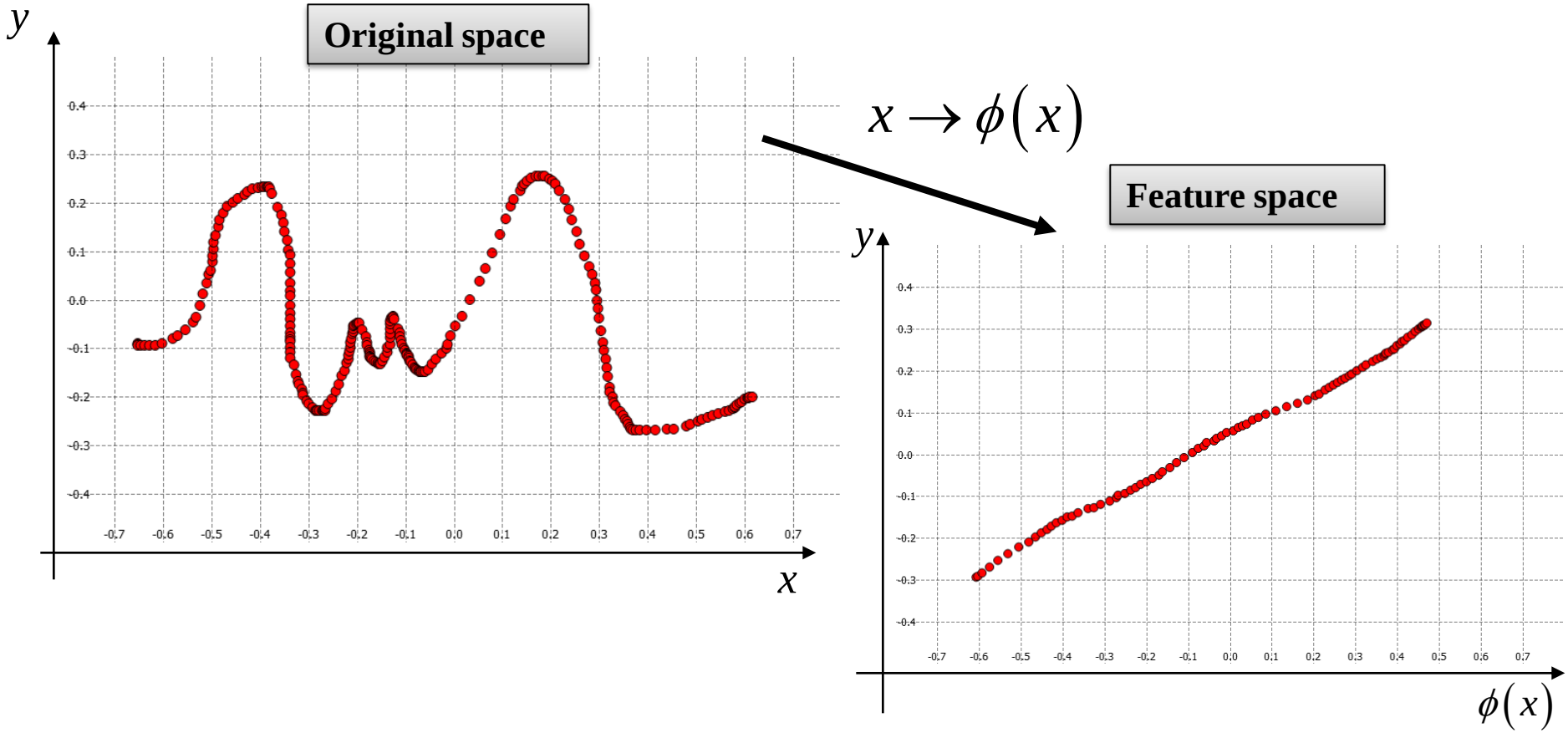
Kernel trick

SVR extensions
ν-SVR, RVR

Recap: Nonlinear regression with SVR



Principle: Assume there exists a transformation ϕ such that the problem becomes linear

→ Perform linear regression in feature space

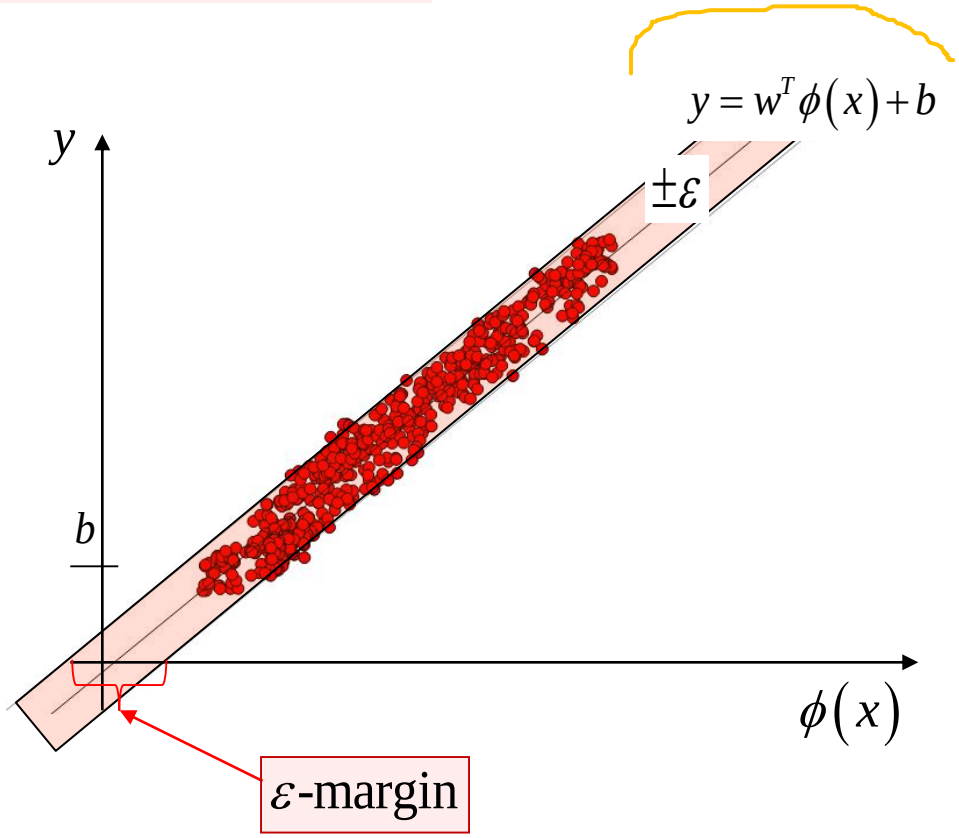
Recap: Nonlinear regression with SVR

The ε-insensitive tube is the region surrounding the regression function that encapsulates acceptable noise.

To maximize the ε-margin, minimize the norm of w.

Assume deterministic noise model: $y \pm \varepsilon$

Consider as correctly fit all points such that $f(x) - y \leq \varepsilon$.



Recap: Nonlinear regression with SVR

This can be expressed as a constraint-based optimization problem of the form:

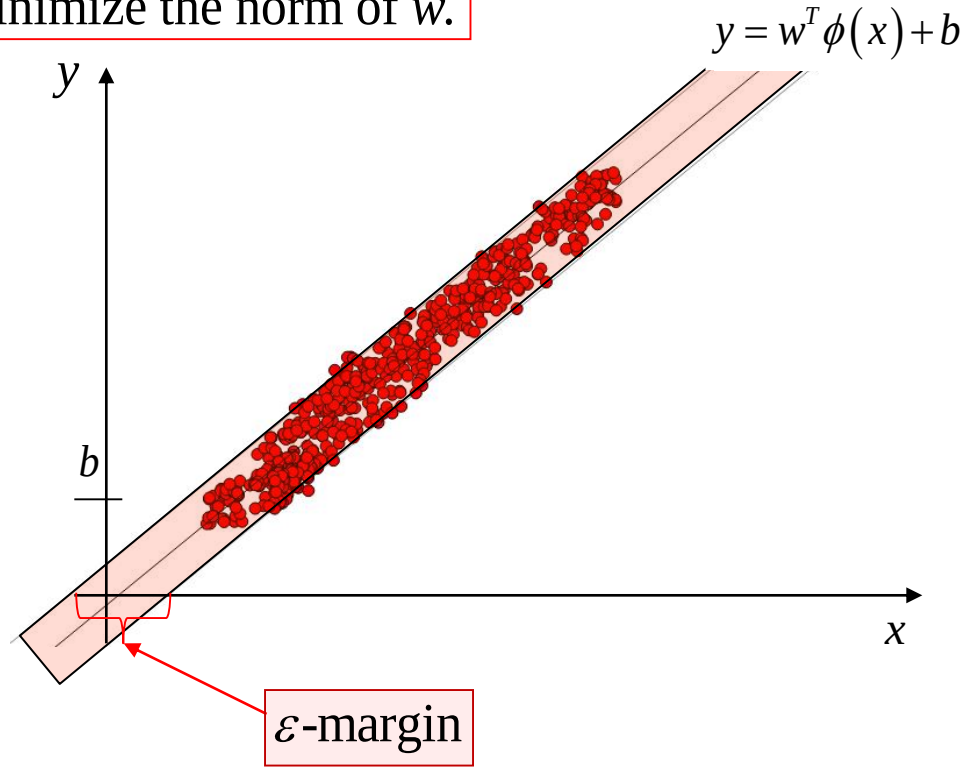
maximize the ε-margin,
minimize the norm of w.

$$\text{minimize } \frac{1}{2}\|w\|^2$$

$$\text{subject to } \begin{cases} \langle w, \phi(x^i) \rangle + b - y^i \leq \varepsilon \\ y^i - \langle w, \phi(x^i) \rangle - b \leq \varepsilon \end{cases}$$

$$\forall i = 1, \ldots M$$

Consider as correctly fit all points such that $|f(x) - y| \leq \varepsilon$.



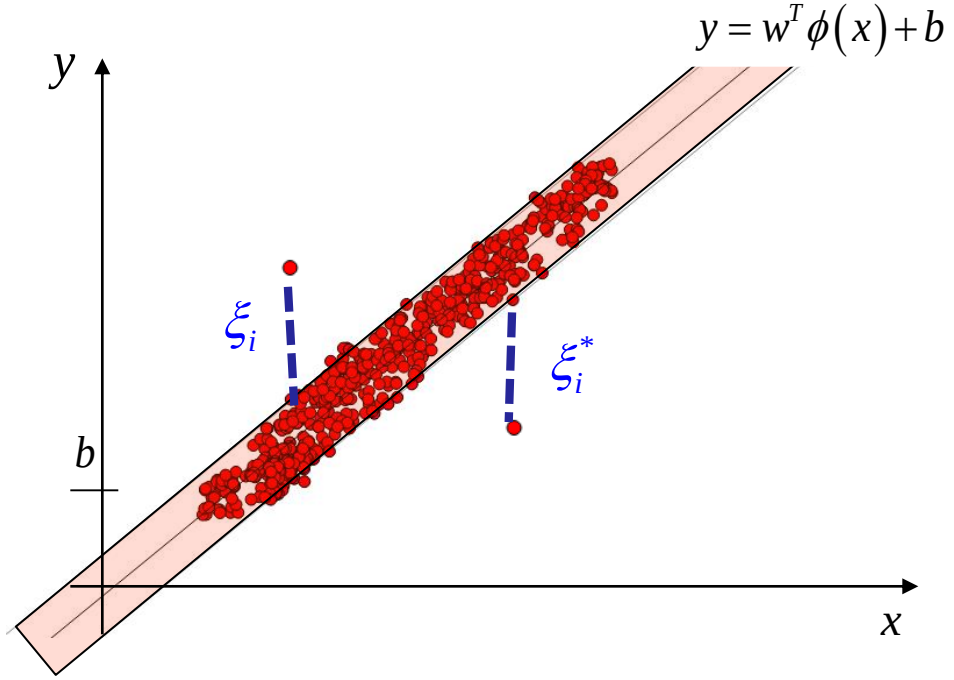
Recap: Nonlinear regression with SVR

Relaxes constraints and introduce slack variables ξ_i, ξ_i^*.

Penalizes for points outside the ε – tube, $C \geq 0$

$$\text{minimize } \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right)$$

$$\text{subject to } \begin{cases} \left\langle w, \phi\left(x^i\right)\right\rangle + b - y^i \leq \varepsilon + \xi_i \\ y^i - \left\langle w, \phi\left(x^i\right)\right\rangle - b \leq \varepsilon + \xi_i^* \\ \xi_i \geq 0, \quad \xi_i^* \geq 0 \end{cases}$$



Recap: SVR Hyperparameters

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right)$$

$$\text{subject to} \begin{cases} \left\langle w, \phi\left(x^i\right)\right\rangle + b - y^i \le \varepsilon + \xi_i \\ y^i - \left\langle w, \phi\left(x^i\right)\right\rangle - b \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \quad \xi_i^* \ge 0 \end{cases}$$

$$y = f(x) = \sum_{i=1}^{M} \alpha_i \, k\left(x^i, x\right) + b$$

Dual optimization problem:

$$\max_{\alpha,\alpha^*} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{M} \alpha_i \alpha_j \cdot \underbrace{\left\langle \phi\left(x^i\right), \phi\left(x^j\right)\right\rangle}_{k\left(x^i, x^j\right)} \\ -\varepsilon\sum_{i=1}^{M}\alpha_i + \sum_{i=1}^{M} y^i \alpha_i \end{cases}$$

$$\text{subject to} \sum_{i=1}^{M}\alpha_i = 0 \quad \text{and} \quad \alpha_i \in \left[0, \frac{C}{M}\right]$$

Recap: SVR Hyperparameters

Value for hyperparameter C is unbounded. C>0.

$$\text{minimize } \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right)$$

$$\text{subject to } \begin{cases} \left\langle w, \phi\left(x^i\right)\right\rangle + b - y^i \le \varepsilon + \xi_i \\ y^i - \left\langle w, \phi\left(x^i\right)\right\rangle - b \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \quad \xi_i^* \ge 0 \end{cases}$$

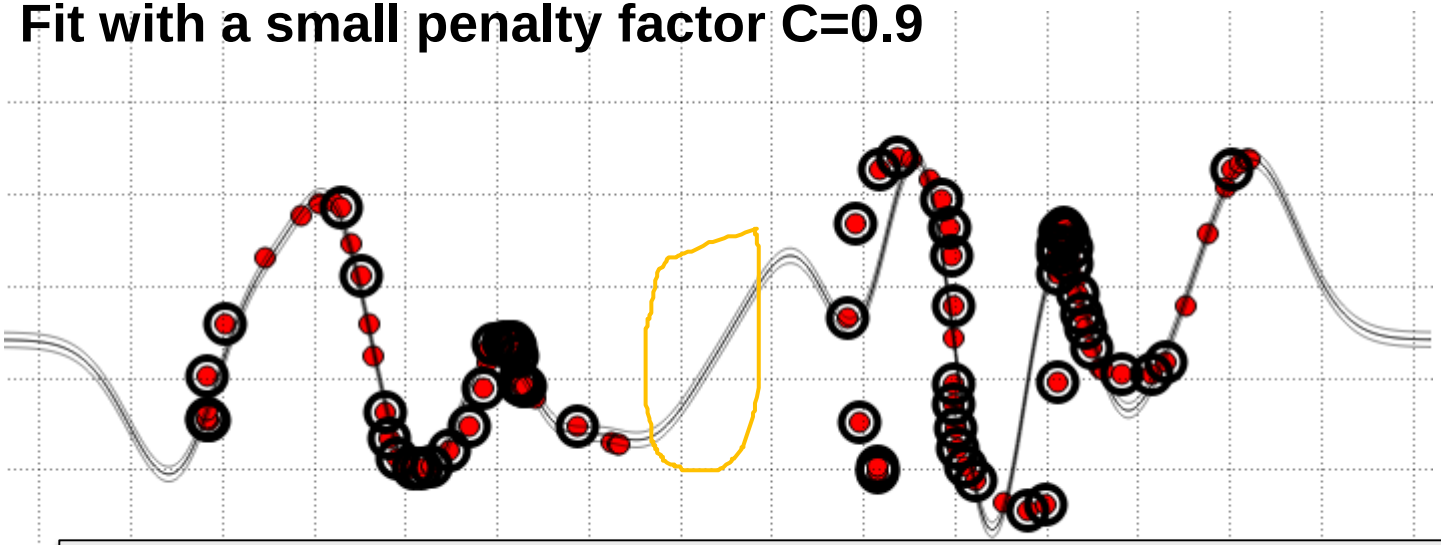
Dual optimization problem:

$$\max_{\alpha,\alpha^*} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{M}\alpha_i\alpha_j \cdot \left\langle \phi\left(x^i\right), \phi\left(x^j\right)\right\rangle \\ -\varepsilon\sum_{i=1}^{M}\alpha_i + \sum_{i=1}^{M} y^i\alpha_i \end{cases}$$

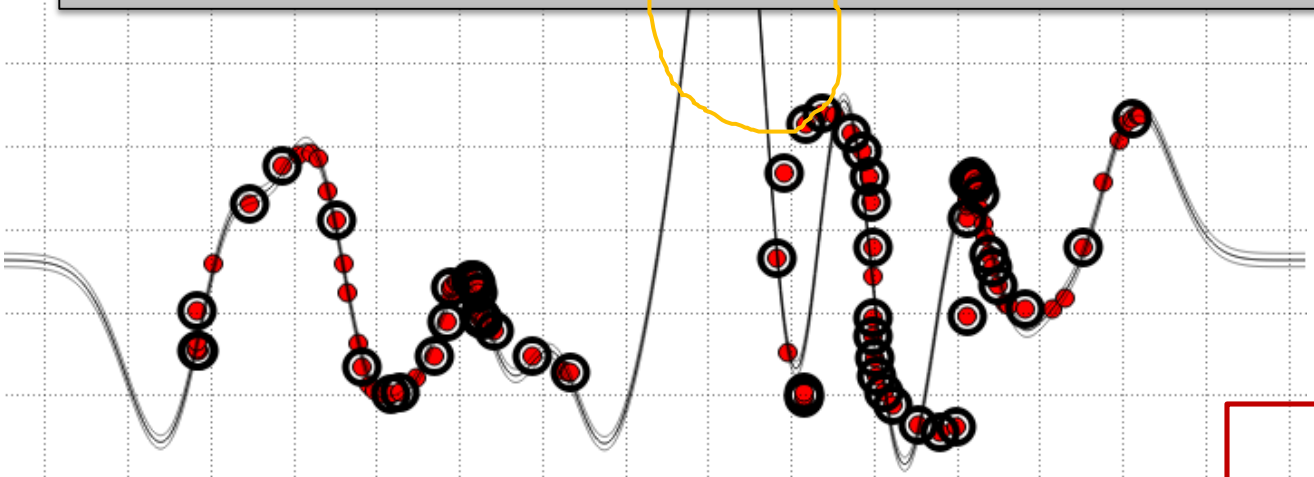
$$\text{subject to } \sum_{i=1}^{M}\alpha_i = 0 \text{ and } \alpha_i \in \left[0, \frac{C}{M}\right]$$

C becomes an upper bound on the absolute value of the α!

Fit with a small penalty factor C=0.9



When performing grid search on the open parameters, choose C to fit within the deviation of *y and not order of magnitude bigger as in the above example!*



Fit with a penalty factor C=10

C becomes an upper bound on the absolute value of the α!

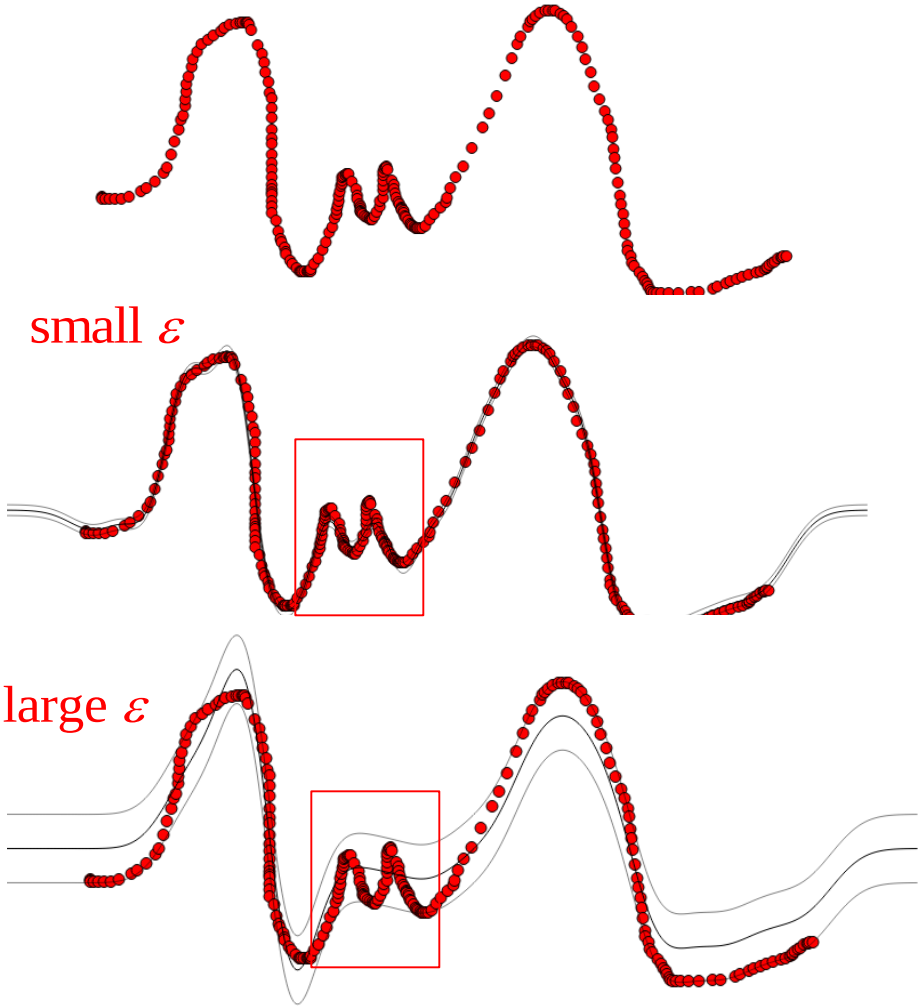
$$\alpha_i \in \left[0, \frac{C}{M}\right]$$

Recap: SVR Hyperparameters

$\varepsilon > 0$ controls the precision of the fit

$$\text{minimize } \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right)$$

$$\text{subject to } \begin{cases} \left\langle w, \phi\left(x^i\right)\right\rangle + b - y^i \le \varepsilon + \xi_i \\ y^i - \left\langle w, \phi\left(x^i\right)\right\rangle - b \le \varepsilon + \xi_i^* \\ \xi_i \ge 0, \quad \xi_i^* \ge 0 \end{cases}$$



When performing grid search on the open parameters, choose ε to be within the noise observed on *y* and the kernel width (for RBF kernels) within the variance on x.

Support Vector Regression: ν-SVR

The standard version of SVR is referred to as ε-SVR.

ν-SVM exploits and rewrites the problem as a convex optimization expression:

$$\min_{w,\xi,\varepsilon}\left(\frac{1}{2}\|w\|^2+\left[\nu\varepsilon+\frac{1}{M}\sum_{j=1}^{M}\left(\xi_j+\xi_j^*\right)\right]\right)$$ under constraints

$$\left(w^T\cdot x^j+b\right)-y^j\geq\varepsilon+\xi_j,$$

$$y^j-\left(w^T\cdot x^j+b\right)\geq\varepsilon+\xi_j^*,$$

$$\varepsilon\geq 0,\quad 0\leq\nu\leq 1,\quad \xi_j\geq 0,\quad \xi_j^*\geq 0.$$

Penalizes increases of ε.

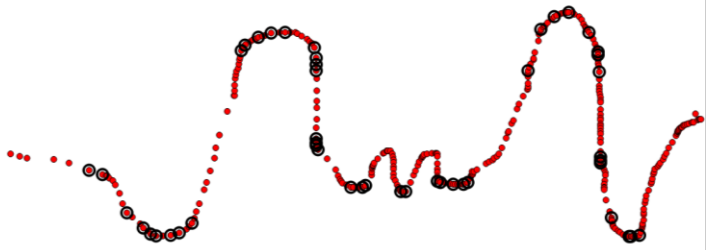
As ε decreases, the number of points outside the $\varepsilon-$tube increases. The last term penalizes decreases of ε.

Problem still convex!

ν is a lower bound on the fraction of support vectors (see classification version – ν-SVM)

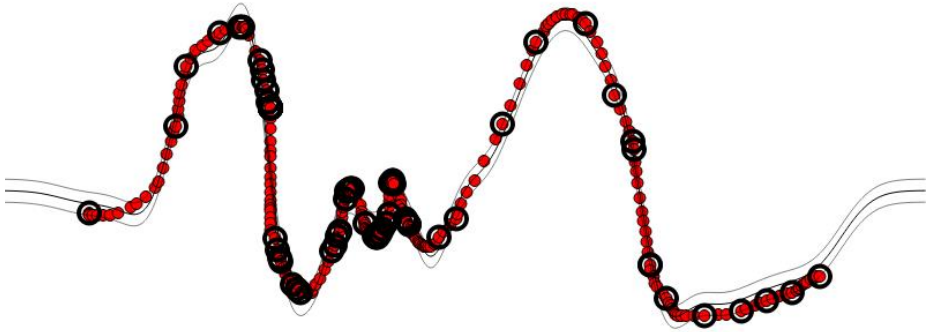
Support Vector Regression: ν-SVR

As the number of data grows, so does the number of support vectors.

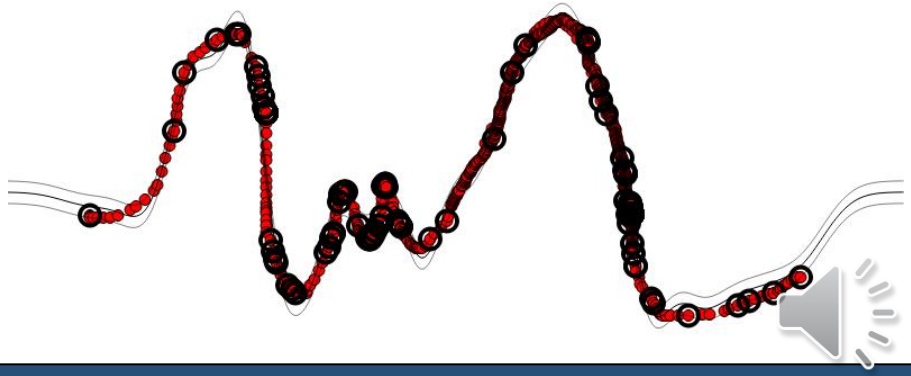


ν-SVR – 49 SVs

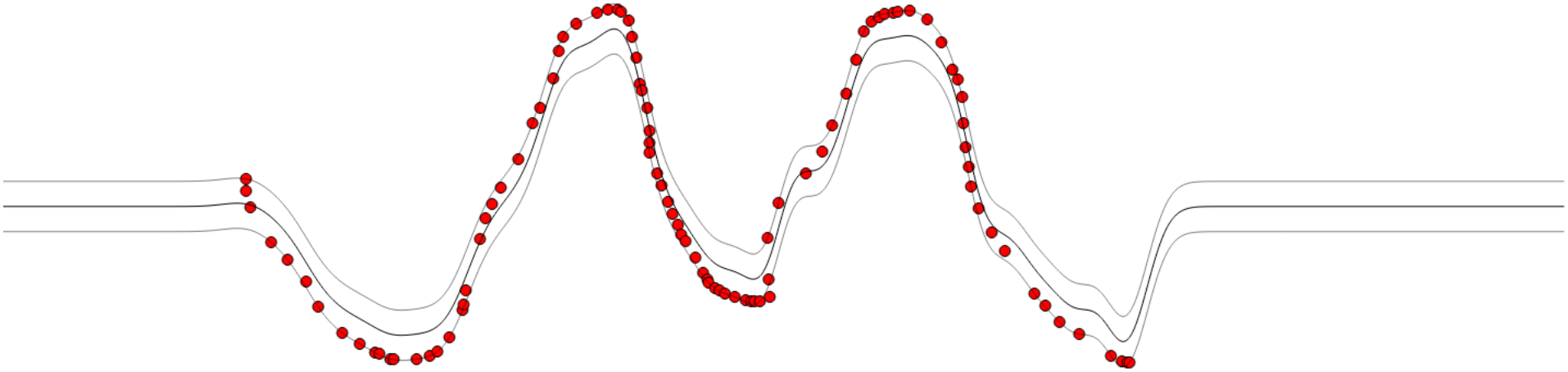
500 datapoints – 79 SVs



1200 datapoints – 117 SVs

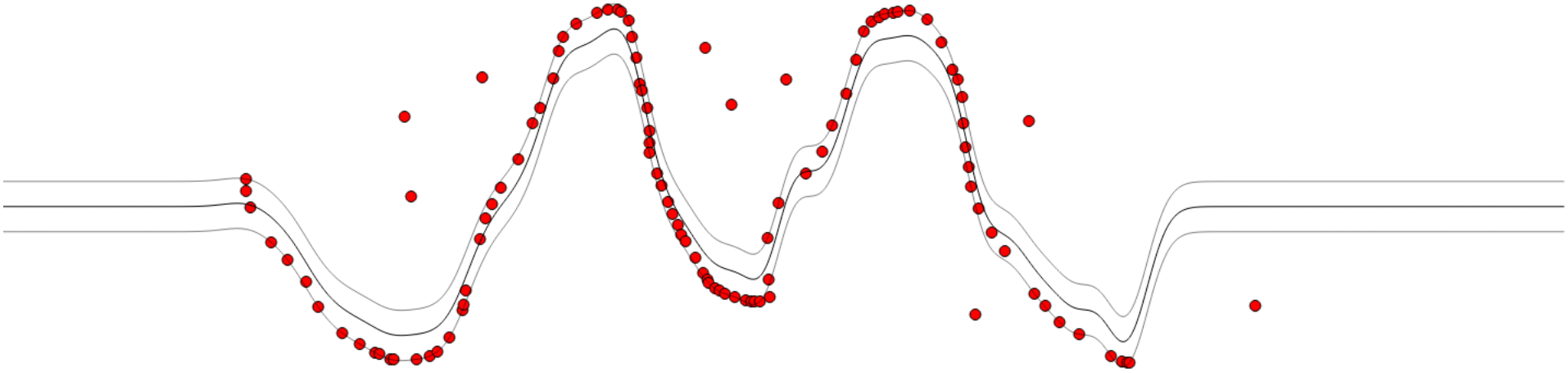


Support Vector Regression: ν-SVR



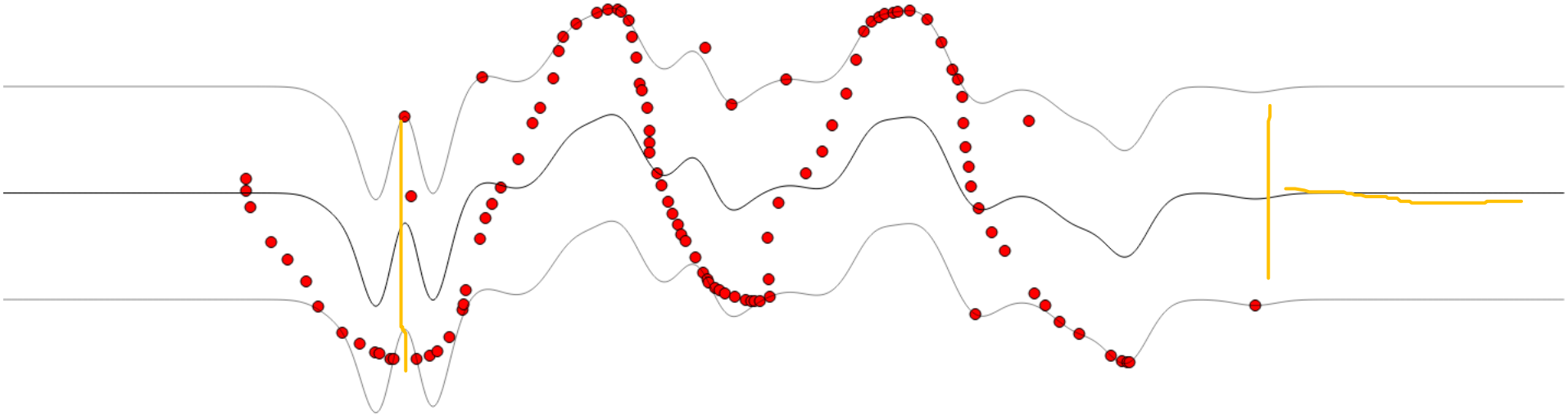
Effect of the automatic adaptation of ε using ν-SVR

Support Vector Regression: ν-SVR



Effect of the automatic adaptation of ε using ν-SVR

Support Vector Regression: ν-SVR



Effect of the automatic adaptation of ε using ν-SVR

Relevance Vector Regression (RVR)

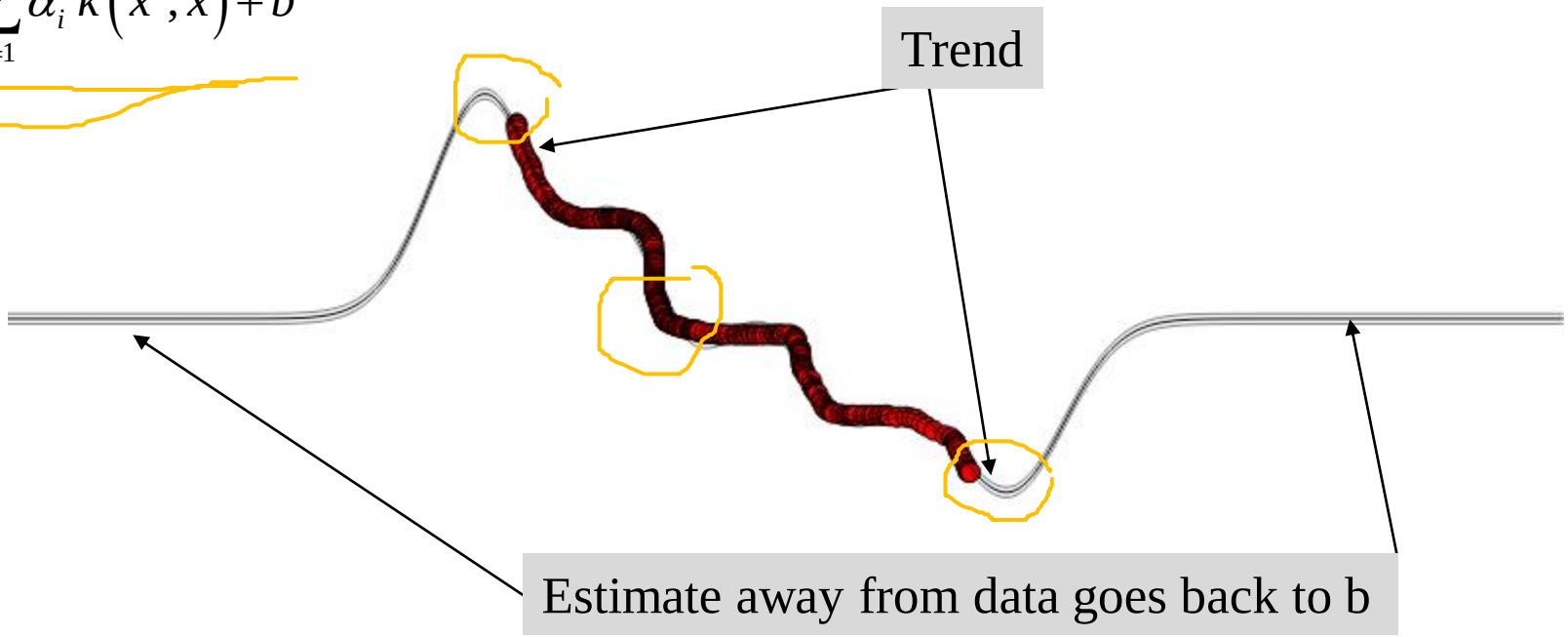
Relevance Vector Regression (RVR) is the regression version of RVM.

- It provides a sparser version of SVR.
- It offers a probabilistic interpretation of the solution.
(see Tipping 2011, supplementary material to the class)

ε-SVR - prediction

It is a deterministic model. It does not entail a notion of likelihood.

$$y = \sum_{i=1}^{M} \alpha_i \, k\left(x^i, x\right) + b$$



Relevance Vector Regression (RVR)

Same principle as that described for RVM (see slides on SVM and extensions). The derivation of the parameters however differ (see Tipping 2011 for details).

To recall, we start from the solution of SVM.

$$y(x) = f(x) = \sum_{i=1}^{M} \alpha_i \underbrace{k(x, x^i)}_{\psi_i(x)} + b$$

Rewrite the solution of SVM as a linear combination over M basis functions

A sparse solution has a majority of entries with alpha zero.

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_M \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ 0 \\ \alpha_3 \\ 0 \\ 0 \\ . \\ \alpha_M \end{bmatrix}$$

In the (binary) classification case, $y \in [0;1]$.

In the regression case, $y \in \mathbb{R}$.

Relevance Vector Regression (RVR)

Rewrite the solution of SVR in a compact form such that the problem is linear in the parameters:

$$y(x) = f(x) = \sum_{i=1}^{M} \alpha_i \underbrace{k(x, x^i)}_{\psi_i(x)} + \underbrace{b}_{\alpha_0 \psi_o(x)}$$

$$y(x) = \alpha^T \Psi(x), \qquad \Psi(x) = \left[\psi_0(x)\ \psi_1(x) \ldots\ldots \psi_M(x)\right]^T, \ \psi_0(x) = 1.$$

Open parameters α.

Relevance Vector Regression (RVR)

Take a Bayesian approach and assume that all samples y_i are i.i.d and that they are measurements of the real value $\alpha^T \Psi\left(x^i\right)$ subjected to white noise ε, i.e.:

$$y_i(x) = \alpha^T \Psi\left(x^i\right) + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma_\varepsilon\right)$$

Likelihood of the model under the dataset $\{X, \mathbf{y}\}$ composed of M pairs of data points: $\{X, \mathbf{y}\} = \left\{x^i, y^i\right\}_{i=1}^{M}$

$$\mathbf{y} = \alpha^T X + N\left(0, \sigma_\varepsilon\right)$$

$$\Rightarrow \mathbf{y} \sim p\left(\mathbf{y} \mid X, \alpha, \sigma_\varepsilon\right)$$

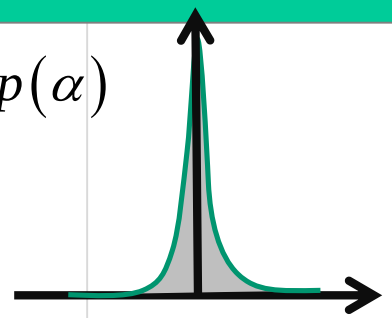
Relevance Vector Regression (RVR)

Estimates the posterior distribution on α, given the data using Bayes' rule: $p(\alpha, \sigma_\varepsilon \mid \mathbf{y}, X)$

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{marginal likelihood}}$$

$$\Rightarrow p(\alpha, \sigma_\varepsilon \mid \mathbf{y}, X) = \frac{p(\mathbf{y} \mid X, \alpha, \sigma_\varepsilon)\, p(\alpha, \sigma_\varepsilon)}{p(\mathbf{y} \mid X)}$$

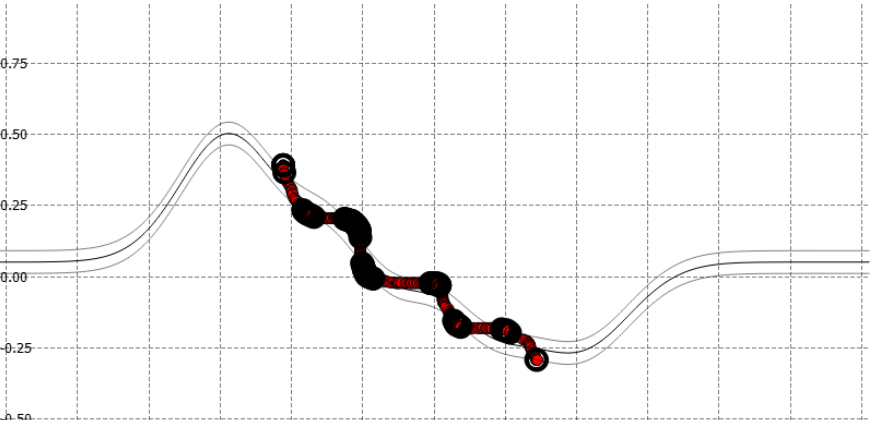
Sparsity is obtained when the distribution is sharply peaked at zero.



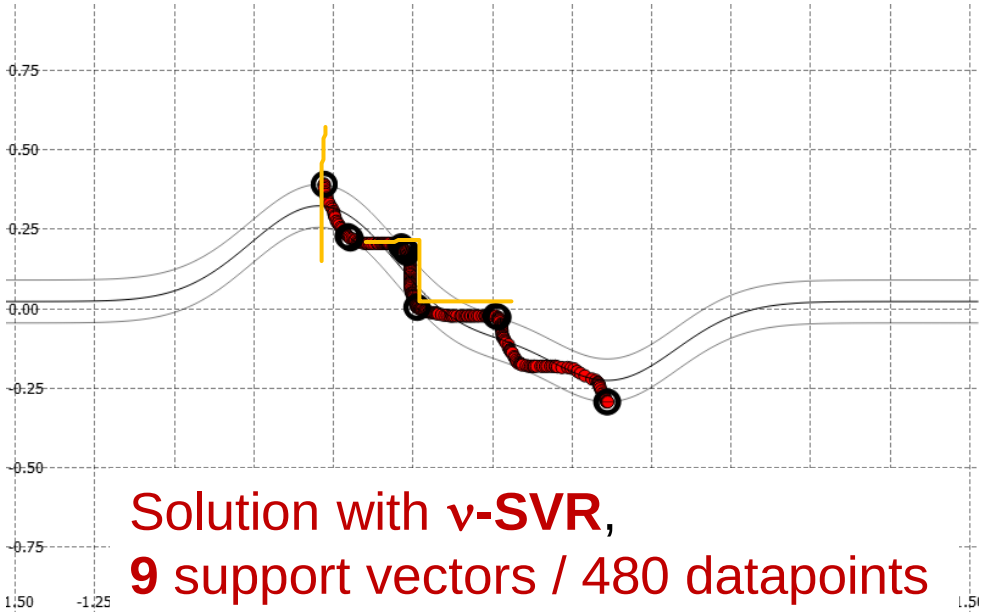
Unlike Ridge Regression Process: No closed-form solution, but iterative procedure.

Unlike SVR, not convex optimization any more!

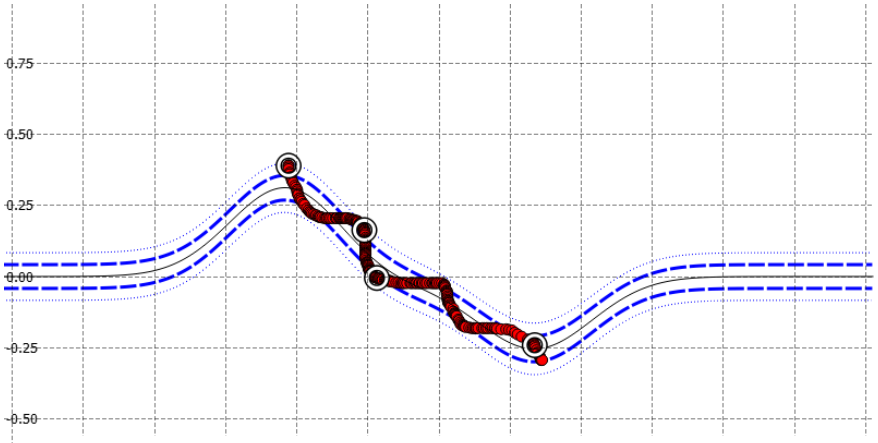
Comparison ε-SVR, ν-SVR, RVR: RBF kernel



Solution with **ε-SVR**,
87 support vectors / 480 datapoints

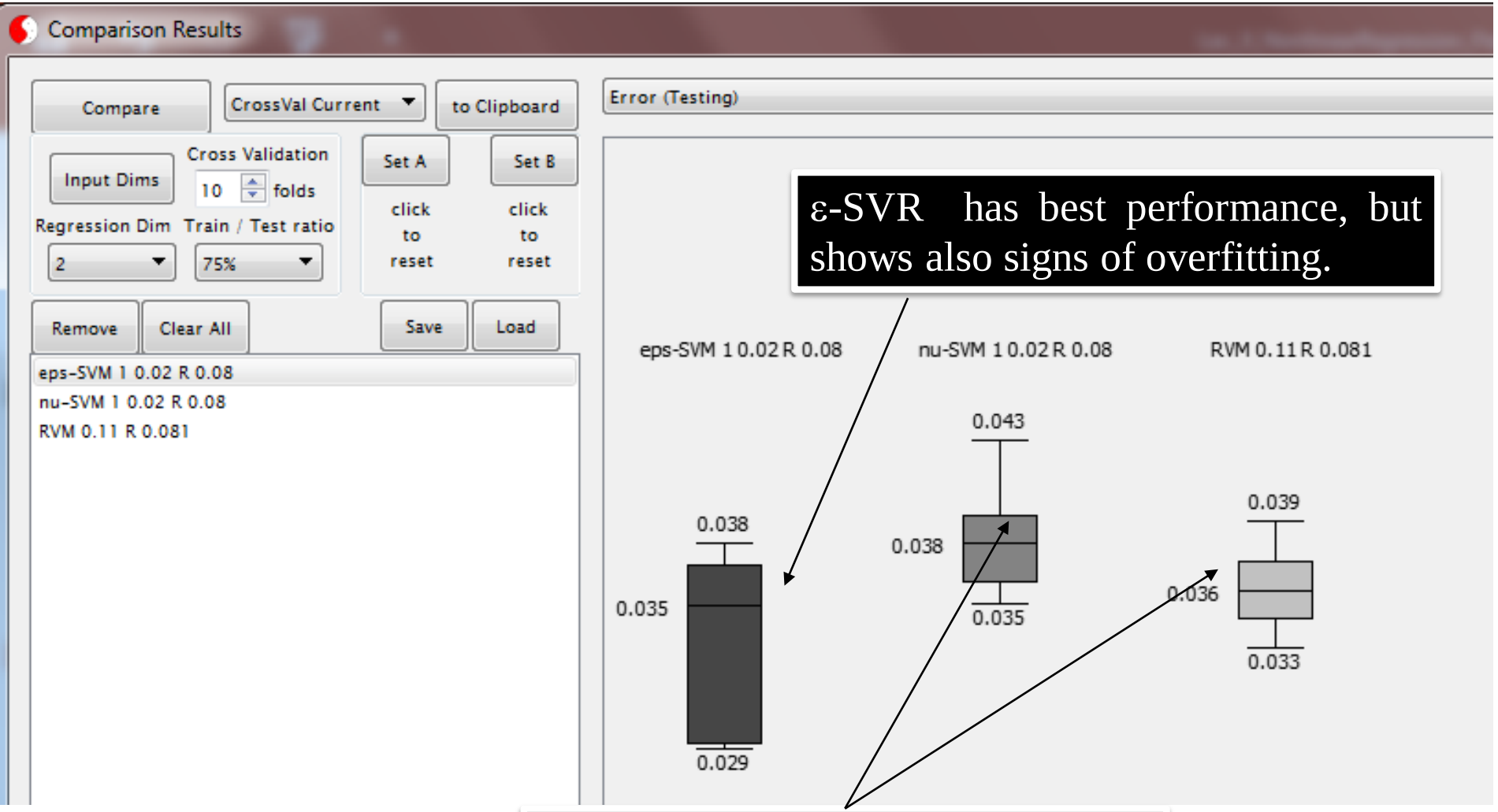


Solution with **ν-SVR**,
9 support vectors / 480 datapoints



Solution with **RVR**,
3 support vectors / 480 datapoints

Comparison ε-SVR, ν-SVR, RVR: RBF kernel



RVR has the least number of SVs and better performance than ν-SVR.

Limitations and extensions of ε-SVR

Main limitations:

- ❖ Difficult to determine hyperparameters
- ❖ Lack of constraints to limit number of support vectors
- ❖ Lack of probabilistic prediction

Extensions:

- ❑ ν-SVR:
  - ○ Removes choice of noise model.
  - ○ Reduces number of support vectors.

- ❖ RVR:
  - ○ Reduces number of support vectors.
  - ○ Measure of the goodness of the fit through likelihood**.**