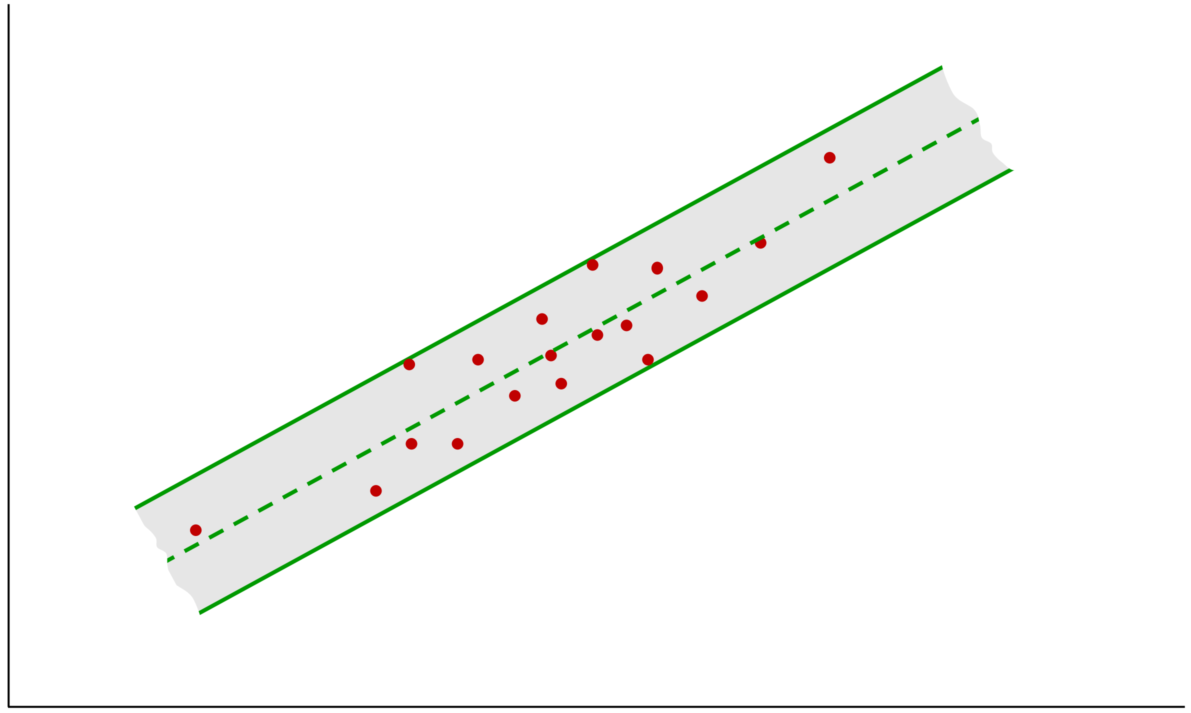
… back to example



how is this constructed?

minimize size while keeping points inside

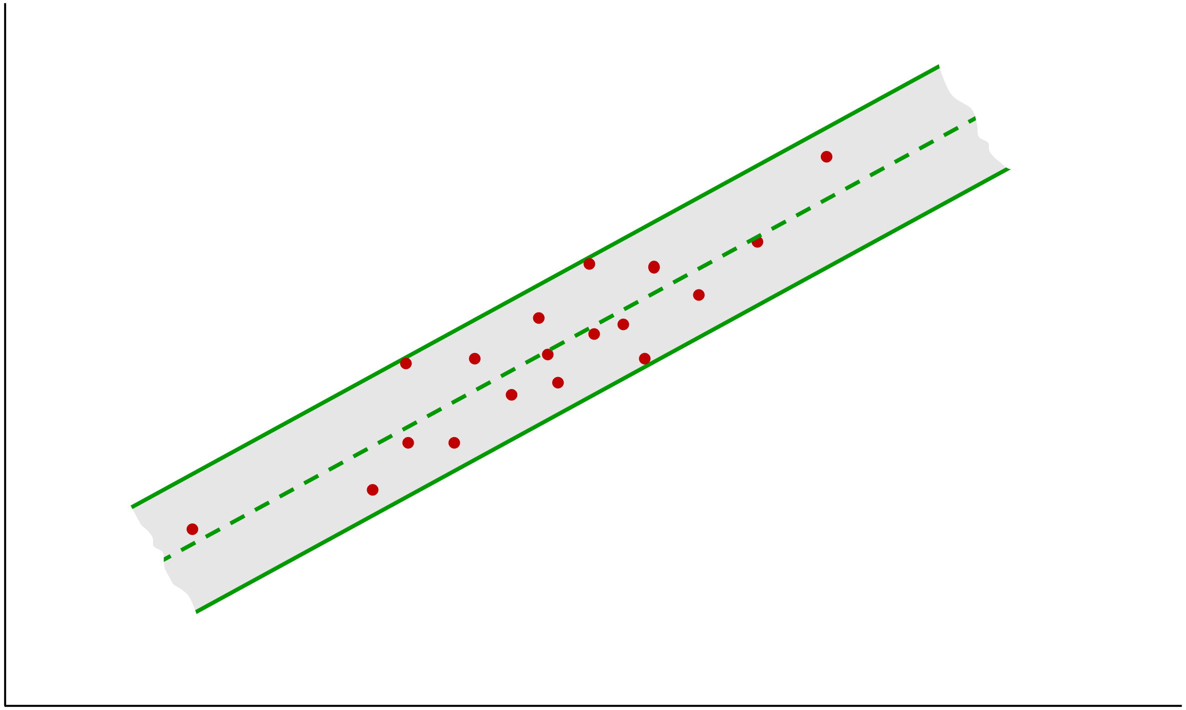
… back to example



how is this constructed?

minimize size while keeping points inside
(inherently different from enlarging from LS estimate)

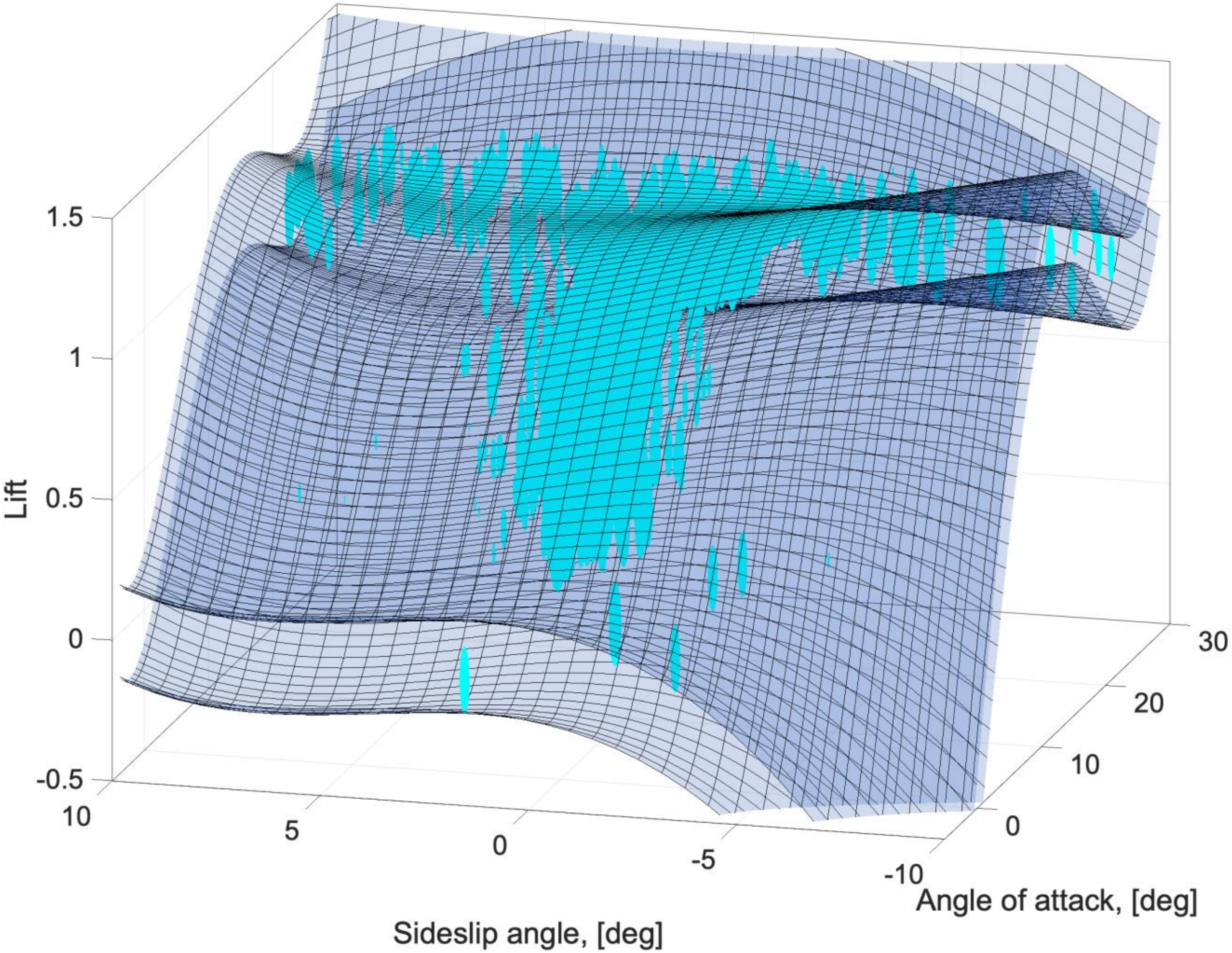
… back to example



in formulas:

$$\min_{\theta} \sum_{i=1}^{N} (y_i - f_\theta(u_i))^2 \quad \Longrightarrow \quad \min_{\theta} \max_{i=1,\ldots,N} |y_i - f_\theta(u_i)|$$

another example



→ *min-max modelling was introduced by Leonhard Euler some half a century before least squares*

conservative? … well, relax:

conservative? … well, relax:

$$\min_{\theta} \max_{\{95\% \; of \;\; the \;\; i's\}} |y_i - f_\theta(u_i)|$$

discarding

permit exclusion of some “odd” data points

conservative? … well, relax:

$$\min_{\theta} \max_{\{95\%\ of\ the\ i's\}} |y_i - f_\theta(u_i)|$$

<u>discarding</u>
permit exclusion of some "odd" data points

$$\min_{\theta} \max_{i=1,\ldots,N} |y_i - f_\theta(u_i)|$$

$$\Downarrow$$

$$\min_{\theta,\gamma} \gamma$$

$$\text{subject to: } |y_i - f_\theta(u_i)| \leq \gamma$$

conservative? … well, relax:

$$\min_{\theta} \max_{\{95\% \ of \ the \ i's\}} |y_i - f_\theta(u_i)|$$

discarding

permit exclusion of some “odd” data points

$$\min_{\theta} \max_{i=1,\dots,N} |y_i - f_\theta(u_i)|$$

⇓

$$\min_{\theta,\gamma} \gamma$$

$$\text{subject to: } |y_i - f_\theta(u_i)| \le \gamma$$

⇒

$$\min_{\theta,\gamma,\xi_i \ge 0} \gamma + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{subject to: } |y_i - f_\theta(u_i)| - \gamma \le \xi_i$$

relaxation

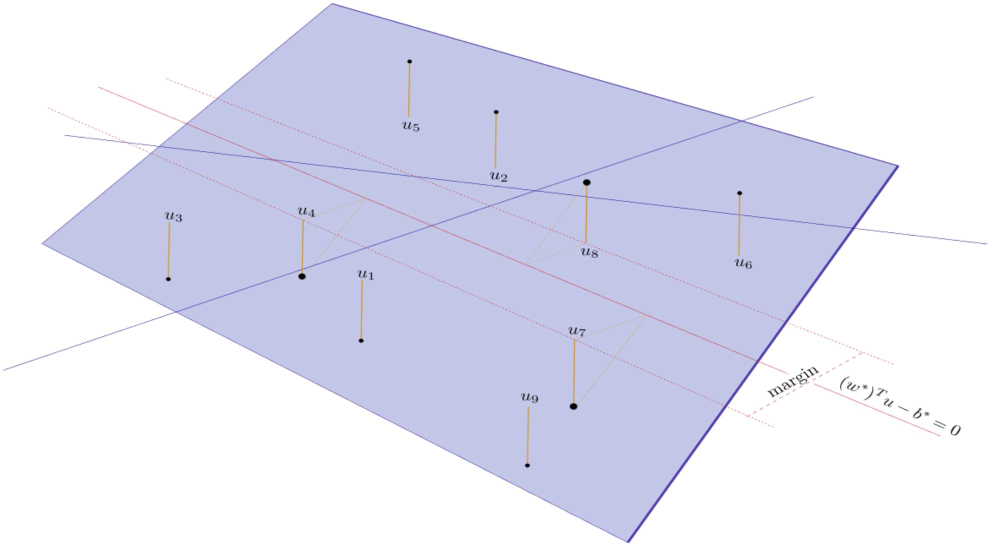
→ *interestingly, these approaches are quite popular in the ML community*

⟶ *interestingly, these approaches are quite popular in the ML community*

for example, SVM:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi_i \geq 0} \|w\|^2 + \rho \sum_{i=1}^{N} \xi_i$$

$$\text{subject to: } 1 - y_i(\langle w, u_i \rangle - b) \leq \xi_i$$



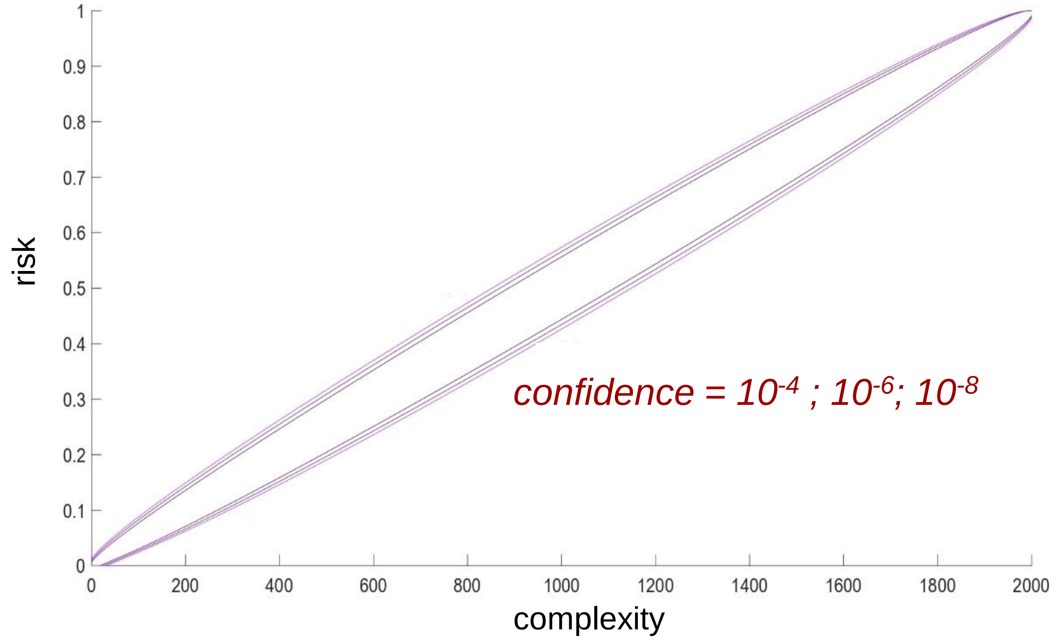
a truly beautiful theory

a truly beautiful theory

in the i.i.d. case, these methods come with an enthralling theory.

a truly beautiful theory

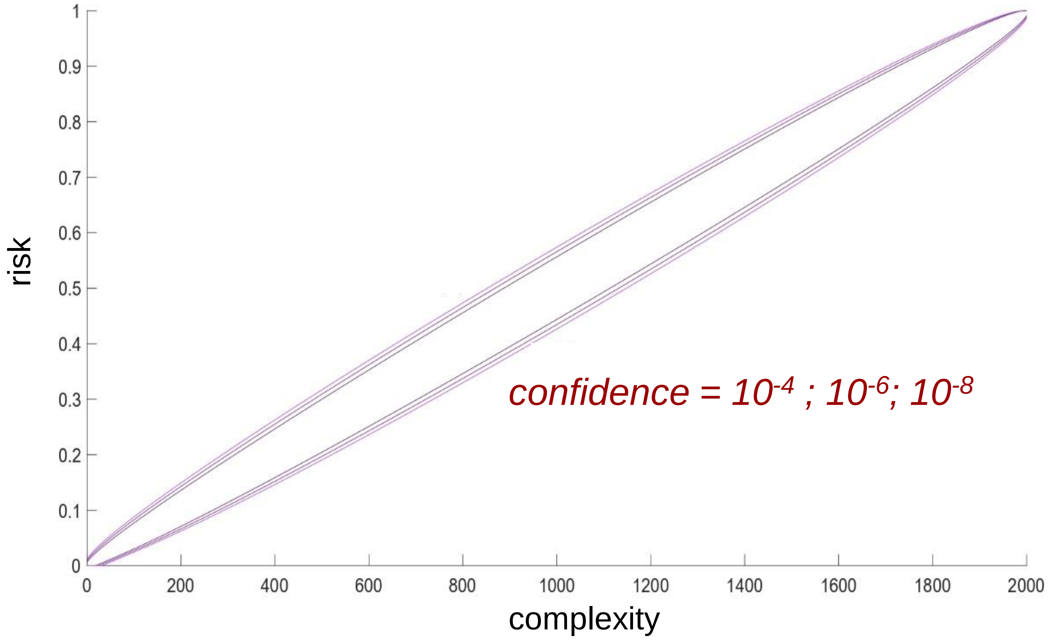
in the i.i.d. case, these methods come with an enthralling theory.



confidence = 10^{-4} ; 10^{-6}; 10^{-8}

a truly beautiful theory

in the i.i.d. case, these methods come with an enthralling theory.

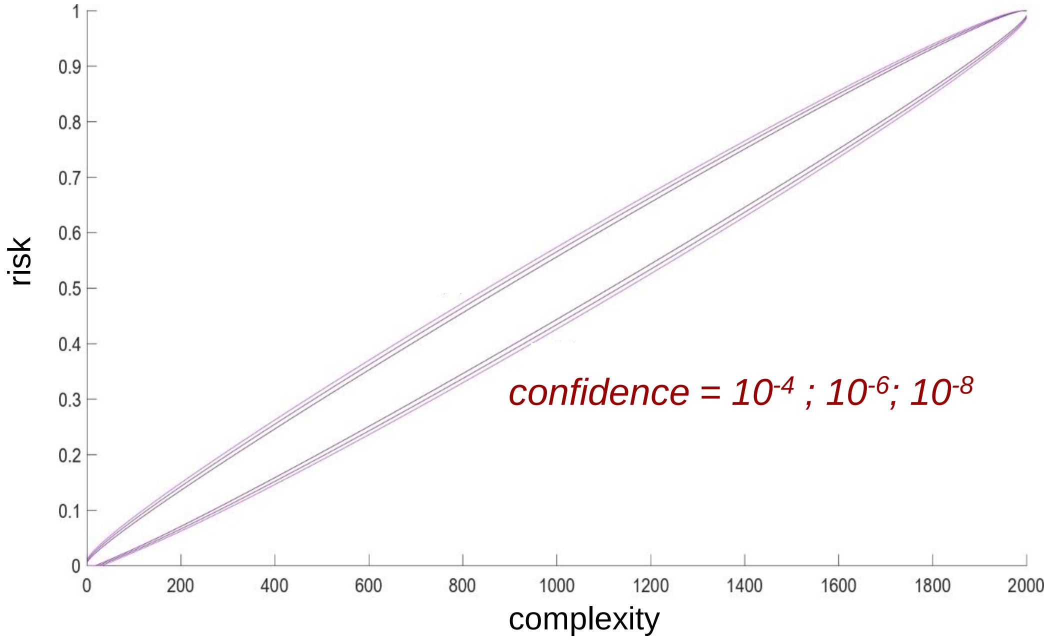


confidence = 10^{-4} ; 10^{-6}; 10^{-8}

holds distribution-free!

a truly beautiful theory

in the i.i.d. case, these methods come with an enthralling theory.



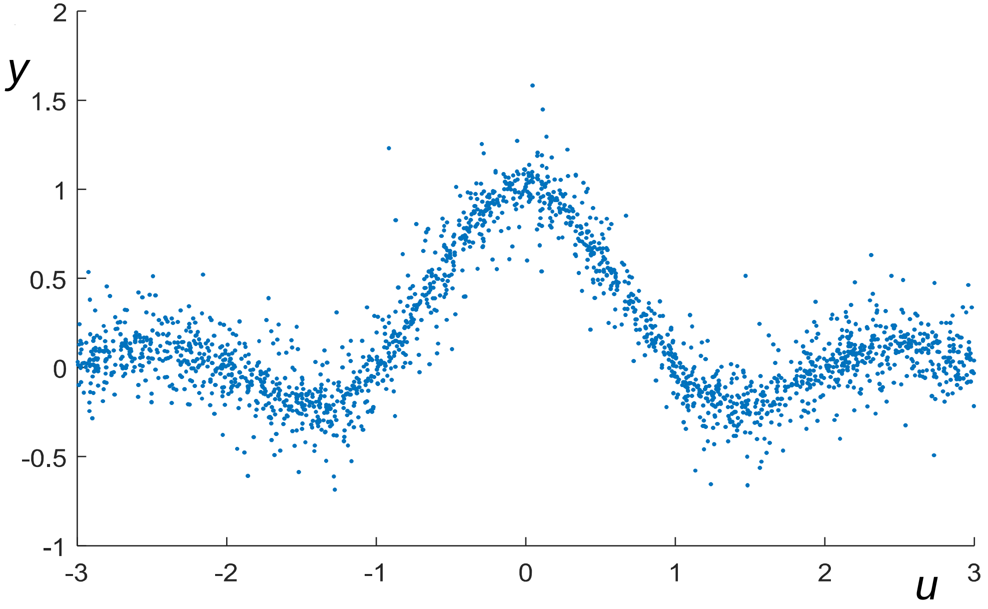
confidence = 10^{-4} ; 10^{-6}; 10^{-8}

holds distribution-free!

→ *develop trust in the model*

→ *tune hyper-parameters*

Example: SVR

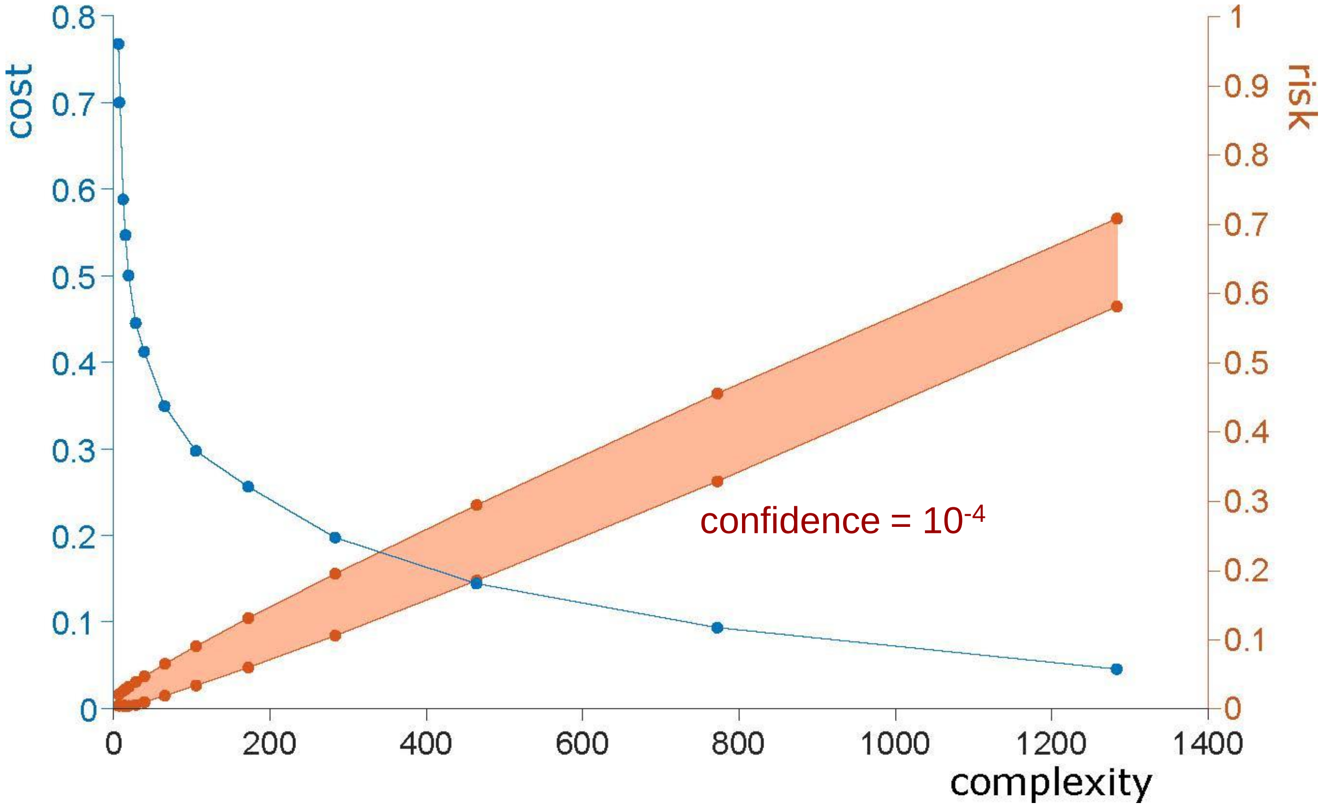


$$\min_{w,\gamma,b,\xi_i\geq 0} (\gamma + 0.01\|w\|^2) + \rho \sum_{i=1}^{N} \xi_i$$

subject to: $|y_i - \langle w, \phi_i \rangle - b| - \gamma \leq \xi_i, \quad i = 1, \ldots, 2000$

$\langle \phi_i, \phi_j \rangle = \exp(-(u_i - u_j))^2$ (Gaussian kernel)

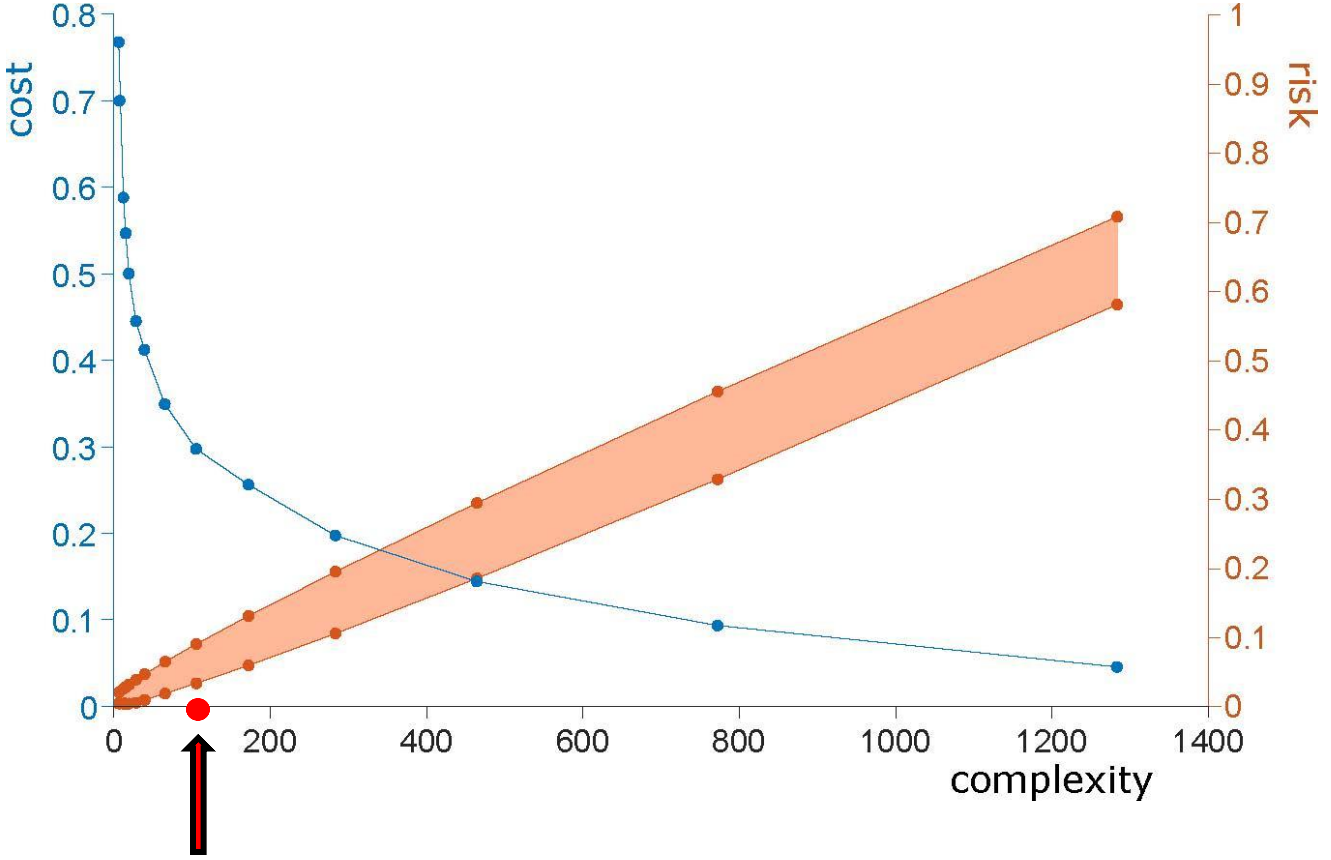
Example: SVR



confidence = 10^{-4}

$$\rho = \left(\frac{3}{5}\right)^{\ell}, \quad \ell = 0, \ldots, 14$$

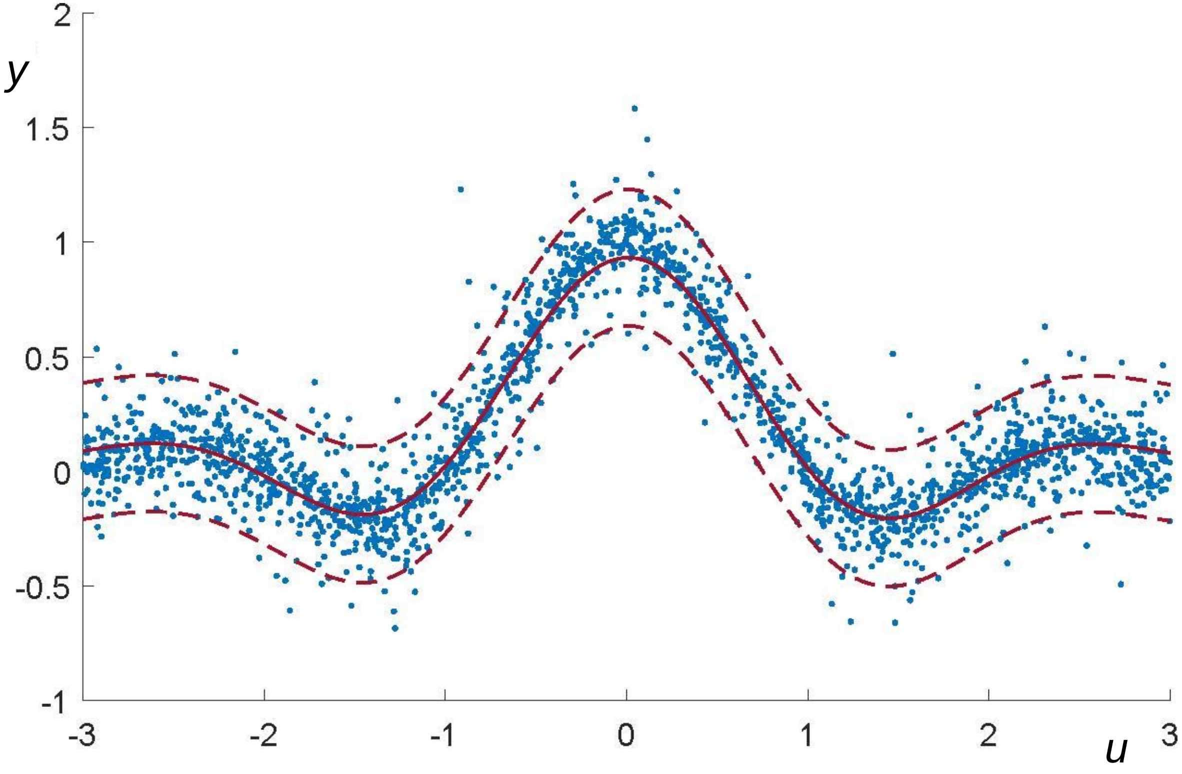
Example: SVR



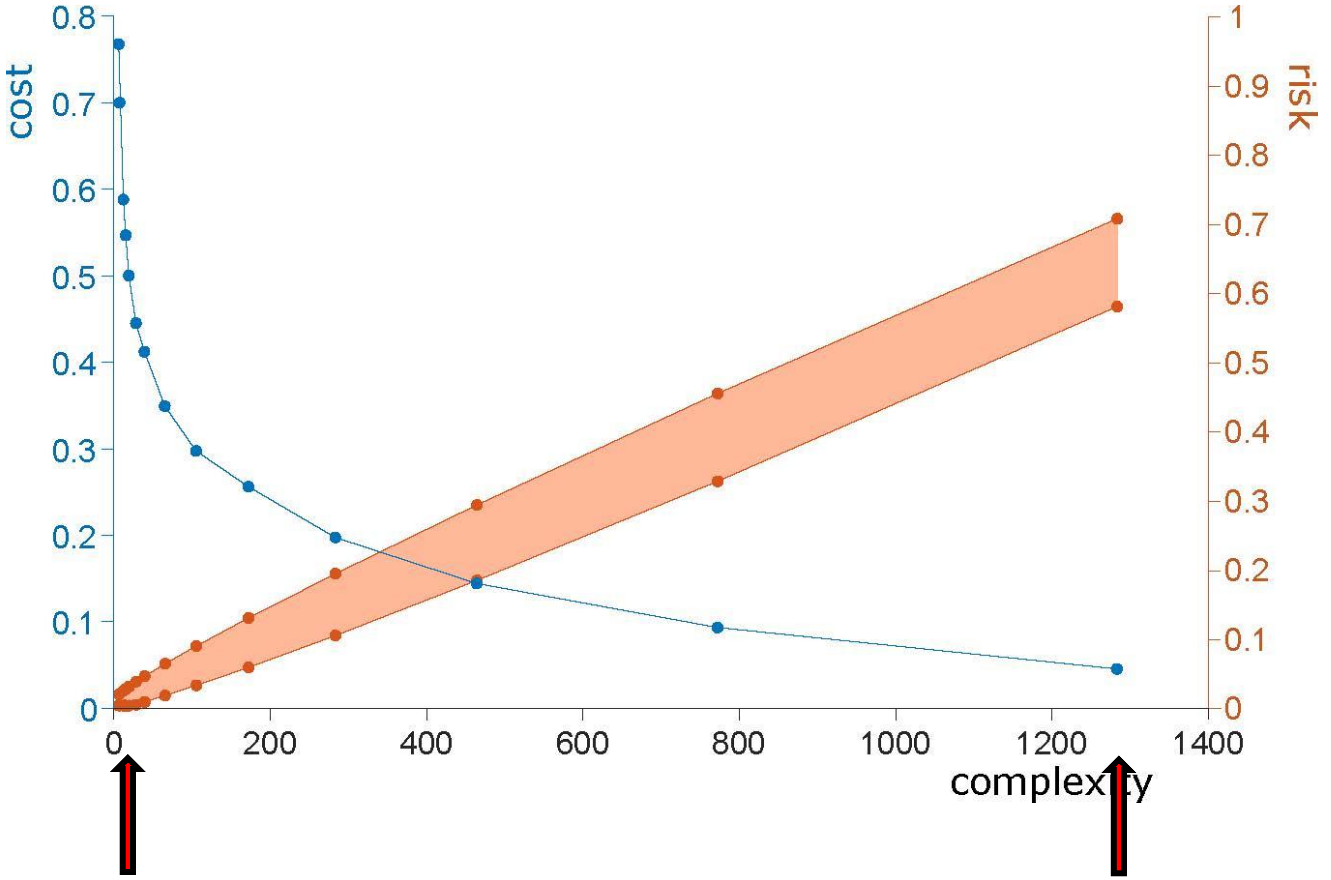
$\rho = (3/5)^9$

$\text{risk} \in [0.032, 0.08] \qquad \gamma^* = 0.3$

Example: SVR



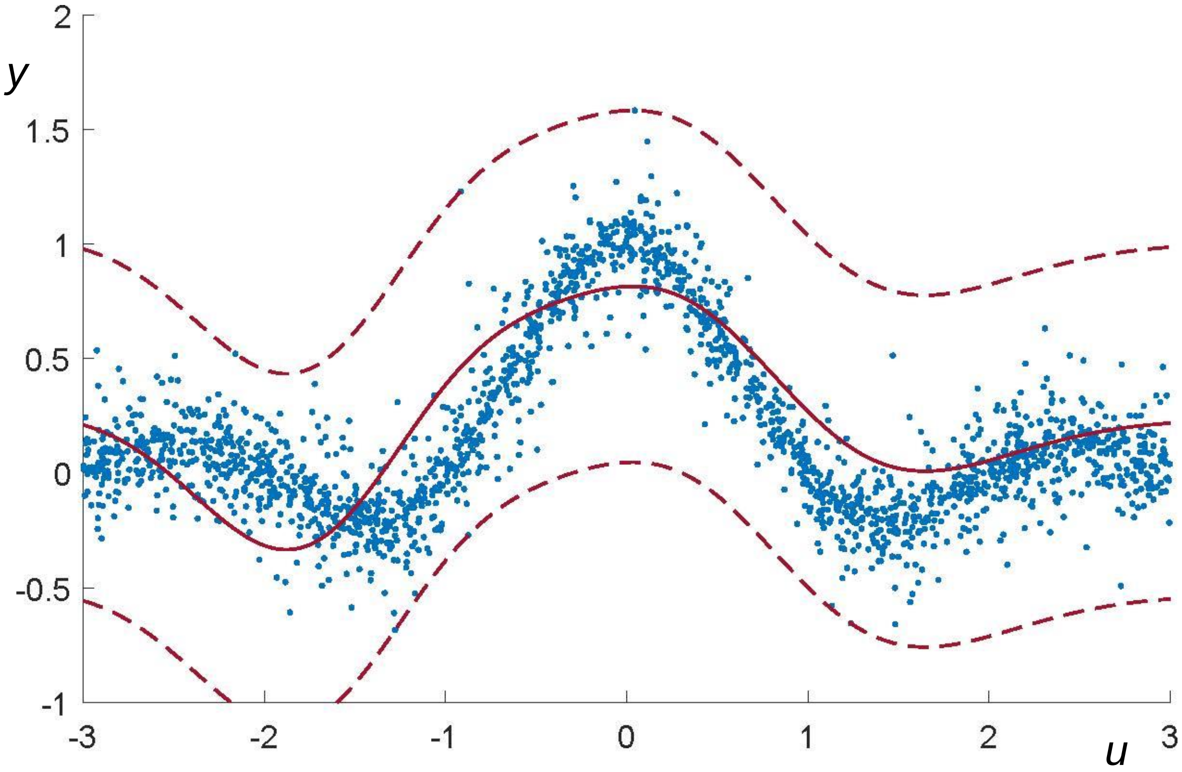
Example: SVR



$\rho = (3/5)^0$

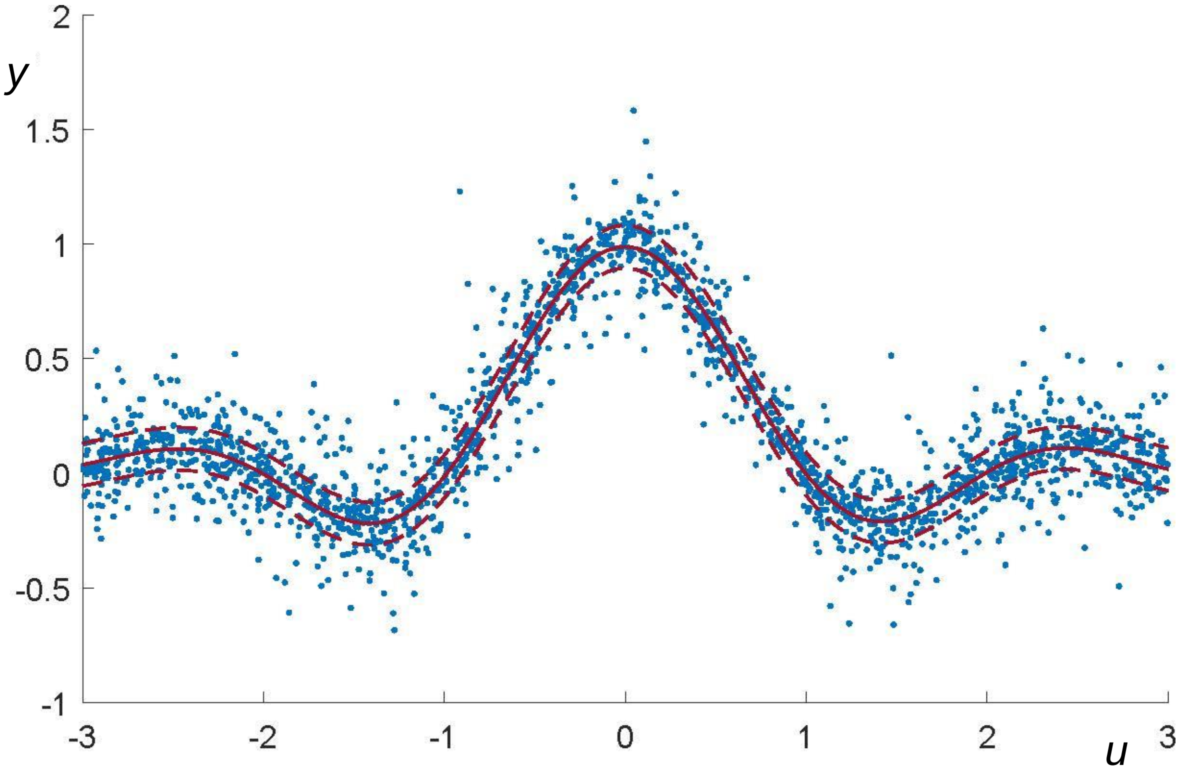
$\rho = (3/5)^{14}$

Example: SVR



$\rho = (3/5)^0$

Example: SVR

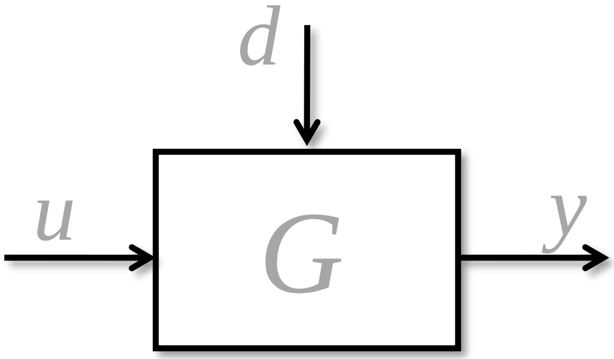


$\rho = (3/5)^{14}$

should we try to import these methods into identification and identification for control?

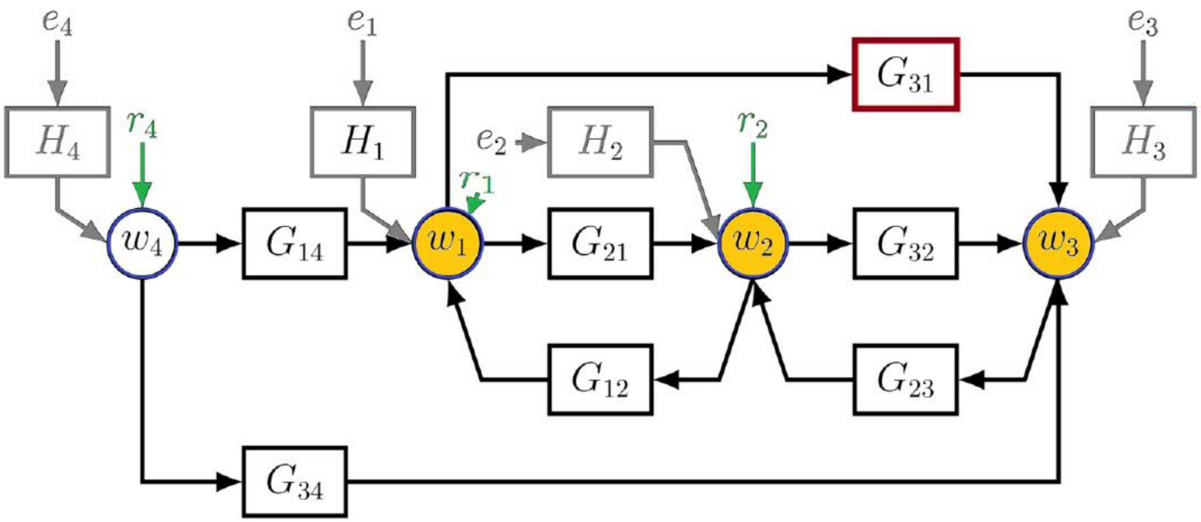
should we try to import these methods into identification and identification for control?

the "big challenge": move away from i.i.d



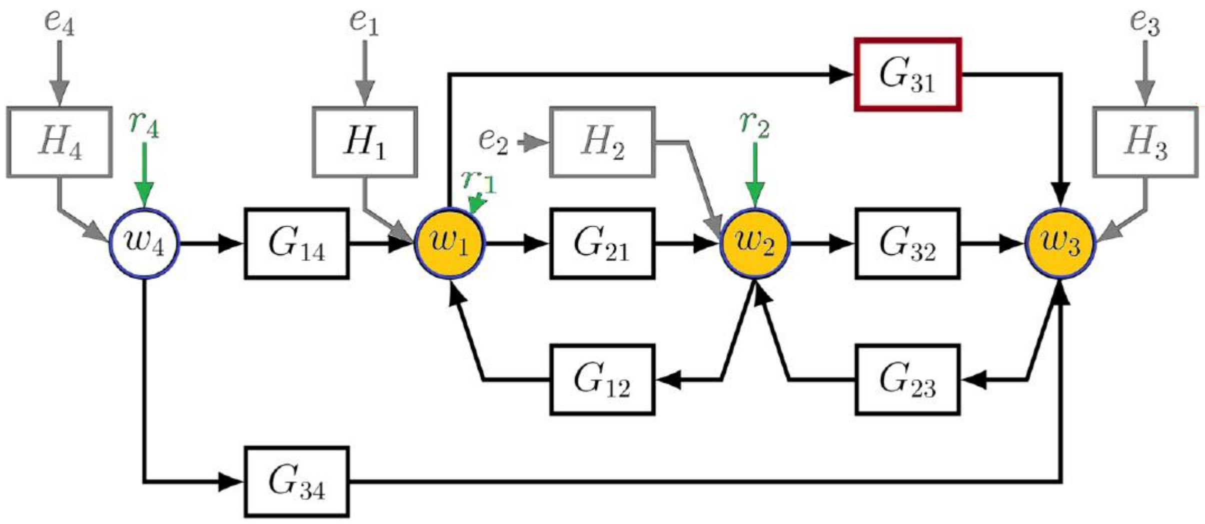
should we try to import these methods into identification and identification for control?

the "big challenge": move away from i.i.d



should we try to import these methods into identification and identification for control?

the "big challenge": move away from i.i.d



something one of us knows very well!

should we try to import these methods into identification and identification for control?

the "big challenge": move away from i.i.d

a smart guy!

should we try to import these methods into identification and identification for control?

the "big challenge": move away from i.i.d

a smart guy!

Ad maiora, Paul!