J Comb Optim (2014) 28:218–232
DOI 10.1007/s10878-013-9678-9

Value-at-risk support vector machine: stability to outliers

Peter Tsyurmasto · Michael Zabarankin · Stan Uryasev

Published online: 5 December 2013

Abstract A support vector machine (SVM) stable to data outliers is proposed in three closely related formulations, and relationships between those formulations are established. The SVM is based on the value-at-risk (VaR) measure, which discards a specified percentage of data viewed as outliers (extreme samples), and is referred to as VaR-SVM. Computational experiments show that compared to the ν-SVM, the VaR-SVM has a superior out-of-sample performance on datasets with outliers.

Keywords Support vector machine · Classification · Conditional value-at-risk · Value-at-risk · Risk management · Optimization

1 Introduction

Nowadays, support vector machines (SVMs) are ubiquitous in various applications ranging from biomedicine (Kazama et al. 2002) and bioinformatics (Byvatov and Schneider 2003) to image recognition (Guo et al. 2000) and credit scoring (Huang et al. 2007). Also, SVM classification is closely related to structural risk minimization (Vapnik 1999). The idea of an SVM is to map a training dataset from two classes into a multidimensional space and to separate the classes with maximal margin.

P. Tsyurmasto (✉) · S. Uryasev
Department of Industrial and Systems Engineering, University of Florida,
303 Weil Hall, PO Box 116595, Gainesville, FL 32611-6595, USA
e-mail: tsyurmasto@ufl.edu

S. Uryasev
e-mail: uryasev@ufl.edu

M. Zabarankin
Department of Mathematical Sciences, Stevens Institute of Technology,
Castle Point on Hudson, Hoboken, NJ 07030, USA
e-mail: mzabaran@stevens.edu

The pioneering *hard-margin SVM* (Boser et al. 1992) requires each training sample to be classified correctly, whereas the *soft-margin SVM* (C-SVM) (Cortes and Vapnik 1995) involves a parameter C trading off the training error with the margin size. The ν-*SVM* (Schölkopf et al. 2000) replaces the parameter C in the C-SVM by a parameter $\nu \in [0, 1]$ imposing an upper bound for the percentage of data missclassifications, and the *extended* ν-*SVM* (Eν-SVM) (Pérez-Cruz et al. 2003) extends a range of ν for which a nontrivial solution exists.

The SVM literature offers several SVMs to deal with data outliers and noisy data. For example, to reduce the effect of outliers, the *fuzzy* SVM (Lin and Wang 2002) associates a fuzzy membership with each training sample in the C-SVM, although, it does not specify how to select a proper fuzzy membership function for a particular dataset. The *robust* SVM (Song et al. 2002) and the *center* SVM (Zhang 1999) use centers of classes along with support vectors to construct a classification boundary. However, when the sample distribution is not Gaussian and highly skewed, the mean (center) of the class may not be a representative or may fall outside of the class. For noisy and corrupt data, SVM classification relies on methods of robust optimization (Ben-Tal et al. 2009; Xanthopoulos et al. 2013a,b) and the rough set theory, see Zhang and Wang (2008). For example, in Trafalis and Gilbert (2006), the magnitude of noise is assumed to be bounded, and for relatively small bounds, the separating hyperplane remains almost unaffected. However, as the bound exceeds a certain threshold, the misclassification error considerably increases. Also, the upper bound on the noise can not be directly estimated from the data.

We propose an SVM that requires the "hard-margin" constraint to hold with probability $\alpha \in (0, 1]$. The SVM is reformulated with the value-at-risk (VaR) measure[1] and is referred further to as VaR-SVM. For $\alpha = 1$, the VaR-SVM reduces to the hard-margin SVM, whereas for $\alpha < 1$, it discards $(1 - \alpha) \cdot 100$ % of the training samples viewed as outliers (extreme samples). A similarity between SVM classification and optimization of monetary risk measures was observed in Goto and Takeda (2005), Takeda and Sugiyama (2008) and Sakalauskas et al. (2012). For example, the ν-SVM can be reformulated as an SVM with the conditional value-at-risk (CVaR) measure, which controls the average of the largest $(1 - \nu) \cdot 100$ % of misclassification errors. However, in contrast to the VaR-SVM, the ν-SVM discards no data samples. Computational experiments with artificial datasets show that compared to the ν-SVM, the VaR-SVM is stable to outliers or *statistically robust*.[2]

The paper is organized into six sections. Section 2 reformulates the well-known SVMs as SVMs with risk functionals. Section 3 introduces the VaR-SVM, whereas Sect. 4 specializes the VaR-SVM for the nonlinear case with the radial-basis-function (RBF) kernel. Section 5 compares the VaR-SVM with the ν-SVM on artificial and real-life datasets. Section 6 concludes the work.

[1] VaR with a confidence level α is the α-percentile of the distribution of loss and is widely used to monitor and control the market risk of financial instruments (Jorion 1997; Duffie and Pan 1997).

[2] A statistically robust SVM should not be confused with an SVM relying on methods of robust optimization (Ben-Tal et al. 2009): the former is unaffected if a certain percentage of the training data is changed, whereas the latter finds the separation hyperplane under the assumption that the training data are not clearly specified but rather are known to be from some set.

2 Reformulation of SVMs with risk functionals

Let $\{(\xi_1, y_1), \ldots, (\xi_l, y_l)\}$ be a training dataset of samples $\xi_i \in \mathbb{R}^m$ with class labels $y_i \in \{-1, +1\}$ for $i = 1, \ldots, l$. The original samples $\{\xi_1, \ldots, \xi_l\} \subset \mathbb{R}^m$ are transformed into $\{\phi(\xi_1), \ldots, \phi(\xi_l)\} \subset \mathbb{R}^n$ by mapping $\phi : \mathbb{R}^m \to \mathbb{R}^n$. The goal is to construct a hyperplane $w^\top x + b = 0, w \in \mathbb{R}^n, b \in \mathbb{R}$, that separates samples $\{\phi(\xi_1), \ldots, \phi(\xi_l)\}$ with class label $+1$ from those with class label -1 in the $\mathbb{R}^n$ space. In this case, sample ξ_i is classified correctly if $y_i(w^\top \phi(\xi_i) + b) \geq 0$ and incorrecly if $y_i(w^\top \phi(\xi_i) + b) < 0$.

Let $\Omega = \{\omega_1, \ldots, \omega_l\}$ be a finite sample space with equal[3] probabilities of outcomes, i.e. $\Pr(\omega_i) = 1/l, i = 1, \ldots, l$, and let $\xi : \Omega \to \mathbb{R}^n$ and $y : \Omega \to \{-1, +1\}$ be discrete random variables such that $\xi(\omega_i) = \phi(\xi_i)$, $y(\omega_i) = y_i$ for $i = 1, \ldots, l$. For each outcome $\omega \in \Omega$ and *decision variables* $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, a *loss function* is defined by

$$\mathcal{L}_\omega(w, b) = -y(\omega) \cdot \left[w^\top \xi(\omega) + b \right], \tag{1}$$

which is a random variable with realizations $\left\{-y_i \cdot [w^\top \xi_i + b]\right\}_{i=1}^l$ assuming equal probabilities $1/l$, so that sample ξ_i is classified correctly if $\mathcal{L}_{\omega_i}(w, b) \leq 0$ and incorrectly if $\mathcal{L}_{\omega_i}(w, b) > 0$.

Typically, a random loss X is translated into a real-valued number through risk functionals such as

- *Worst-case loss* $\sup_{\omega \in \Omega} X$.
- *Partial moment* $\mathbb{E}\,[X - \eta]_+$, which is the expected loss exceeding some specified threshold $\eta \in \mathbb{R}$, where $[\,\cdot\,]_+ = \max\{\cdot, 0\}$.
- *Conditional value-at-risk (CVaR)* $\mathrm{CVaR}_\alpha(X)$, defined as the average of the α-tail of the probability distribution of X for a specified confidence level $\alpha \in [0, 1]$.

All well-known SVMs admit a concise formulation with the above risk functionals.

The hard-margin SVM (Boser et al. 1992)

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^\top \phi(\xi_i) + b) \geq 1, \qquad i = 1, \ldots, l,$$

can be expressed with the worst-case loss functional by

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \sup_{\omega \in \Omega} \mathcal{L}_\omega(w, b) \leq -1. \tag{2}$$

The soft-margin SVM (C-SVM) (Cortes and Vapnik 1995)

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \left[-y_i \left(w^\top \phi(\xi_i) + b \right) + 1 \right]_+ \right), \quad C > 0,$$

[3] The approach can be readily extended to the case of arbitrary probabilities of outcomes: $\Pr(\omega_i) = p_i$, $i = 1, \ldots, l$.

can be rewritten with the partial moment functional as

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \left(\frac{1}{2}\|w\|^2 + C'\,\mathbb{E}\,[\mathcal{L}_\omega(w,b)+1]_+\right), \quad C' = C\cdot l. \tag{3}$$

The ν-SVM (Schölkopf et al. 2000), originally formulated by

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R},\, \rho\geq 0} \left(\frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\Big[\rho - y_i(w^\top\phi(\xi_i)+b)\Big]_+\right), \tag{4}$$

can be recast in the form

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \left(\frac{1}{2}\|w\|^2 + \nu\,\mathrm{CVaR}_{1-\nu}\left(\mathcal{L}_\omega(w,b)\right)\right), \tag{5}$$

which follows from Rockafellar–Uryasev's optimization formula (Rockafellar and Uryasev 2000) for CVaR:

$$\mathrm{CVaR}_{1-\nu}\left(\mathcal{L}_\omega(w,b)\right) = \min_{\rho\in\mathbb{R}}\left(-\rho + \frac{1}{\nu l}\sum_{i=1}^{l}\Big[\rho - y_i(w^\top\phi(\xi_i)+b)\Big]_+\right) \tag{6}$$

and the fact that the condition $\rho \geq 0$ is redundant as shown in Burges (2000). The relationship between the ν-SVM and CVaR minimization was first reported in Goto and Takeda (2005). For $\nu > 0$, the SVM (5) is closely related to the formulation

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \mathrm{CVaR}_{1-\nu}(\mathcal{L}_\omega(w,b)) \leq -1. \tag{7}$$

The problems (5) and (7) are both convex, and Corollary 2 will show that if exist, their solutions coincide up to a positive multiplier for $\nu > 0$.

Finally, Takeda and Sugiyama (2008) showed that the Eν-SVM (Pérez-Cruz et al. 2003) can be formulated by

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \mathrm{CVaR}_{1-\nu}\left(\mathcal{L}_\omega(w,b)\right) \quad \text{s.t.} \quad \|w\| = 1. \tag{8}$$

The three risk functionals: worst-case loss, partial moment, and CVaR have different levels of tolerance to the misclassification error. The worst-case loss functional is the most conservative among the three, and the corresponding SVM (2) can be interpreted as a robust optimization problem (Zhang and Wang 2008). In the SVM formulations (5), (7), and (8), the greater the parameter ν is, the more tolerant to misclassifications the SVMs are: $\nu = 0$ and $\nu = 1$ correspond to the most and least conservative cases, respectively. In fact, for $\nu = 0$, the SVMs (7) and (8) are equivalent to the SVM (2) with the worst-case loss functional, which is the hard-margin SVM.

3 VaR-SVM

The SVMs (2), (3), (5), and (8) are sensitive to outliers (extreme samples), since the supremum, partial moment, and CVaR all rely on the right tail of the loss distribution that contains extreme data samples. Specifically, the partial moment is the average of the losses exceeding -1, whereas CVaR averages $(1-\alpha)\cdot 100\,\%$ of the largest losses, and the supremum is the largest single loss. However, SVM's sensitivity to extreme data samples can be reduced by using risk functionals that discard the largest values in the right tail of the loss distribution.

The hard-margin SVM (2) has the equivalent formulation

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \Pr[\mathcal{L}_\omega(w,b)\le -1]=1, \tag{9}$$

which suggests that the constraint $\mathcal{L}_\omega(w,b)\le -1$ can be required to hold with probability $\alpha\in(0,1]$, i.e.

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \Pr[\mathcal{L}_\omega(w,b)\le -1]\ge \alpha. \tag{10}$$

With *value-at-risk* (VaR), or percentile function, defined by

$$\begin{aligned}\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b)) &= \min_{z\in\mathbb{R}} \{z \mid \Pr[\mathcal{L}_\omega(w,b)\le z]\ge\alpha\}\\ &\equiv \min\left\{ z \,\middle|\, \frac{1}{l}\sum_{i=1}^{l} \mathbb{1}_{\{-y_i[w^\top\phi(\xi_i)+b]\le z\}}\ge\alpha \right\},\end{aligned} \tag{11}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function equal to 1 if the condition in curly brackets is true and equal to 0 otherwise, the problem (10) can be rewritten in the form

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b))\le -1, \tag{12}$$

which will be referred to as *VaR-SVM*. The parameter α in the VaR-SVM indicates that $(1-\alpha)\cdot 100\,\%$ of data is considered as outliers and, thus, is discarded, see Fig. 1. In contrast to the ν-SVM, the VaR-SVM is unaffected by outliers in the α-tail of the loss distribution.

Observe that the VaR-SVM (12) resembles (7), which is closely related to (5). An SVM with VaR in place of CVaR in (5) is formulated by

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \left(\frac{1}{2}\|w\|^2 + C\cdot\ \mathrm{VaR}_\alpha\left(\mathcal{L}_\omega(w,b)\right)\right), \quad C>0, \tag{13}$$

where C is a given constant. It will be shown that if exist, solutions to (12) and (13) coincide up to a positive multiplier.

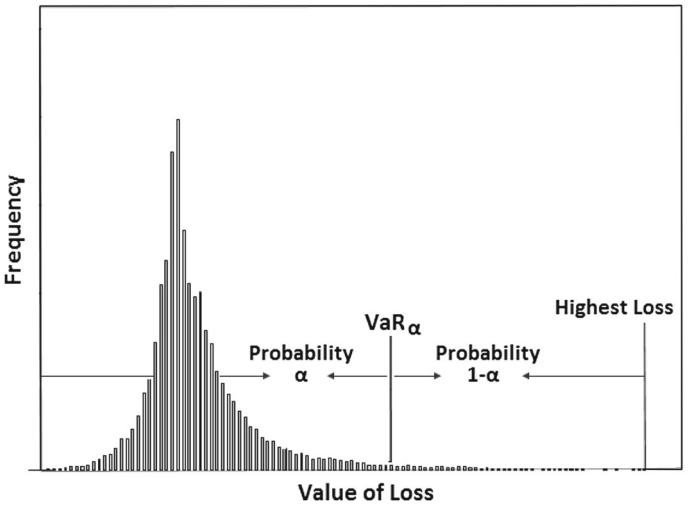

Fig. 1 Histogram for the probability density function of the loss distribution: VaR_α is the α-percentile of the loss distribution

The VaR-SVM minimizes the α-quantile of the distance $d_\omega(w, b)$ from the vector $\xi(\omega)$ to the separating hyperplane $H = \{x \in \mathbb{R}^n \mid w^\top x + b = 0\}$ in $\mathbb{R}^n$. For each outcome $\omega \in \Omega$ and decision variables $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, $d_\omega(w, b)$ is defined by

$$d_\omega(w, b) = \frac{\mathcal{L}_\omega(w, b)}{\|w\|}. \tag{14}$$

The distance assumes values $-y_i(w^\top \phi(\xi_i) + b)/\|w\|$, $i = 1, \ldots, l$, with equal probabilities $1/l$. It is nonpositive for correctly classified $\xi(\omega)$ and is positive for incorrectly classified $\xi(\omega)$. Figure 2 shows the histogram of $d_\omega(w, b)$ for the data $(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)$ from German Credit Data[4] for some fixed w and b. The blue-colored and red-colored samples are the samples classified correctly and incorrectly, respectively. The goal is to minimize the number of red-colored samples by varying the parameters (w, b) in (14), so that the VaR-SVM is formulated by

$$\min_{w \in \mathbb{R}^n,\ b \in \mathbb{R}} \mathrm{VaR}_\alpha \left(\frac{\mathcal{L}_\omega(w, b)}{\|w\|} \right). \tag{15}$$

Next two theorems establish relationships between optimization problems (12), (13), and (15).

Theorem 1 *If (w^*, b^*) is an optimal solution of (15) and $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, then $(-1/\zeta)(w^*, b^*)$ is optimal for (12). If (w^*, b^*) is an optimal solution of (12) with $w^* \neq 0$, then $(\lambda w^*, \lambda b^*)$ is optimal for (15) for each $\lambda > 0$.*

[4] The dataset was taken from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html).

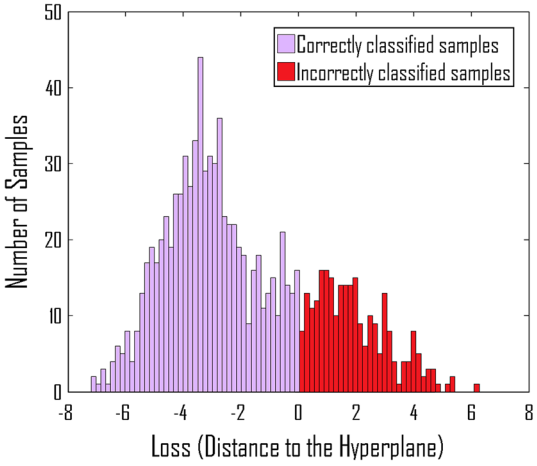

Fig. 2 Histogram of the distances (14) of the data samples $(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)$ to the separating hyperplane for German Credit Data with some fixed decision variables

Proof If (w^*, b^*) is optimal for (12) then $\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) = -1$. Indeed, suppose that $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < -1$, then $-1/\zeta \in (0, 1)$ and $(\tilde{w}, \tilde{b}) = (-1/\zeta)(w^*, b^*)$ is feasible for (12) with $\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w}, \tilde{b})) = -1$, but $\|\tilde{w}\| < \|w^*\|$, which contradicts the optimality of (w^*, b^*).

With the assumption that $w^* \neq 0$, this fact and the positive homogeneity of $\mathrm{VaR}_\alpha(\cdot)$ imply that the problem (12) is equivalent to

$$\min_{w\in\mathbb{R}^n,\, b\in\mathbb{R}} \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w, b)}{\|w\|}\right) \quad \text{s.t.} \quad \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w, b)) = -1. \tag{16}$$

If (w^*, b^*) is optimal for (15) and $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, then the positive homogeneity of $\mathrm{VaR}_\alpha(\cdot)$ implies that $(-1/\zeta)(w^*, b^*)$ is optimal for (15) and also that $(-1/\zeta)\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) = -1$. Consequently, $(-1/\zeta)(w^*, b^*)$ is feasible for (16) and, thus, it is optimal for (16).

Now suppose that (w^*, b^*) is optimal for (16) but is not optimal for (15), i.e. there exists $(\tilde{w}, \tilde{b})$ such that

$$\mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(\tilde{w}, \tilde{b})}{\|\tilde{w}\|}\right) < \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w^*, b^*)}{\|w^*\|}\right).$$

Then $\mu(\tilde{w}, \tilde{b})$ with $\mu = (-1/\zeta)$ is feasible for (16) and

$$\mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(\mu\tilde{w}, \mu\tilde{b})}{\|\mu\tilde{w}\|}\right) = \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(\tilde{w}, \tilde{b})}{\|\tilde{w}\|}\right) < \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w^*, b^*)}{\|w^*\|}\right),$$

which contradicts the assumption that (w^*, b^*) is optimal for (16).

Also, positive homogeneity of VaR_α implies that $(\lambda w^*, \lambda b^*)$ is optimal for (15) for each $\lambda > 0$. □

Theorem 2 *If* (w^*, b^*) *is an optimal solution of* (15) *and* $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, *then* $\frac{Cr^*}{\|w^*\|}(w^*, b^*)$ *is optimal for* (13), *where* $r^* = -\zeta/\|w^*\|$. *If* (w^*, b^*) *is an optimal solution of* (13) *with* $w^* \neq 0$, *then* $(\lambda w^*, \lambda b^*)$ *is optimal for* (15) *for each* $\lambda > 0$.

Proof Let (w^*, b^*) be an optimal solution of (15) such that $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, and let $r^* = -\zeta/\|w^*\|$, which is positive. Then

$$\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w, b))}{\|w\|} \geq \frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*))}{\|w^*\|} = -r^*$$

for all $w \neq 0$ and b. Thus, for any $w \neq 0$ and b, the objective function of (13) is bounded from below by

$$\begin{aligned}\frac{1}{2}\|w\|^2 + C \cdot \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w, b)) &= \frac{1}{2}\|w\|^2 + C\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w, b))}{\|w\|}\|w\| \\ &\geq \frac{1}{2}\|w\|^2 - Cr^*\|w\| \geq -\frac{1}{2}(Cr^*)^2. \qquad (17)\end{aligned}$$

Observe that for $(\tilde{w}, \tilde{b}) = \frac{Cr^*}{\|w^*\|}(w^*, b^*)$, the inequality (17) reduces to the equality

$$\begin{aligned}\frac{1}{2}\|\tilde{w}\|^2 + C \cdot \mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w}, \tilde{b})) &= \frac{1}{2}(Cr^*)^2\frac{\|w^*\|^2}{\|w^*\|^2} + C(Cr^*)\underbrace{\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*))}{\|w^*\|}}_{=-r^*} \\ &= -\frac{1}{2}(Cr^*)^2. \qquad (18)\end{aligned}$$

If $w = 0$, then the left-hand side of (17) is equal to $C \cdot \mathrm{VaR}_\alpha(-y(\omega)b) = C|b|\,\mathrm{VaR}_\alpha(-y(\omega)\,\mathrm{sign}\,b) \geq 0$, since the existence of an optimal solution of (15) implies $\mathrm{VaR}_\alpha(\pm y(\omega)) \geq 0$. Indeed, let $\mathrm{VaR}_\alpha(-y(\omega)) < 0$ and let $w = w_0$ be fixed with $\|w_0\| = 1$, then

$$\lim_{b\to\infty} \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w_0, b)}{\|w_0\|}\right) = \lim_{b\to\infty} |b|\underbrace{\mathrm{VaR}_\alpha\left(-y(\omega)\cdot[\delta_\omega(b)+1]\right)}_{<-\epsilon \text{ for sufficiently small } \delta_\omega(b)} = -\infty,$$

where $\delta_\omega(b) = w_0^\top \xi(\omega)/|b| \to 0$ as $b \to \infty$, and ϵ is a positive number. Similarly, it can be shown that $\mathrm{VaR}_\alpha(y(\omega)) \geq 0$, so that (17) holds for $w = 0$, and consequently, $(\tilde{w}, \tilde{b})$ is optimal for (13).

Now suppose that (w^*, b^*) is optimal for (13) with $w^* \neq 0$ but that it is not optimal for (15), i.e. there exists $(\tilde{w}, \tilde{b})$ such that

$$\mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(\tilde{w}, \tilde{b})}{\|\tilde{w}\|}\right) < \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w^*, b^*)}{\|w^*\|}\right).$$

Similarly to the inequality (17), we obtain

$$\begin{aligned}\frac{1}{2}\|w^*\|^2 + C\cdot \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*,b^*)) &= \frac{1}{2}\|w^*\|^2 + C\cdot \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right)\|w^*\| \\ &> \frac{1}{2}\|w^*\|^2 + C\cdot \frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w},\tilde{b}))}{\|\tilde{w}\|}\|w^*\| \\ &= \frac{1}{2}\|w^*\|^2 - C\tilde{r}\|w^*\| \geq -\frac{1}{2}(C\tilde{r})^2, \end{aligned} \tag{19}$$

where $\tilde{r} = -\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w},\tilde{b}))/\|\tilde{w}\|$. Since (w^*, b^*) is optimal for (13), the inequality (19) yields

$$\frac{1}{2}\|w\|^2 + C\cdot \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b)) > -\frac{1}{2}(C\tilde{r})^2 \tag{20}$$

for any w and b. Similarly to (18), it can be shown that for $(\hat{w},\hat{b}) = \frac{C\tilde{r}}{\|\tilde{w}\|}(\tilde{w},\tilde{b})$,

$$\frac{1}{2}\|\hat{w}\|^2 + C\cdot \mathrm{VaR}_\alpha(\mathcal{L}_\omega(\hat{w},\hat{b})) = -\frac{1}{2}(C\tilde{r})^2,$$

which contradicts (20), so that (w^*, b^*) is optimal for (15), and the positive homogeneity of $\mathrm{VaR}_\alpha(\cdot)$ implies that $(\lambda w^*, \lambda b^*)$ is optimal solution of (15) for each $\lambda > 0$ as well. □

The relationship between problems (12) and (13) follows from Theorems 1 and 2.

Corollary 1 *If (w^*, b^*) is optimal for (12) with $w^* \neq 0$, then $\frac{C}{\|w^*\|^2}(w^*, b^*)$ is optimal for (13). Conversely, if (w^*, b^*) is optimal for (13) with $w^* \neq 0$ and $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, then $(-1/\zeta)(w^*, b^*)$ is optimal for (12).*

Remark 3 Corollary 1 implies that solving of (12) reduces to solving the unconstrained optimization problem (13).

Remark 4 The parameters (w, b) of the separating hyperplane $w^\top x + b = 0$ are determined up to a positive multiplier $\lambda > 0$. Therefore, when (12) and (13) have non-zero optimal solutions, they determine the same separating hyperplane. Also, (13) determines the same separating hyperplane for different values of $C > 0$.

Corollary 2 *Theorems 1 and 2 and Corollary 1 can be readily extended to establish a relationship between the problems (5) and (7). Namely, if (w^*, b^*) is optimal for (7) with $w^* \neq 0$, the $\frac{C}{\|w^*\|^2}(w^*, b^*)$ is optimal for (5). Conversely, if (w^*, b^*) is optimal for (5) with $w^* \neq 0$ and $\zeta = \mathrm{CVaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, then $(-1/\zeta)(w^*, b^*)$ is optimal for (7).*

4 Nonlinear VaR-SVM

Convex SVMs are usually solved in dual formulations, where the transformation ϕ is implicitly specified by a kernel function $K(\xi, \xi')$ (Schölkopf and Smola 2001). However, the VaR-SVM is not convex and cannot be solved through its dual. This section shows how to compare the VaR-SVM with, for example, the ν-SVM, which is solved in the dual formulation with the Gaussian (RBF) kernel.

There exists a linear transformation ψ of the original set of samples $\{\xi_1, \ldots, \xi_l\} \subset \mathbb{R}^m$ such that the scalar products of $\psi(\xi_j)$ and $\psi(\xi_j)$ are equal to those produced by $K(\xi, \xi')$, i.e. $\langle \psi(\xi_i), \psi(\xi_j) \rangle = K(\xi_i, \xi_j) \equiv \langle \phi(\xi_i), \phi(\xi_j) \rangle$ for all i and j (see, e.g., Cristianini and Shawe-Taylor 2000), so that the solution of a convex SVM with the transformed samples $\{\psi(\xi_1), \ldots, \psi(\xi_l)\} \subset \mathbb{R}^n$ coincides with that for the SVM dual with the kernel $K(\xi, \xi')$ corresponding to the original unknown transformation ϕ (Chapelle 2007).

For the samples $\{\xi_1, \ldots, \xi_l\}$, the kernel $K(\xi, \xi')$ yields a positive definite kernel matrix $\mathbf{K} = \{K(\xi_i, \xi_j)\}_{i,j=1}^{l}$, which can be decomposed as

$$\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top \equiv (\mathbf{V}\Lambda^{1/2})(\mathbf{V}\Lambda^{1/2})^\top, \tag{21}$$

where $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_l\}$ is a diagonal matrix with eigenvalues $\lambda_1 > 0, \ldots, \lambda_l > 0$ and $V = (v_1, \ldots, v_l)$ is an orthogonal matrix with corresponding eigenvectors $v_1, \ldots, v_l$ of $\mathbf{K}$. The representation (21) implies that $\psi : \xi_i \to (\mathbf{V}\Lambda^{1/2})_i, i = 1, \ldots, l$, is the sought linear transformation, where $(\mathbf{V}\Lambda^{1/2})_i$ is row i of the matrix $(\mathbf{V}\Lambda^{1/2})$. Thus, non-linear SVMs with the kernel K should be compared to the VaR-SVM with the transformed samples $\{\psi(\xi_1), \ldots, \psi(\xi_l)\}$.

5 Numerical experiments

The VaR-SVM (13) and the ν-SVM (5), both with ϕ being the identity function, are compared on artificial and real-life datasets with outliers. Although VaR optimization is an NP-complete problem (see e.g., Yang et al. 2002), it can be readily handled by various heuristic algorithms and smoothing techniques (see, e.g., Gaivoronski and Pflug 2005; Larsen et al. 2002) included in standard optimization packages. Specifically, we performed computations with Matlab using Portfolio Safeguard (PSG) solver[5], which has special heuristics to handle non-convex optimization, see Sect. 9.16 in Zabarankin and Uryasev (2013). With PSG, solving of the problems (13) and (5) involves three stages:

(1) *Formulating the optimization problems with precoded VaR and CVaR functions.* A typical meta-code uses 5–10 operators [see Appendix 1 for the PSG meta-code for (13)].
(2) *Data processing for the PSG functions in a required format.* Typically, VaR and CVaR functions are defined on the matrix of transformed training samples $\{(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)\}$.
(3) *Running PSG solver with the meta-code and processed data.*

[5] http://www.aorda.com/aod/welcome.action/psg.action

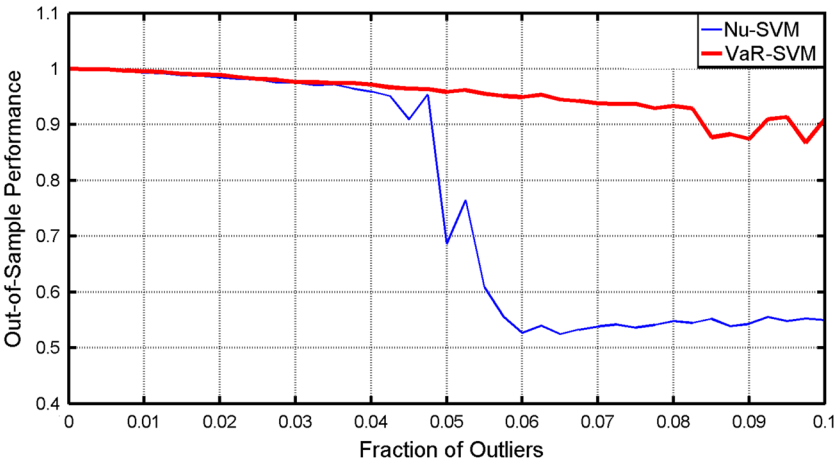

Fig. 3 Out-of-sample (OOS) performance of ν-SVM and of VaR-SVM as a function of the percentage of outliers for the artificial dataset

5.1 Artificial dataset

An artificial dataset consists of $l_1 = 400$ samples (with class label $+1$) and $l_2 = 400$ samples (with class label -1) generated from Gaussian distributions with mean vector μ and covariance matrix Σ: $(\mu_+, \Sigma_+) = (5e, 5I)$ and $(\mu_-, \Sigma_-) = (15e, 5I)$, where e is a vector of ones in $\mathbb{R}^{10}$ and I is an identity matrix in $\mathbb{R}^{10\times 10}$. The outliers are modeled to follow a Gaussian distribution $(\mu_{out}, \Sigma_{out})$ with $\mu_{out} = 100e$, $\Sigma_{out} = 20I$ and with class label $+1$. The percentage of outliers varies in the range 0–10 % of the original 800 samples. The VaR-SVM and the ν-SVM are then compared on out-of-sample (OOS) with the following parameters:

- The entire dataset is split randomly into training samples and testing (OOS) samples as 2:1;
- Number of random splitting of the dataset into training and testing is 10;
- The parameters ν in (5) and α in (13) are selected from a grid 0:0.0025:1 to minimize the OOS error, and parameter C in (13) is set to 0.5.

Figure 3 shows OOS performance of the VaR-SVM (13) and of the ν-SVM (5) as a function of the percentage of outliers for the artificial dataset. For a small number of outliers (<3 %), OOS performance is approximately the same for both classifiers and is close to 1. However, for a larger number of outliers (>4 %), the OOS performance plots deviate substantially. For the ν-SVM, the OOS performance dramatically drops down to almost 0.5 due to sensitivity of the ν-SVM to outliers. In contrast, for the VaR-SVM, the OOS performance stabilizes at the level of about 0.9. The 10 % of misclassifications correspond to 10 % of outliers. These results confirm that the VaR-SVM is stable to outliers.

Table 1 Experimental results for Liver Disorders dataset with outliers

Percentage of outliers %	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	Training		Testing		Training		Testing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	65.78	3.12	**63.65**	3.88	71.22	2.37	**69.65**	3.14
1	60.65	3.13	**59.35**	3.55	71.43	2.43	**68.35**	3.29
5	59.73	1.36	**58.87**	2.72	69.78	2.73	**66.52**	3.89
10	59.17	2.94	**58.78**	3.11	70.43	3.64	**65.74**	3.17

Table 2 Experimental results for Heart Disease dataset with outliers

Percentage of outliers %	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	Training		Testing		Training		Testing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	83.82	1.12	**82.96**	1.22	85.82	1.13	**81.84**	1.51
1	79.16	1.27	**78.29**	2.94	82.91	1.56	**81.61**	2.67
5	77.11	2.01	**76.53**	2.19	82.24	2.04	**81.12**	1.52
10	71.68	2.16	**70.71**	2.65	82.15	1.13	**81.53**	2.14

5.2 Real-life data sets

The problems (13) and (5) are solved with datasets from UCI Machine Learning Repository[6]: Liver Disorders, Heart Disease, Indian Diabetes, German Credit Data and Ionosphere. The original samples are normalized to have zero mean and unit standard deviation, and outliers are generated by artificially multiplying the fraction of 1, 5, and 10 % of the normalized samples by 1,000. Testing accuracy is evaluated with 10-fold cross validation. In both (5) and (13), ν and α are taken to be 0, 0.005, 0.01, ..., 0.095, 0.1. To determine the "best" values of ν and α among the selected ones, the training dataset is randomly split 100 times into two parts with the ratio of 2:1, where the first part (2/3 of the dataset) is used to find w and b. Then among the values of 0, 0.005, 0.01, ..., 0.095, and 0.1, the "best" ones for ν and α are those for which the average misclassification error on the second part (1/3 of the dataset) over 100 splits is minimal. Tables 1–5 show that as the percentage of outliers increases, the performance of the ν-SVM degrades significantly, whereas the performance of the VaR-SVM is almost unaffected. The running time for the VaR-SVM is slightly greater than that for the ν-SVM, though, it is of the same order (see Table 6).

[6] http://archive.ics.uci.edu/ml/datasets.html

Table 3 Experimental results for German Credit Dataset with outliers

Percentage of outliers %	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	Training		Testing		Training		Testing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	78.04	0.71	**74.86**	1.31	78.02	0.77	**73.42**	0.48
1	72.89	2.85	**72.49**	4.40	77.38	0.08	**74.14**	0.07
5	64.12	9.12	**62.70**	1.43	71.24	1.65	**70.93**	1.35
10	61.17	4.12	**60.03**	5.39	69.27	1.06	**70.39**	2.20

Table 4 Experimental results for Indian Diabetes Dataset with outliers

Percentage of outliers %	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	Training		Testing		Training		Testing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	78.14	2.13	**77.54**	1.84	78.13	0.10	**76.56**	2.31
1	76.93	1.34	**74.15**	1.82	77.91	2.02	**76.84**	1.78
5	64.00	3.69	**60.86**	4.13	76.33	2.46	**73.95**	3.22
10	60.18	8.19	**57.66**	7.63	75.00	2.92	**73.16**	2.97

Table 5 Experimental results for Ionosphere Dataset with outliers

Percentage of outliers %	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	Training		Testing		Training		Testing	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	73.41	1.29	**69.96**	1.88	75.88	1.36	**69.25**	1.73
1	65.18	2.13	**63.17**	1.23	76.05	1.03	**70.60**	2.68
5	63.95	2.56	**61.28**	3.02	73.12	1.73	**68.01**	2.92
10	64.22	1.97	**60.36**	1.05	71.31	2.18	**67.36**	1.92

Table 6 Running time of ν-SVM and VaR-SVM for different datasets with processor Intel(R) Core(TM)2 Quad CPU @2.83 GHz

Data set	# Samples	# Features	Running time (s)	
			ν-SVM	VaR-SVM
Liver Disorders	345	6	1.12	0.93
Heart Disease	294	13	0.80	1.27
Indian Diabetes	345	6	1.12	0.93
German	1,000	24	2.03	3.07
Ionosphere	796	14	1.76	2.49

6 Conclusions

The VaR-SVM has been proposed to overcome sensitivity of several well-known SVMs (hard-margin SVM, soft-margin SVM, ν-SVM, and Eν-SVM) to data outliers (extreme samples). Compared to the ν-SVM, the VaR-SVM has a superior OOS performance on artificial and real-life datasets. However, the VaR-SVM is non-convex, since VaR is not convex. This calls for the search of convex functionals having similar stability to outliers.

Acknowledgments We are grateful to the referees for their comments and suggestions, which helped to improve the quality of the paper. This research was supported by AFOSR Grant FA9550-11-1-0258, New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization.

Appendix 1: A PSG meta-code for VaR-SVM

The PSG meta-code, data, and solutions for the optimization problem (13) are available at the University of Florida Optimization Test Problems webpage,[7] see Problem 1b. For convenience, the PSG meta-code is presented below.

```
1  Problem: problem_var_svm, type = minimize
2  objective: objective_svm
3  quadratic_matrix_quadratic(matrix_quadratic)
4  var_risk_1(0.5,matrix_prior_scenarios)
5  box_of_variables: upperbounds =1,000, lowerbounds = −1,000
6  Solver: VAN, precision = 6, stages = 6
```

Command **minimize** instructs the solver that (13) is a minimization problem, whereas **objective** is a declaration of the objective function stated in lines 3 and 4: commands **quadratic** and **var_risk_1** refer to the quadratic and VaR terms in (13), respectively, and **matrix_quadratic** and **matrix_prior_scenarios** are the names of text files (*.txt) that store corresponding data matrices. The coefficients C and α are set to 0.5.

References

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization. Princeton University Press, Princeton

Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, ACM, pp 144–152

Byvatov E, Schneider G (2003) Support vector machine applications in bioinformatics. Appl Bioinformatics 2:67

Chapelle O (2007) Training a support vector machine in the primal. Neural Comput 19:1155–1178

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

[7] http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/case-study-nu-support-vector-machine-based-on-tail-risk-measures/

Crisp D, Burges C (2000) A geometric interpretation of ν-SVM classifiers. Adv Neural Inf Process Syst 12(12):244
Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
Duffie D, Pan J (1997) An overview of value at risk. J Deriv 4:7–49
Gaivoronski AA, Pflug G (2005) Value-at-risk in portfolio optimization: properties and computational approach. J Risk 7:1–31
Goto J, Takeda A (2005) Linear decision model based on conditional geometric score. In: Abstract collection of spring meeting for reading research paper by the Operations Research Society of Japan
Guo G, Li SZ, Chan K (2000) Face recognition by support vector machines. In: Proceedings of fourth IEEE international conference on automatic face and gesture recognition, pp 196–201
Huang C-L, Chen M-C, Wang C-J (2007) Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl 33:847–856
Jorion P (1997) Value at risk: the new benchmark for controlling market risk, vol 2. McGraw-Hill, New York
Kazama J, Makino T, Ohta Y, Tsujii J (2002) Tuning support vector machines for biomedical named entity recognition. In: Proceedings of the ACL-02 workshop on natural language processing in the biomedical domain, vol 3. Association for Computational Linguistics, pp 1–8
Larsen N, Mausser H, Uryasev S (2002) Algorithms for optimization of value-at-risk. Appl Optim 70:19–46
Lin C, Wang S (2002) Fuzzy support vector machines. IEEE Trans Neural Netw 13:464–471
Pérez-Cruz F, Weston J, Herrmann D, Scholkopf B (2003) Extension of the nu-SVM range for classification. NATO Sci Ser III Comput Syst Sci 190:179–196
Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. J Risk 2:21–42
Sakalauskas L, Tomasgard A, Wallace S (2012) Advanced risk measures in estimation and classification. Proceedings, Vilnius, pp 114–118
Schölkopf B, Smola A, Williamson R, Barlett P (2000) New support vector algorithms. Neural Comput 12:1207–1245
Schölkopf B, Smola A (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge
Song Q, Hu W, Xie W (2002) Robust support vector machine with bullet hole image classification. IEEE Trans Syst Man Cybern C Appl Rev 32:440–448
Takeda A, Sugiyama M (2008) ν-support vector machine as conditional value-at-risk minimization. In: Proceedings of the 25th international conference on machine learning, ACM, pp 1056–1063
Trafalis TB, Gilbert RC (2006) Robust classification and regression using support vector machines. Eur J Oper Res 173:893–909
Vapnik V (1999) The nature of statistical learning theory. Springer, Berlin
Xanthopoulos P, Guarracino MR, Pardalos PM (2013a) Robust generalized eigenvalue classifier with ellipsoidal uncertainty. Ann Oper Res 1–16
Xanthopoulos P, Pardalos PPM, Trafalis TB (2013b) Robust data mining. Springer, New York
Yang X, Tao S, Liu R, Cai M (2002) Complexity of scenario-based portfolio optimization problem with VaR objective. Int J Found Comput Sci 13:671–679
Zabarankin M, Uryasev S (2013) Statistical decision problems: selected concepts and portfolio safeguard case studies. Springer, Berlin
Zhang X (1999) Using class-center vectors to build support vector machines. In: Neural networks for signal processing IX. Proceedings of the 1999 IEEE signal processing society workshop, pp 3–11
Zhang J, Wang Y (2008) A rough margin based support vector machine. Inf Sci 178:2204–2214