View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repository@Hull - CRIS

On the Validity of Minimin and Minimax Methods for Support Vector Regression with Interval Data

Andrea Wiencierz
Department of Mathematics
University of York
andrea.wiencierz@york.ac.uk

Marco E. G. V. Cattaneo
Department of Mathematics
University of Hull
m.cattaneo@hull.ac.uk

Abstract

In the recent years, generalizations of support vector methods for analyzing interval-valued data have been suggested in both the regression and classification contexts. Standard Support Vector methods for precise data formalize these statistical problems as optimization problems that can be based on various loss functions. In the case of Support Vector Regression (SVR), on which we focus here, the function that best describes the relationship between a response and some explanatory variables is derived as the solution of the minimization problem associated with the expectation of some function of the residual, which is called the risk functional. The key idea of SVR is that even when considering an infinite-dimensional space of arbitrary regression functions, given a finite-dimensional data set, the function minimizing the risk can be represented as the finite weighted sum of kernel functions. This allows to practically determine the SVR estimate by solving a much simpler optimization problem, even in the case of nonlinear regression. In case that only interval-valued observations of the variables of interest are available, it has been suggested to minimize the minimal or maximal risk values that are compatible with the imprecise data, yielding precise SVR estimates on the basis of interval data. In this paper, we show that also in the case of an interval-valued response the optimal function can be represented as the finite weighted sum of kernel functions. Thus, the minimin and minimax SVR estimates can be obtained by minimizing the corresponding simplified expressions of the empirical lower and upper risks, respectively.

Keywords. Support Vector Regression, interval data, Representer Theorem.

1 Introduction

In this paper, we deal with the generalization of Support Vector Regression (SVR) to interval data. By SVR we denote a class of kernel-based methods for the statistical problem of regression analysis. These methods originated in the field of Machine Learning (Vapnik, 1998, 1995) and recently also gained attention in the field of Statistics (see, e.g., Hable, 2012; Christmann et al., 2009; Hofmann et al., 2008; Steinwart and Christmann, 2008). The typical goal of a regression analysis is to describe the relationship between a response variable $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and a number $d \in \mathbb{N}$ of explanatory variables $X \in \mathcal{X} \subseteq \mathbb{R}^d$ by a function $f : \mathcal{X} \to \mathbb{R}$. The sought-after function f is usually assumed to be a member of a particular space $\mathcal{F}$ of considered regression functions, for example, the space of all (affine) linear functions.

To identify which functions in $\mathcal{F}$ best describe the relationship between the random variables in $(X, Y) = V$, the considered regression functions are assessed by a loss function. Most common loss functions are characteristics of the distribution of (some function of) the residual R_f, which we here define by

$$R_f = |Y - f(X)|$$

for each $f \in \mathcal{F}$. In the SVR methodology, the expectation of some usually convex error function is considered as loss function, which is called risk functional. If the probability distribution P_V of the random vector V is known, the distribution of R_f can be derived from it and the best regression functions can be identified by minimizing the chosen loss function. Yet, usually the true distribution of the investigated variables is unknown, but it is assumed that P_V lies in some specific set of probability measures $\mathcal{P}_V$. Thus, the evaluation of each regression function also varies over possible distributions of V.

Given the realization of an independent sample of random variables $V_1 = (x_1, y_1), \ldots, V_n = (x_n, y_n)$, with $n \in \mathbb{N}$, where $V_i \sim P_V$ for all $i \in \{1, \ldots, n\}$, we can learn something about the distribution of the variables of interest. In SVR, the empirical distribution $\hat{P}_V$ of the observations is used as a point estimate of P_V and the (regularized) risk under this particular distribu-

tion is minimized to obtain the regression estimate. The SVR estimate is in general unique. Moreover, the so-called Representer Theorem states that the function minimizing the risk given the observations can be represented as the finite weighted sum of kernel functions. This is a key result for SVR, as it allows to practically determine the SVR estimate by solving a relatively simple optimization problem, even in the case of nonlinear regression. Further details of the SVR methodology are presented in the next section.

If the variables of interest are not observed as precise numbers but only upper and lower bounds to the values are available, the empirical distribution $\hat{P}_V$ is not revealed by the observable data. We denote the random sets describing the observables by $V_1^*, \ldots, V_n^*$ and their probability distribution by P_{V^*}. If the observed intervals are assumed to cover the unknown precise values with probability one, bounds for the empirical risk can be derived from the empirical distribution $\hat{P}_{V^*}$ of the imprecise data. How can we use this information to obtain an SVR estimate in this situation? Starting from the simplified representation of the optimal function in standard SVR, Utkin and Coolen (2011) proposed to follow a minimin or a minimax approach and to minimize either the lower or the upper (regularized) risk in order to obtain a precise regression estimate.

In this paper, we investigate the validity of their starting from the simplified representation in the generalized data situation. At first, we introduce the formal framework of the SVR methodology in detail and formally discuss Utkin and Coolen (2011)'s SVR generalization. Then, we consider the Representer Theorem in the more general data situation. We find that also in this case the optimal function can be represented as the finite weighted sum of kernel functions. Finally, after applying the discussed SVR methods to an interesting problem in the area of winemaking, a short outlook concludes the paper.

2 Methodological Framework of SVR

In this section, the formal framework of SVR with precise data is presented. In the SVR methodology, the set $\mathcal{P}_V$ is assumed to contain all probability measures on $\mathcal{V} = \mathcal{X} \times \mathcal{Y}$. In this paper, we additionally assume that $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. Furthermore, in SVR, the loss assigned to a possible regression function f and a distribution P_V is the risk $\mathcal{E}_{P_V}(f)$. Presupposing measurability, the risk functional $\mathcal{E}_{P_V}$ on $\mathcal{F}$ can be defined for each $P_V \in \mathcal{P}_V$ as

$$\mathcal{E}_{P_V} \,:\, f \mapsto \mathcal{E}_{P_V}(f) = \mathbb{E}_{P_V}(\psi(R_f)), \qquad (1)$$

where ψ is a convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ satisfying $\psi(0) = 0$ and $\mathbb{E}_{P_V}$ denotes the expectation with respect to P_V. For example, if ψ is defined by $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, the loss associated with a pair (f, P_V) is given by $\mathcal{E}_{P_V}(f) = \mathbb{E}_{P_V}(R_f^2)$. Thus, we obtain the loss function corresponding to Least Squares regression. Another famous example is the function defined by $\psi(r) = \max\{0, r - \nu\}$, for all $r \in \mathbb{R}_{\geq 0}$ and some $\nu \geq 0$, which was introduced by Vapnik (1995, Section 6.1) and represents the so-called ν-insensitive loss.

The convexity of the mapping ψ implies convexity of the risk functional $\mathcal{E}_{P_V}$, that is, the risk functional satisfies for each $\rho \in [0, 1]$

$$\mathcal{E}_{P_V}\left(\rho\, f + (1-\rho)\, f'\right) \leq \rho\, \mathcal{E}_{P_V}(f) + (1-\rho)\, \mathcal{E}_{P_V}(f'),$$

for all $f, f' \in \mathcal{F}$ (see also Steinwart and Christmann, 2008, Lemma 2.13). As explained later, this property is crucial to the existence of a unique optimal regression function.

In the SVR framework, the space $\mathcal{F}$ of considered regression functions from $\mathcal{X}$ to $\mathbb{R}$ is supposed to be a Reproducing Kernel Hilbert Space (RKHS) with associated scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$. An RKHS is uniquely associated with its reproducing kernel function. A kernel function κ is a positive semi-definite function on $\mathcal{X} \times \mathcal{X}$, that is, $\sum_{i=1}^n \sum_{j=1}^n \alpha_i\, \alpha_j\, \kappa(x_i, x_j) \geq 0$, for all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $n \in \mathbb{N}$. Here, we only consider kernel functions that are moreover measurable and bounded. If κ is the reproducing kernel function of the RKHS $\mathcal{F}$, for each $x \in \mathcal{X}$ we have $\kappa(\cdot, x) \in \mathcal{F}$ and

$$f(x) = \langle f, \kappa(\cdot, x) \rangle_{\mathcal{F}},$$

for all $f \in \mathcal{F}$. From this property called reproducing property follows that $\kappa(x, x') = \langle \kappa(\cdot, x), \kappa(\cdot, x') \rangle_{\mathcal{F}}$, for all $x, x' \in \mathcal{X}$. A simple example for an RKHS and its reproducing kernel is the function space associated with the linear kernel defined by $\kappa(x, x') = \langle x, x' \rangle + 1$, for all $x, x' \in \mathcal{X}$, which is the Hilbert space of all (affine) linear functions from $\mathcal{X}$ to $\mathbb{R}$. Another common kernel function is the so-called Gaussian kernel, which is defined for all $x, x' \in \mathcal{X}$ by

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

with $\sigma > 0$. The associated RKHS is a very large function space that is dense in the space of all continuous (real-valued) functions on $\mathcal{X}$. For more details on kernels and RKHSs, see, for example, Steinwart and Christmann (2008, Chapter 4).

To avoid obtaining too wiggly functions as descriptions of the relationship of interest when the regression analysis is based on a finite sample of observations,

the risk is further supplemented by an additive penalty for the complexity of the functions $f \in \mathcal{F}$. Hence, in the SVR methodology, instead of $\mathcal{E}_{P_V}$ the regularized risk functional $\mathcal{E}_{P_V,\lambda}$ is minimized, which is defined for all $f \in \mathcal{F}$ by

$$\mathcal{E}_{P_V,\lambda}(f) = \mathcal{E}_{P_V}(f) + \lambda \|f\|_{\mathcal{F}}^2,$$

where $\lambda > 0$ is a fixed parameter regulating the penalization and $\|\cdot\|_{\mathcal{F}}$ is the norm induced by the scalar product in $\mathcal{F}$. The regularization can be interpreted as minimizing $\mathcal{E}_{P_V}$ under the restriction $\|f\|_{\mathcal{F}}^2 \leq c$, for some $c \in \mathbb{R}_{\geq 0}$, but instead of choosing the bound c explicitly, we fix the value of the corresponding Lagrange multiplier λ in the constrained optimization problem.

As the functional $f \mapsto \lambda \|f\|_{\mathcal{F}}^2$ is strictly convex by general properties of norms and $\mathcal{E}_{P_V}$ is convex because of ψ, we have that $\mathcal{E}_{P_V,\lambda}$ is also a strictly convex functional on $\mathcal{F}$. Exploiting the strict convexity of $\mathcal{E}_{P_V,\lambda}$, it can be shown that an optimal function always exists and is unique, provided that some regularity conditions are fulfilled (see, e.g., Steinwart and Christmann, 2008, Lemma 5.1 and Theorem 5.2).

Given observations $(x_1, y_1), \ldots, (x_n, y_n)$ of an independent and identically distributed random sample $V_1, \ldots, V_n$, the SVR methodology consists in estimating P_V by the corresponding empirical distribution $\hat{P}_V$, before identifying the regression estimate $f_{\hat{P}_V,\lambda} \in \mathcal{F}$ by the minimization of $\mathcal{E}_{\hat{P}_V,\lambda}$, for some $\lambda > 0$. Like in the general case, there always exists a unique minimizer of the regularized risk for $\hat{P}_V$. Moreover, the so-called Representer Theorem states that this unique function $f_{\hat{P}_V,\lambda}$ can be represented as the linear combination of the corresponding functions $\kappa(\cdot, x_1), \ldots, \kappa(\cdot, x_n)$, that is, there exist weights $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that

$$f_{\hat{P}_V,\lambda}(x) = \sum_{j=1}^{n} \alpha_j \, \kappa(x, x_j), \tag{2}$$

for all $x \in \mathcal{X}$ (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5). This expression is sometimes called support vector expansion of $f_{\hat{P}_V,\lambda}$ and the optimal function $f_{\hat{P}_V,\lambda}$ is often referred to as a Support Vector Machine (SVM). This term can be explained historically, because Vapnik (1998, 1995) proposed to use functions for ψ that have the property that some of the resulting $\alpha_1, \ldots, \alpha_n$ are zero. The vectors x_j for which $\alpha_j \neq 0$ are called support vectors, whence the notion SVM. One example for such a representing function ψ is the function associated with the ν-insensitive loss mentioned before. Nevertheless, in general, SVMs are not sparse in this sense (see, e.g., Steinwart and Christmann, 2008, Section 11.1).

The result of the Representer Theorem expressed in (2) is extremely useful for the practical computation of SVR estimates as it simplifies the associated optimization problems and allows to solve them even when large RKHSs of arbitrary smooth regression functions are considered, like, for example, the RKHS associated with the Gaussian kernel. Given a data set $(x_1, y_1), \ldots, (x_n, y_n)$ with empirical distribution $\hat{P}_V$ and a fixed $\lambda > 0$, Equation (2) tells us that $f_{\hat{P}_V,\lambda}$ is an element of the set $\mathcal{F}_n \subset \mathcal{F}$, with

$$\mathcal{F}_n = \left\{ \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, x_j) \,:\, \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\}.$$

Furthermore, for all functions $f_\alpha = \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, x_j)$, with $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}} \in \mathbb{R}^n$, the squared norm is given by $\|f_\alpha\|_{\mathcal{F}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j)$. Hence, the regularized risk associated with $\hat{P}_V$ can be written for each $f_\alpha \in \mathcal{F}_n$ as

$$\begin{aligned} \mathcal{E}_{\hat{P}_V,\lambda}(f_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \psi\big(\big|y_i - \textstyle\sum_{j=1}^{n} \alpha_j \, \kappa(x_i, x_j)\big|\big) \\ + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j). \end{aligned}$$

As $\mathcal{E}_{\hat{P}_V,\lambda}$ is convex, the SVM $f_{\hat{P}_V,\lambda}$ can be obtained by solving a convex optimization problem over $\alpha \in \mathbb{R}^n$, for which there are numerous efficient algorithms (see, e.g., Boyd and Vandenberghe, 2004). For the selection of an appropriate regularization parameter $\lambda > 0$ and of other hyper-parameters like the parameter σ of the Gaussian kernel, different strategies can be applied, for instance, cross-validation (see, e.g., Steinwart and Christmann, 2008, Section 11.3). Since we are mainly interested in the generalization of a key theoretical result about SVR to the situation with interval data, we neglect the latter issues in this paper and always consider these parameters fixed.

3 SVR with Interval Data

In this section, we investigate whether the SVR methodology can be used for regression analysis when the variables of interest cannot be observed as precise numbers but only (bounded) intervals covering the values of interest are available. Utkin and Coolen (2011) proposed a generalization of the SVR methodology to this situation. As we will see later, the suggested methods of Utkin and Coolen (2011) work well for interval-valued observations of the response variable Y, but cannot directly be extended to interval-valued observations of the variables in X. Therefore, we also consider here only the situation where instead of V the random set $V^* \in \mathcal{V}^* \subseteq 2^{\mathcal{V}}$ is observed, whose possible realizations are of the form $\{X\} \times [\underline{Y}, \overline{Y}]$, with $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $\underline{Y}, \overline{Y} \in \mathcal{Y} \subset \mathbb{R}$ such that $\underline{Y} \leq \overline{Y}$.

3.1 Utkin and Coolen (2011)'s SVR Generalization

Now, we discuss the generalization of SVR proposed by Utkin and Coolen (2011) in detail. Since in the considered data situation the precise variables are not observable, it is impossible to evaluate the considered regression functions $f \in \mathcal{F}$ by $\mathcal{E}_{\hat{P}_V}(f)$, i.e., by the risk associated with the empirical distribution of the precise data. However, the probability distribution of the imprecise data P_{V^*} can be estimated on the basis of the observations.

When the probability distribution P_{V^*} of the observable data is known, as we assume that the interval $[\underline{Y}, \overline{Y}]$ covers the precise unobservable Y with probability one, we know that the unknown probability distribution of the precise data lies in the set $[P_{V^*}] \subseteq \mathcal{P}_V$ containing all distributions of the precise data, P_V, that satisfy for all measurable events $A \subseteq \mathcal{V}$ the inequalities

$$\begin{aligned} &P_V(V \in A) \geq P_{V^*}(V^* \subseteq A) \quad \text{and} \\ &P_V(V \in A) \leq P_{V^*}(V^* \cap A \neq \varnothing). \end{aligned} \tag{3}$$

By consequence, for all $f \in \mathcal{F}$, the unknown risk $\mathcal{E}_{P_V}(f)$ lies in the interval $[\underline{\mathcal{E}}_{P_{V^*}}(f), \overline{\mathcal{E}}_{P_{V^*}}(f)]$, where

$$\begin{aligned} \underline{\mathcal{E}}_{P_{V^*}}(f) &= \min_{P'_V \in [P_{V^*}]} \mathcal{E}_{P'_V}(f) \quad \text{and} \\ \overline{\mathcal{E}}_{P_{V^*}}(f) &= \max_{P'_V \in [P_{V^*}]} \mathcal{E}_{P'_V}(f). \end{aligned}$$

Hence, in the regression problem with interval-valued response, the set $[\underline{\mathcal{E}}_{P_{V^*}}(f), \overline{\mathcal{E}}_{P_{V^*}}(f)]$ of all possible risk values constitutes the loss evaluation for each $f \in \mathcal{F}$. Of course, it is in general impossible to directly determine an optimal function with respect to this imprecise criterion. The central idea of the regression methodology proposed by Utkin and Coolen (2011) is to use the minimin or the minimax rule to solve this problem, that is, to minimize either the lower risk $\underline{\mathcal{E}}_{P_{V^*}}$ or the upper risk $\overline{\mathcal{E}}_{P_{V^*}}$ in order to identify a single optimal regression function.

To derive expressions of the lower and upper risks, Utkin and Coolen (2011) describe, for each regression function $f \in \mathcal{F}$, the set of compatible probability distributions of the residual R_f given P_{V^*} by a so-called p-box and apply results from Utkin and Destercke (2009). Introduced by Ferson et al. (2003, Section 2), the notion p-box designates a convex set of probability measures for a univariate random quantity that is bounded by a lower and an upper cumulative distribution function. In the situation considered here, given P_{V^*}, also the marginal distribution of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, where

$$\begin{aligned} \underline{R}_f &= \min_{(x,y) \in V^*} |y - f(x)| \quad \text{and} \\ \overline{R}_f &= \max_{(x,y) \in V^*} |y - f(x)|, \end{aligned}$$

is known for each $f \in \mathcal{F}$. According to (3), the marginal distribution of the imprecise residual implies lower and upper bounds to the probabilities of all measurable events associated with the marginal distribution of the precise residual R_f. If we consider these lower and upper bounds for all events of the form $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, we obtain a lower and an upper cumulative distribution function that constitute a p-box. As the p-box covers all probability distributions of R_f that comply with the bounds at least for the intervals $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, some of the probability measures included in the p-box may not satisfy (3) for all measurable events, and thus, may be incompatible with the marginal distribution of the imprecise residual. However, the p-box obtained in the described way from the random set $[\underline{R}_f, \overline{R}_f]$, with $f \in \mathcal{F}$, is the tightest outer approximation by a p-box of the set of probability distributions of R_f implied by this random set (see, e.g., Destercke et al., 2008). In fact, in the present situation, for each $f \in \mathcal{F}$, the upper bound of the associated p-box corresponds to the cumulative distribution function of the lower endpoint of the interval-valued residual $[\underline{R}_f, \overline{R}_f]$, while the lower bound of the p-box corresponds to the cumulative distribution function of the upper endpoint. This can be seen by considering the corresponding bounds to the probabilities of the events $(-\infty, r]$, with $r \in \mathbb{R}_{\geq 0}$, used to derive the p-box for all $f \in \mathcal{F}$, that is,

$$\begin{aligned} &P_V(R_f \leq r) \geq P_{V^*}([\underline{R}_f, \overline{R}_f] \subseteq (-\infty, r]) \quad \text{and} \\ &P_V(R_f \leq r) \leq P_{V^*}([\underline{R}_f, \overline{R}_f] \cap (-\infty, r] \neq \varnothing) \end{aligned}$$

It can easily be checked that the probability distributions corresponding to the bounds of the p-box comply with (3) for arbitrary measurable events, and thus, are elements of $[P_{V^*}]$. Since, according to its definition in (1), the risk functional $\mathcal{E}_{P_V}$ is the expectation of a convex function in R_f with minimum at zero, it is straightforward to conclude that $\underline{\mathcal{E}}_{P_{V^*}}$ and $\overline{\mathcal{E}}_{P_{V^*}}$ coincide with the expected errors associated with the marginal distributions of the lower and of the upper residual, that is, of $\underline{R}_f$ and of $\overline{R}_f$, respectively (see also Utkin and Destercke, 2009, Proposition 3).

Now consider that the realization of an independent sample of random sets $V_1^* = A_1, \ldots, V_n^* = A_n$ is observed, where $V_i^* \sim P_{V^*}$ for all $i \in \{1, \ldots, n\}$. Then, by analogy with standard SVR, P_{V^*} is estimated by the empirical distribution $\hat{P}_{V^*}$ of the imprecise data,

and furthermore, the complexity of the estimated functions is restricted by an additive penalty term. Hence, the optimization criteria considered in the minimin and minimax generalizations of SVR are the regularized lower and upper risk, respectively. For a fixed penalization parameter $\lambda > 0$, the regularized lower and upper risks associated with the empirical distribution $\hat{P}_{V^*}$ can, for each $f \in \mathcal{F}$, be expressed as follows:

$$\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} \min_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \|f\|_{\mathcal{F}}^2,$$
$$\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f) = \frac{1}{n}\sum_{i=1}^{n} \max_{(x_i,y_i)\in A_i} \psi(|y_i - f(x_i)|) + \lambda \|f\|_{\mathcal{F}}^2, \tag{4}$$

where ψ is again the convex mapping from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ representing the chosen loss.

Utkin and Coolen (2011) deduce from these expressions of the regularized empirical lower and upper risks solvable formulations of the optimization problems corresponding to both suggested strategies in the special case of linear regression for different choices of the loss function. We do not restrict the approach to this special case here and continue to consider more general RKHSs of regression functions. Moreover, Utkin and Coolen (2011) start from the support vector expansion (2) of the solution of the optimization problem corresponding to standard SVR. However, it first has to be verified that the Representer Theorem applies to or that its statements can be transferred to the setting with interval data. Only in this case, the simple expression (2) can be used for the optimal regression function in (4), providing the favorable starting point for solving the corresponding optimization problems.

3.2 The Representer Theorem for SVR with Interval-Valued Response

As mentioned in the previous subsection, the Representer Theorem implies that if an SVR analysis of a precise data set $V_1 = (x_1, y_1), \ldots, V_n = (x_n, y_n)$ with empirical distribution $\hat{P}_V$ is based on a convex representing function ψ, then, for all $\lambda > 0$, there exists a unique function minimizing $\mathcal{E}_{\hat{P}_V,\lambda}$, which can be represented as (2) (see, e.g., Steinwart and Christmann, 2008, Theorem 5.5). In the proof of this theorem as it is presented in Steinwart and Christmann (2008, Theorem 5.5), the first steps are to show strict convexity and continuity of $\mathcal{E}_{\hat{P}_V,\lambda}$, which provide existence and uniqueness of the minimizing function $f_{\hat{P}_V,\lambda} \in \mathcal{F}$, by the corresponding arguments of the proofs of Theorem 5.2 and Lemma 5.1 of Steinwart and Christmann (2008), respectively. Then, the representation of $f_{\hat{P}_V,\lambda}$ as the kernel expansion of (2) is derived by exploiting properties of the function spaces $\mathcal{F}_n$ and $\mathcal{F}$ in addition to the existence and the uniqueness of the function $f_{\hat{P}_V,\lambda}$.

The generalized SVR methods discussed in this section differ from the standard SVR methods only in the expressions of their risks. Hence, we have to derive the crucial properties of convexity and continuity for the lower and upper risks to be able to transfer the arguments proving the simplified expression of $f_{\hat{P}_V,\lambda}$ to the situation with interval-valued response. In the following lemma, we derive for the general case that the regularized lower and upper risks have unique minimizers, before we prove Theorem 1, stating that the functions minimizing the regularized empirical lower and upper risks can be expressed as in Equation (2).

Lemma 1. *The regularized lower and upper risk functionals*

$$\underline{\mathcal{E}}_{P_{V^*},\lambda} : f \mapsto \underline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \quad \textit{and}$$
$$\overline{\mathcal{E}}_{P_{V^*},\lambda} : f \mapsto \overline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2$$

have unique minimizers $f_{P_{V^*},\lambda}^{\text{minimin}}$ and $f_{P_{V^*},\lambda}^{\text{minimax}}$ in $\mathcal{F}$, respectively.

Proof. Since κ is bounded, convergence in the norm $\|\cdot\|_{\mathcal{F}}$ implies convergence in the norm $\|\cdot\|_\infty$, because using the Cauchy–Schwarz inequality,

$$\begin{aligned}\|f\|_\infty &= \sup_{x\in\mathcal{X}} \|f(x)\| = \sup_{x\in\mathcal{X}} \|\langle f, \kappa(\cdot,x)\rangle_{\mathcal{F}}\| \\ &\leq \sup_{x\in\mathcal{X}} \|f\|_{\mathcal{F}} \sqrt{\langle \kappa(\cdot,x), \kappa(\cdot,x)\rangle_{\mathcal{F}}} \\ &= \|f\|_{\mathcal{F}} \sup_{x\in\mathcal{X}} \sqrt{\kappa(x,x)}\end{aligned}$$

for all $f \in \mathcal{F}$. Therefore, the functionals $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are continuous on $\mathcal{F}$ (with respect to the norm $\|\cdot\|_{\mathcal{F}}$), because they are the sum of the continuous functional $\lambda\|\cdot\|_{\mathcal{F}}^2$ with the lower and upper previsions of $\psi(R_f)$, respectively, and ψ is uniformly continuous on the relevant domain (since it is convex, and $\mathcal{Y}$ is bounded).

Moreover, $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are strictly convex functionals on $\mathcal{F}$, since $\lambda\|\cdot\|_{\mathcal{F}}^2$ is strictly convex, and the unregularized lower and upper risk functionals $\underline{\mathcal{E}}_{P_{V^*}}$ and $\overline{\mathcal{E}}_{P_{V^*}}$ can be shown to be convex. The proof for the upper risk functional is simple, since $\overline{\mathcal{E}}_{P_{V^*}}$ is the maximum of the convex functionals $\mathcal{E}_{P'_V}$ with $P'_V \in [P_{V^*}]$. By contrast, the proof for the lower risk functional is more involved. We start by noting that for each possible realization $A = \{x\} \times [\underline{y}, \overline{y}] \in \mathcal{V}^*$ of the random set V^*, the function

$$\underline{r}_A : z \mapsto \min_{\underline{y}\leq y\leq \overline{y}} |y - z| = \begin{cases} \underline{y} - z & \text{if } z < \underline{y}, \\ 0 & \text{if } \underline{y} \leq z \leq \overline{y}, \\ z - \overline{y} & \text{if } \overline{y} < z, \end{cases}$$

on $\mathbb{R}$ is convex, and therefore $\psi \circ \underline{r}_A$ is convex too, since ψ is convex and nondecreasing. This implies that

$$\min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V \,|\, V^*} \left(\psi(R_f) \,|\, V^* = A\right) = \left(\psi \circ \underline{r}_A\right)\left(f(x)\right)$$

is a convex functional of f, and so is

$$\begin{aligned}\underline{\mathcal{E}}_{P_{V^*}}(f) &= \min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V}\left(\psi(R_f)\right) \\ &= \mathbb{E}_{P_{V^*}}\left(\min_{P'_V \in [P_{V^*}]} \mathbb{E}_{P'_V \,|\, V^*}\left(\psi(R_f) \,|\, V^*\right)\right).\end{aligned}$$

So far we have proven that $\underline{\mathcal{E}}_{P_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{P_{V^*},\lambda}$ are continuous and strictly convex functionals on $\mathcal{F}$. The desired result is implied by Theorem A.6.9 of Steinwart and Christmann (2008), since the sets

$$\left\{f \in \mathcal{F} : \underline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \leq \underline{\mathcal{E}}_{P_{V^*}}(0)\right\} \quad \text{and}$$
$$\left\{f \in \mathcal{F} : \overline{\mathcal{E}}_{P_{V^*}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \leq \overline{\mathcal{E}}_{P_{V^*}}(0)\right\}$$

are nonempty and bounded (with respect to the norm $\|\cdot\|_{\mathcal{F}}$). □

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

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

are the unique minimizers of $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ *and* $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ *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 $i \in \{1, \ldots, n\}$. 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 $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Moreover, for each $i \in \{1, \ldots, n\}$, 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, $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f') \leq \underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f)$ and $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f') \leq \overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}(f)$, and the desired result is implied by Lemma 1. □

Hence, $f(x_i)$ can indeed be replaced by a support vector expansion in the expressions of $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ and $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ given in (4), and the derivation of solvable formulations of the corresponding optimization problems can be based on the thereby simplified expressions of the risks.

However, the above results cannot directly be generalized to accounting also for interval-valued observations of the explanatory variables. This is because, when V^* is of the form $[\underline{X}^{(1)}, \overline{X}^{(1)}] \times \ldots \times [\underline{X}^{(d)}, \overline{X}^{(d)}] \times [\underline{Y}, \overline{Y}]$, in general $\underline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ is no longer convex, and moreover, Theorem 1 does not apply to $\overline{\mathcal{E}}_{\hat{P}_{V^*},\lambda}$ anymore.

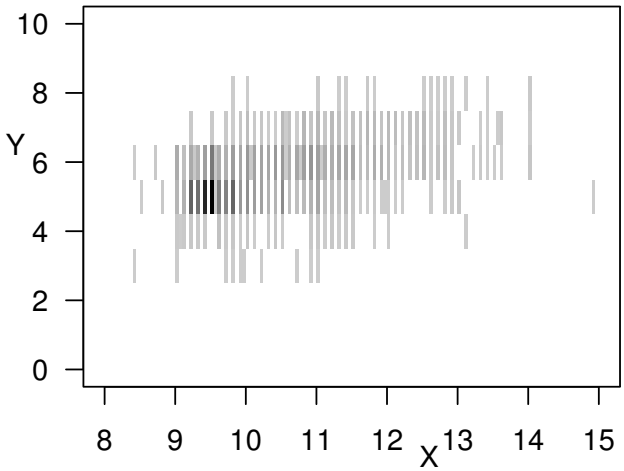

Figure 1: Histogram plot of the red wine data set with $n = 1\,599$ observations. The darker a line segment the more observations overlap this line segment.

4 SVR Analysis of Wine Quality

In this section, we analyze a data set collected to study the quality of Vinho Verde wines from Portugal. The data were obtained from wine samples that were tested by the official certification entity of the system of protected designation of origin of the Vinho Verde wines between May 2004 and February 2007. For each of the included 1 599 red and 4 898 white wines, 11 physicochemical characteristics and an evaluation of the sensory quality are available. The data set was initially analyzed by Cortez et al. (2009) and is freely available from the UC Irvine Machine Learning Repository (Lichman, 2013). Here, we focus on the subsample of red Vinho Verde wines and study the relationship between taste and alcohol content.

In the data set, the sensory quality of the wine is measured on a discrete scale ranging from 0 – *very bad* to 10 – *excellent*. These discrete quality measurements should, in fact, be considered as coarse observations of an underlying continuous variable taking values in $[0, 10]$. Therefore, instead of analyzing the discrete values as if they were precise measurements of the wine quality, we consider them to be interval data and replace the discrete values $0, 1, \ldots, 9, 10$ by the intervals $[0, 0.5], [0.5, 1.5], \ldots, [8.5, 9.5], [9.5, 10]$, respectively. The alcohol content of the wines is available as volume percent of alcohol, which we here assume to be measured with sufficient precision.

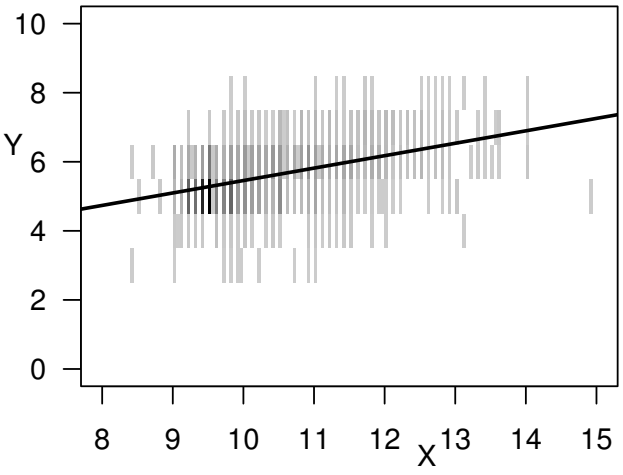

Figure 2: Minimax function of the generalized SVR analysis with linear kernel, $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$, and $\lambda = 0.0001$.

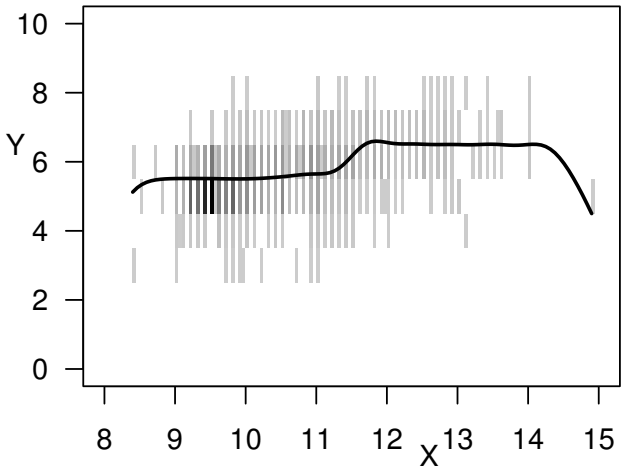

Figure 3: Minimax function of the generalized SVR analysis with Gaussian kernel, $\psi(r) = r$ for all $r \in \mathbb{R}_{\geq 0}$, and $\lambda = 0.000001$.

Hence, we analyze the relationship between the precisely observed alcohol content and the imprecisely observed sensory quality of the red Vinho Verde wine. Thus, as we consider only one explanatory variable here, the imprecise data are line segments. The analyzed data set is displayed in Figure 1, where X is the alcohol level in percent by volume and Y corresponds to the sensory quality. All graphs and computations are realized in the statistical software environment R (R Core Team, 2014), resorting amongst others to functions provided by the packages `kernlab` (Karatzoglou et al., 2004) and `quadprog` (Turlach and Weingessel, 2013).

A red wine lover would probably hypothesize that the higher the alcohol content of a red wine, the stronger and possibly better the taste of the wine. As also the data suggest a positive linear relationship, in the first instance, we choose the linear kernel function for the SVR analysis, although SVR is not limited to linear regression. Furthermore, we consider the Least Squares loss, i.e., we set $\psi(r) = r^2$ for all $r \in \mathbb{R}_{\geq 0}$. This configuration of SVR corresponds to what is also known as Ridge regression. As the minimax approach appears to be more cautious, we consider the corresponding generalized SVR method of Utkin and Coolen (2011) here. Finally, for the estimation, the regularization parameter λ is set to 0.0001. The estimated regression line confirms the surmise of a positive relationship between alcohol content and sensory quality of the Vinho Verde red wines and is displayed in Figure 2.

As the assumption of a linear relationship is very strict, we alternatively consider the minimax SVR method based on the Gaussian kernel with parameter σ equal to 1. Furthermore, we consider the absolute loss here represented by ψ defined as $\psi(r) = r$ for all $r \in \mathbb{R}_{\geq 0}$ and set $\lambda = 0.000001$. The estimated regression function is depicted in Figure 3 and shows an increasing tendency in those areas of the observation space $\mathcal{V} = [8, 15] \times [0, 10]$ where most observations are. Hence, also the more general SVR analysis provides evidence for a positive relationship between alcohol content and sensory quality of red Vinho Verde wines.

5 Conclusion and Outlook

In this paper, we investigated the generalized SVR methods for regression with interval data that were initially proposed by Utkin and Coolen (2011). These methods consist in minimizing either the minimal or the maximal regularized risk compatible with the empirical distribution of the imprecise data. In this paper, we proved that the corresponding optimal functions can be represented as the weighted sum of kernel functions and thereby provide the so far lacking justification for the regression methods derived in Utkin and Coolen (2011). Hence, the minimin and minimax SVR methods constitute sensible adaptations of the SVR methodology to interval data and yield interesting results when applied to real data as in the previous section.

We here focused on the data situation where only for the response variable there are interval-valued observations, while the explanatory variables are precisely observed. Unfortunately, our findings cannot simply be generalized to account also for interval-valued observations of the explanatory variables, because then the regularized lower risk is no longer necessarily convex and the Representer Theorem cannot be transferred to the regularized upper risk anymore. This means that

for the minimin SVR method there is not necessarily a unique optimal function and that the optimal minimax function cannot be expanded as in Equation (2). This indeed limits the applicability of the minimin and minimax SVR methods to the more restrictive setting considered in this paper. Moreover, the meaning of the estimated regression functions is less clear than in the precise data case.

Furthermore, it can be argued that, in the context of the statistical analysis of imprecise data, methods yielding precise results are in general problematic, because a reasonable statistical method should reflect the imprecision of the data in its result. In addition, a responsible statistical analysis should always take the involved statistical uncertainty into account. A regression methodology for imprecise data allowing to express these two types of uncertainty at the same time constitutes the so-called Likelihood-based Imprecise Regression (LIR) methodology introduced by Cattaneo and Wiencierz (2012). In the LIR methodology, each possible regression function is evaluated by the whole set of loss values that are plausible in the light of the data and then the set of all undominated regression functions is considered as the imprecise result of the regression analysis, which can furthermore be interpreted as a confidence set. As it can be shown that, for each $f \in \mathcal{F}$, the interval $[\underline{\mathcal{E}}_{\hat{P}_{V^*}}(f), \overline{\mathcal{E}}_{\hat{P}_{V^*}}(f)]$ is the Maximum Likelihood estimate of $\mathcal{E}_{P_V}(f)$ in the situation considered in Section 3, Utkin and Coolen (2011)'s SVR methods can be further generalized by embedding them in the LIR framework.

References

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cattaneo, M., and Wiencierz, A. (2012). Likelihood-based Imprecise Regression. *International Journal of Approximate Reasoning* 53, 1137–1154.

Christmann, A., Van Messem, A., and Steinwart, I. (2009). On consistency and robustness properties of Support Vector Machines for heavy-tailed distributions. *Statistics and Its Interface* 2, 311–327.

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.

Destercke, S., Dubois, D., and Chojnacki, E. (2008). Unifying practical uncertainty representations: I. Generalized p-boxes. *International Journal of Approximate Reasoning* 49, 649–663.

Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D., and Sentz, K. (2003). *Constructing Probability Boxes and Dempster-Shafer Structures*. Technical Report SAND2002-4015, Sandia National Laboratories.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis* 106, 92–117.

Hofmann, T., Schölkopf, B., and Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistics* 36, 1171–1220.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software* 11, 1–20.

Lichman, M. (2013). *UCI Machine Learning Repository*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Used R version 2.15.2.

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

Turlach, B., and Weingessel, A. (2013). *quadprog: Functions to solve Quadratic Programming Problems*. R package version 1.5-5.

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

Utkin, L., and Destercke, S. (2009). Computing expectations with continuous p-boxes: Univariate case. *International Journal of Approximate Reasoning* 50, 778–798.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.