On the validity of minimin and minimax methods for Support Vector Regression with interval data

Andrea Wiencierz
Department of Mathematics
University of York

Marco Cattaneo
Department of Mathematics
University of Hull

ISIPTA '15, Pescara, Italy
22 July 2015

Support Vector Regression (SVR) with precise data

data: $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} \overset{\text{compact}}{\subset} \mathbb{R}^d \times \mathbb{R}$

Reproducing Kernel Hilbert Space: set $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$, e.g., with the Gaussian kernel κ_σ defined for all $x, x' \in \mathcal{X}$ and $\sigma > 0$ by

$$\kappa_\sigma(x, x') = \exp\left(-\tfrac{1}{\sigma^2} \|x - x'\|^2\right),$$

$\mathcal{F}$ is dense in the space $\mathcal{C}(\mathcal{X})$ of continuous functions

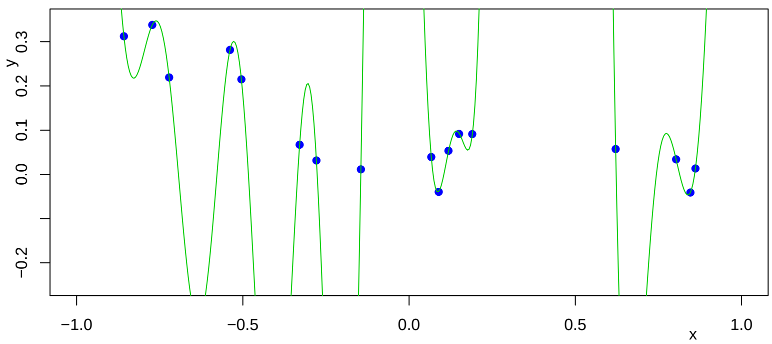
regression function: find the function $f \in \mathcal{F}$ that best describes the relationship between the variables of interest in the light of the data

general idea: function $f \in \mathcal{F}$ minimizing the (empirical) risk

$$\mathcal{E}(f) = \frac{1}{n} \sum_{i=1}^{n} \psi\left(|y_i - f(x_i)|\right),$$

where $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is convex with $\psi(0) = 0$, e.g., for the linear loss ψ is defined by $\psi(r) = r$ for all $r \in \mathbb{R}_{\geq 0}$

$\rightsquigarrow$ estimated functions are too wiggly when considering large $\mathcal{F}$

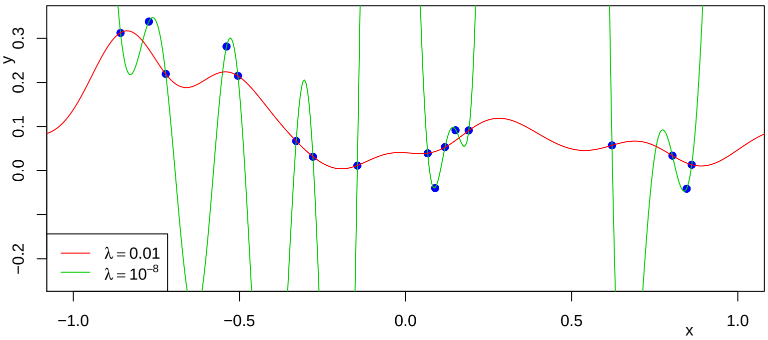


Unpenalized regression function based on Gaussian kernel and linear loss with precise data $(x_i, y_i) \in \mathbb{R}^2$ where $i \in \{1, \dots, 17\}$

SVR estimate: function $f \in \mathcal{F}$ minimizing the regularized risk

$$\mathcal{E}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \psi\left(|y_i - f(x_i)|\right) + \lambda \|f\|_{\mathcal{F}}^2,$$

where $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is convex with $\psi(0) = 0$, and $\lambda \in \mathbb{R}_{>0}$



Unpenalized regression function vs. SVR estimate, both based on Gaussian kernel and linear loss

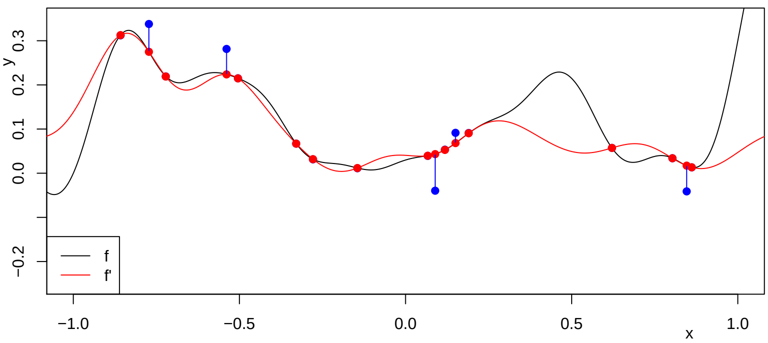
Representer Theorem (RT): the regression function minimizing $\mathcal{E}_\lambda(f)$ exists, is unique, and has the form

$$f = \sum_{j=1}^{n} \alpha_j \, \kappa(\,\cdot\,, x_j),$$

where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and κ is the kernel function associated with $\mathcal{F}$

key result underlying the SVR methodology: the minimization of $\mathcal{E}_\lambda(f)$ becomes a convex optimization task in n variables $\alpha_1, \ldots, \alpha_n$, i.e., the RT makes the theoretical idea practically feasible

core of the proof (see e.g., Steinwart & Christmann (2008)): the structure of $\mathcal{F}$ implies that for each f, the orthogonal projection $f' = \sum_{j=1}^{n} \alpha_j' \, \kappa(\,\cdot\,, x_j)$ of f on the subspace spanned by the functions $\kappa(\,\cdot\,, x_j)$ satisfies $f'(x_i) = f(x_i)$ for all $i \in \{1, \ldots, n\}$, and therefore $\mathcal{E}_\lambda(f') \leq \mathcal{E}_\lambda(f)$



SVR estimate f' based on Gaussian kernel and linear loss vs. another $f \in \mathcal{F}$ with $f(x_i) = f'(x_i)$ for all $i \in \{1, \ldots, 17\}$

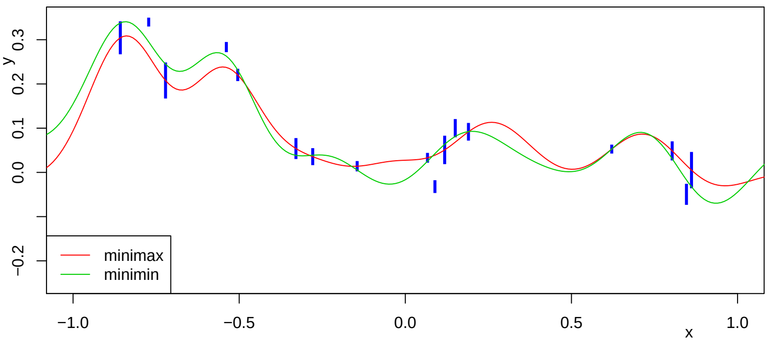
minimin and minimax methods for SVR with interval-valued response

interval data: instead of the values $y_1, \dots, y_n$, only intervals $[\underline{y}_1, \overline{y}_1], \dots, [\underline{y}_n, \overline{y}_n]$ are observed, with $y_i \in [\underline{y}_i, \overline{y}_i]$ for all $i \in \{1, \dots, n\}$

minimin and minimax SVR estimates (Utkin & Coolen (2011)): $f \in \mathcal{F}$ minimizing

$$\underline{\mathcal{E}}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \min_{y_i \in [\underline{y}_i, \overline{y}_i]} \psi\left(|y_i - f(x_i)|\right) + \lambda \|f\|_{\mathcal{F}}^2 \quad \text{and}$$

$$\overline{\mathcal{E}}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \max_{y_i \in [\underline{y}_i, \overline{y}_i]} \psi\left(|y_i - f(x_i)|\right) + \lambda \|f\|_{\mathcal{F}}^2$$



minimin SVR estimate vs. minimax SVR estimate, both based on Gaussian kernel and linear loss

RT for minimin and minimax SVR

Lemma 1. *The regularized lower and upper risk functionals, $\underline{\mathcal{E}}_\lambda$ and $\overline{\mathcal{E}}_\lambda$, respectively have unique minimizers $f_\lambda^{\text{minimin}}$ and $f_\lambda^{\text{minimax}}$ in $\mathcal{F}$, respectively.*

Proof. The proof can be found in the paper. □

Theorem 1. *There are $\alpha_1^{\text{minimin}}, \dots, \alpha_n^{\text{minimin}} \in \mathbb{R}$ and $\alpha_1^{\text{minimax}}, \dots, \alpha_n^{\text{minimax}} \in \mathbb{R}$ such that*

$$f_\lambda^{\text{minimin}} : x \mapsto \sum_{i=1}^{n} \alpha_i^{\text{minimin}} \kappa(x, x_i) \quad \textit{and}$$

$$f_\lambda^{\text{minimax}} : x \mapsto \sum_{i=1}^{n} \alpha_i^{\text{minimax}} \kappa(x, x_i)$$

are the unique minimizers of $\underline{\mathcal{E}}_\lambda$ and $\overline{\mathcal{E}}_\lambda$ in $\mathcal{F}$, respectively.

Proof. Let f' denote the orthogonal projection of a function $f \in \mathcal{F}$ on the subspace $\mathcal{F}_n$ spanned by the functions $\kappa(\cdot, x_i)$ with $i \in \{1, \dots, n\}$. Then $\|f'\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}}$, and f' is of the form $\sum_{i=1}^{n} \alpha_i \, \kappa(\cdot, x_i)$ with $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Moreover, for each $i \in \{1, \dots, n\}$, the orthogonality of $f' - f$ and $\kappa(\cdot, x_i)$ implies $f'(x_i) = f(x_i)$, because

$$f'(x_i) - f(x_i) = \langle f' - f, \, \kappa(\cdot, x_i) \rangle_{\mathcal{F}} = 0.$$

Therefore, $\underline{\mathcal{E}}_\lambda(f') \leq \underline{\mathcal{E}}_\lambda(f)$ and $\overline{\mathcal{E}}_\lambda(f') \leq \overline{\mathcal{E}}_\lambda(f)$, and the desired result is implied by Lemma 1. □

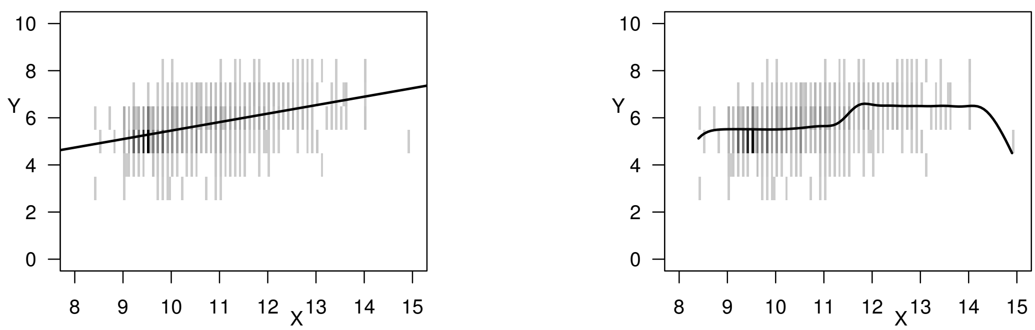
SVR analysis of wine quality

wine data: we analyze the red wine sample ($n = 1\,599$) of the Vinho Verde wine data set initially analyzed by Cortez et al. (2009), which is freely available from the UC Irvine Machine Learning Repository (`http://archive.ics.uci.edu/ml/`)

relationship of interest: between alcohol content (explanatory variable) and sensory quality (interval-valued response) of a red wine

results: both minimax SVR analyses suggest an increasing relationship

- SVR with linear kernel and Least Squares (LS) loss (a.k.a. Ridge regression)
- SVR with Gaussian kernel and linear loss



minimax SVR estimates based on linear kernel and LS loss (left), i.e., $\kappa(x, x') = \langle x, x' \rangle + 1$ for all $x, x' \in \mathcal{X}$ and $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, and based on Gaussian kernel and linear loss (right)

conclusions

main contribution of the paper: generalization of the RT to the case with interval data $[\underline{y}_1, \overline{y}_1], \dots, [\underline{y}_n, \overline{y}_n] \subset \mathbb{R}$, justifying minimin and minimax SVR in this case

no further generalization: the RT for interval-valued response cannot be directly generalized to the case with interval data $[\underline{x}_1, \overline{x}_1], \dots, [\underline{x}_n, \overline{x}_n] \subset \mathbb{R}^d$, in which the following expressions have to be minimized

$$\underline{\mathcal{E}}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \min_{x_i \in [\underline{x}_i, \overline{x}_i]} \psi\left(|y_i - f(x_i)|\right) + \lambda \, \|f\|_{\mathcal{F}}^2 \quad \text{and}$$

$$\overline{\mathcal{E}}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \max_{x_i \in [\underline{x}_i, \overline{x}_i]} \psi\left(|y_i - f(x_i)|\right) + \lambda \, \|f\|_{\mathcal{F}}^2$$

- a regression function minimizing $\underline{\mathcal{E}}_\lambda(f)$ would have the form $f = \sum_{j=1}^{n} \alpha_j \, \kappa(\,\cdot\,, x_j)$, where $\alpha_j \in \mathbb{R}$ and $x_j \in [\underline{x}_j, \overline{x}_j]$ for all $j \in \{1, \dots, n\}$, but in general $\underline{\mathcal{E}}_\lambda$ is not convex
- by contrast, $\overline{\mathcal{E}}_\lambda$ is convex, but a regression function minimizing $\overline{\mathcal{E}}_\lambda(f)$ does not necessarily have the form $f = \sum_{j=1}^{n} \alpha_j \, \kappa(\,\cdot\,, x_j)$, where $\alpha_j \in \mathbb{R}$ and $x_j \in [\underline{x}_j, \overline{x}_j]$ for all $j \in \{1, \dots, n\}$

the even more general case with interval data $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i] \subset \mathbb{R}^d \times \mathbb{R}$ for all $i \in \{1, \dots, n\}$ also presents the above difficulties

references

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 547–553.

Steinwart, I., and Christmann, A. (2008). *Support Vector Machines*. Springer.

Utkin, L., and Coolen, F. (2011). Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA '11 Proceedings*, eds. F. Coolen, G. de Cooman, T. Fetz, and M. Oberguggenberger. SIPTA, 371–380.