Value-at-Risk Support Vector Machine: Stability to Outliers

Peter Tsyurmasto* Michael Zabarankin† Stan Uryasev*

April 30, 2013

RESEARCH REPORT # 2013-2

Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
University of Florida, Gainesville, FL 32611

Abstract

A new robust version of Support Vector Machine (SVM) based on value-at-risk (VaR) measure referred to as VaR-SVM is proposed in three closely related formulations, and relationships between those VaR-SVM formulations is established. In contrast to classical SVMs (hard-margin SVM, soft-margin SVM, and ν-SVM), VaR-SVM is stable to data outliers. Computational experiments confirm that compared to ν-SVM, VaR-SVM has a superior out-of-sample performance on datasets with outliers.

1 Introduction

Nowadays, support vector machines (SVMs) are ubiquitous in a variety of applications such as biomedicine, bioinformatics, image recognition, credit scoring, etc. They are closely related to structural risk minimization [17]. The main idea of SVM is to map a training data from two classes into a higher or infinite dimensional space and separate the classes with maximal margin.

The pioneering *hard-margin SVM* [1] imposes "hard-margin" constraints, implying that each training sample is classified correctly, whereas *soft-margin SVM (C-SVM)* [4] involves a parameter C trading off the training error with the margin size. *ν-SVM* [14] replaces the parameter C by a parameter $\nu \in [0, 1]$ specifying an upper bound for the training error, and *extended ν-SVM* (Eν-SVM) [11] extends the admissible range of ν allowing negative margins.

An SVM constructs a separation boundary based on a small portion of training data called *support vectors*, which makes the SVM sensitive to outliers and can result in overfitting. There are several approaches in SVM literature for handling outliers. The fuzzy SVM [9] associates a fuzzy membership with each training sample in C-SVM to reduce the effect of outliers. Robust SVM [15] and center SVM [20] use centers of classes in addition to support vectors to construct a classification boundary. Yet another approach applies a rough set theory to SVM classification to deal with noisy data [10]. A rough based margin SVM was proposed in [19] and incorporates a rough set method into the SVM classifier.

We propose an SVM based on value-at-risk (VaR) that requires the "hard-margin" constraint to hold with probability $\alpha \in (0, 1]$. When $\alpha = 1$, VaR-SVM reduces to the hard-margin SVM. VaR with confidence level α

*University of Florida, Department of Industrial and Systems Engineering, PO Box 116595, 303 Weil Hall, Gainesville, FL 32611-6595, E-mail: tsyurmasto@ufl.edu, uryasev@ufl.edu

†Stevens Institute of Technology, Department of Mathematical Sciences, Castle Point on Hudson, Hoboken, NJ 07030, E-mail: mzabaran@stevens.edu

is the α-percentile of the loss distribution and is widely used to monitor and control the market risk of financial instruments [8, 6]. The similarity between SVM classification and optimization of monetary risk measures was observed in [7, 16, 12]. Computational experiments with artificial datasets confirm that in contrast to ν-SVM, VaR-SVM is stable to outliers.

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

2 Representation 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 labels $+1$ from those with class label -1 in $\mathbb{R}^n$ space.

Let $\Omega = \{\omega_1, \ldots, \omega_l\}$ be a finite sample space with equal[1] 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

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

It can be interpreted as a random variable with realizations $\{-y_i \cdot [w^\top \xi_i + b]\}_{i=1}^l$ assuming equal probabilities $1/l$. 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* $\mathbb{E}\,[X - C]_+$, which is the expected loss exceeding some specified threshold $C \in \mathbb{R}$, where $[\,\cdot\,]_+ = \max\{\cdot, 0\}$.
- *Conditional value-at-risk (CVaR)* $\mathrm{CVaR}_\alpha(X)$, defined as the expected value 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 [1]

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

can be expressed with the worst-case loss functional by

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

The soft-Margin SVM (C-SVM) [4]

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

can be expressed with the partial moment functional as

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

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

The ν-SVM [14], 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 [**?**] 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 [2]. The relationship between ν-SVM and CVaR-minimization was first reported in [7]. In fact, (5) is equivalent to the formulation

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

Both (5) and (7) are convex problems, and their equivalence is established through the duality theory provided that they both have an optimal solution.

Finally, Takeda and Sugiyama [16] showed that Eν-SVM [11] can be formulated by

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

3 VaR-SVM

The SVMs (2), (3), (5), and (8) are sensitive to outliers, since the supremum, partial moment, and CVaR all rely on the right tail of the loss distribution that contains data outliers. 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 outliers 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

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

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

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

With VaR, or percentile function, defined by

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

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

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

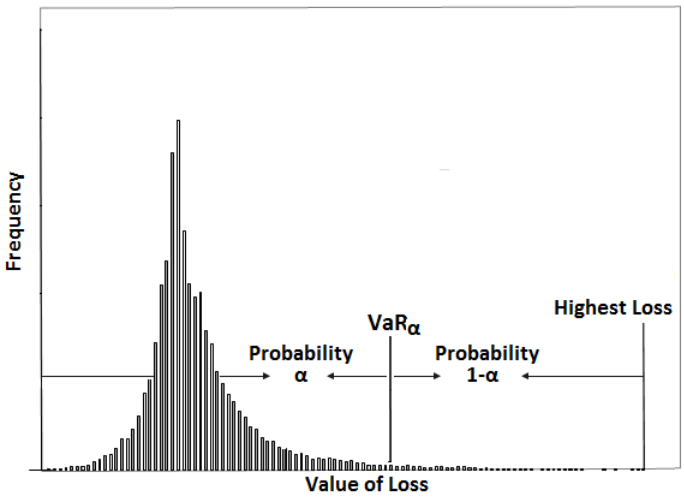

Figure 1: Histogram for the probability density function of the loss distribution: VaR_α corresponds to the α-percentile of the loss distribution.

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

Observe that VaR-SVM (12) resembles (7). Establishing of the equivalence of (5) to (7) relies on CVaR convexity, whereas VaR is not convex and so is the problem (12). Therefore, the similar equivalence of (12) to the unconstrained problem

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

can not be obtained through the duality theory and will be established by other means.

VaR-SVM can be interpreted as minimization of the distances from data samples $\{(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)\}$ to the separating hyperplane aggregated with the percentile function. For each outcome $\omega \in \Omega$ and decision variables $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$,

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

is the Euclidean distance from the vector $\xi(\omega)$ to the hyperplane $H = \{x \in \mathbb{R}^n \,|\, w^\top x + b = 0\}$ in $\mathbb{R}^n$. It has sign -1 when $\xi(\omega)$ is classified correctly i.e. $y(\omega)[w^\top \xi(\omega) + b] > 0$ and $+1$ when $\xi(\omega)$ is classified incorrectly i.e. $y(\omega)[w^\top \xi(\omega) + b] < 0$. The distance (14) assumes realizations

$$-\frac{y_i \cdot [w^T \phi(\xi_i) + b]}{\|w\|}\Bigg|_{i=1}^{l} \tag{15}$$

with equal probabilities $1/l$. Figure 2 shows the histogram of the distance (15) for the data samples $(\phi(\xi_1), y_1)$, $\ldots$, $(\phi(\xi_l), y_l)$ from German Credit Data[2] 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). This problem can be formulated by

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

Next two theorems establish relationships between optimization problems (12), (13), and (16).

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

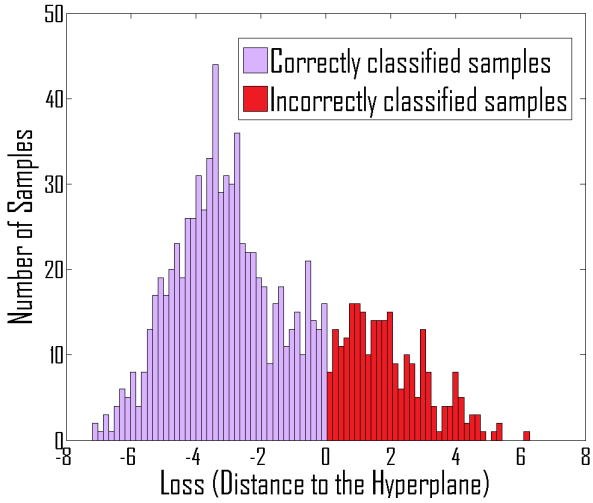

Figure 2: Histogram of the distances (14) of the data samples $(\phi(\xi_1), y_1), \ldots, (\phi(\xi_l), y_l)$ to the separating hyperplane for German Credit Data with some fixed decision variables.

Theorem 1. *If* (w^*, b^*) *is an optimal solution of* (16) *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* (16) *for each* $\lambda > 0$*.*

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

$$\min_{w \in \mathbb{R}^n,\, b \in \mathbb{R}} \mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w, b)}{\|w\|}\right) \qquad \text{s.t.} \qquad \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w, b)) = -1. \tag{17}$$

If (w^*, b^*) is optimal for (16) 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 (16) and also that $(-1/\zeta)\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) = -1$. Consequently, $(-1/\zeta)(w^*, b^*)$ is feasible for (17) and, thus, it is optimal for (17).

Now suppose that (w^*, b^*) is optimal for (17) but is not optimal for (16), 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 (17) 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 (17). □

Also, positive homogeneity of $\mathrm{VaR}_\alpha(\cdot)$ implies that $(\lambda w^*, \lambda b^*)$ is optimal for (16) for each $\lambda > 0$.

Theorem 2. *If* (w^*, b^*) *is an optimal solution of* (16) *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* (16) *for each* $\lambda > 0$*.*

Proof. Let (w^*, b^*) be an optimal solution of (16) such that $\zeta = \mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*, b^*)) < 0$, and let $r^* = -\zeta/\|w^*\| > 0$. 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

$$\frac{1}{2}\|w\|^2 + C\cdot\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b)) = \frac{1}{2}\|w\|^2 + C\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b))}{\|w\|}\|w\| \geq \frac{1}{2}\|w\|^2 - Cr^*\|w\| \geq -\frac{1}{2}(Cr^*)^2. \tag{18}$$

Observe that for $(\tilde{w},\tilde{b}) = \frac{Cr^*}{\|w^*\|}(w^*,b^*)$, the inequality (18) reduces to the equality

$$\frac{1}{2}\|\tilde{w}\|^2 + C\cdot\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w},\tilde{b})) = \frac{1}{2}(Cr^*)^2\frac{\|w^*\|^2}{\|w^*\|^2} + C(Cr^*)\underbrace{\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}}_{=-r^*} = -\frac{1}{2}(Cr^*)^2. \tag{19}$$

If $w = 0$, then the left-hand side of (18) is equal to $C\cdot\mathrm{VaR}_\alpha(-y(\omega)b) = C|b|\,\mathrm{VaR}_\alpha(-y(\omega)\operatorname{sign} b) \geq 0$, since the existence of an optimal solution of (16) 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 (18) 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 (16), 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 (18), we obtain

$$\begin{aligned}\frac{1}{2}\|w^*\|^2 + C\cdot\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w^*,b^*)) &= \frac{1}{2}\|w^*\|^2 + C\cdot\mathrm{VaR}_\alpha\left(\frac{\mathcal{L}_\omega(w^*,b^*)}{\|w^*\|}\right)\|w^*\| \\ &> \frac{1}{2}\|w^*\|^2 + C\cdot\frac{\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w},\tilde{b}))}{\|\tilde{w}\|}\|w^*\| = \frac{1}{2}\|w^*\|^2 - C\tilde{r}\|w^*\| \geq -\frac{1}{2}(C\tilde{r})^2,\end{aligned} \tag{20}$$

where $\tilde{r} = -\mathrm{VaR}_\alpha(\mathcal{L}_\omega(\tilde{w},\tilde{b}))/\|\tilde{w}\|$. Since (w^*, b^*) is optimal for (13), the inequality (20) yields

$$\frac{1}{2}\|w\|^2 + C\cdot\mathrm{VaR}_\alpha(\mathcal{L}_\omega(w,b)) > -\frac{1}{2}(C\tilde{r})^2 \tag{21}$$

for any w and b. Similarly to (19), 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 (21), so that (w^*, b^*) is optimal for (16), and the positive homogeneity of $\mathrm{VaR}_\alpha(\cdot)$ implies that $(\lambda w^*, \lambda b^*)$ is optimal solution of (16) for each $\lambda > 0$ as well. □

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

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

4 Nonlinear VaR-SVM

Convex SVMs are usually solved in dual formulations, where the transformation ϕ is implicitly specified by a kernel function $K(\xi, \xi')$ [13]. However, VaR-SVM is not convex and cannot be solved through its dual. This section shows how to compare VaR-SVM with, for example, 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 features $\{\xi_1, \ldots, \xi_l\} \subset \mathbb{R}^m$ 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., [5]), so that the solution of the primal problem with the transformed features $\{\psi(\xi_1), \ldots, \psi(\xi_l)\} \subset \mathbb{R}^n$ coincides with that for the dual problem with the kernel $K(\xi, \xi')$ corresponding to the original transformation ϕ [3].

For the features $\{\xi_1, \ldots, \xi_l\}$, 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

$$\mathbf{K} = \mathbf{V}\Lambda\mathbf{V}^\top \equiv (\mathbf{V}\Lambda^{\frac{1}{2}})(\mathbf{V}\Lambda^{\frac{1}{2}})^\top, \tag{22}$$

where $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_l\}$ 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 (22) implies that $\psi : \xi_i \to (\mathbf{V}\Lambda^{\frac{1}{2}})_i$, $i = 1, \ldots, l$, is the sought linear transformation, where $(\mathbf{V}\Lambda^{\frac{1}{2}})_i$ is row i of the matrix $(\mathbf{V}\Lambda^{\frac{1}{2}})$. Thus, non-linear SVMs with the kernel K should be compared to VaR-SVM with the transformed features $\{\psi(\xi_1), \ldots, \psi(\xi_l)\}$.

5 Numerical Experiments

Linear VaR-SVM (13) (ϕ is the identity function) and ν-SVM (5) are compared on artificial and real-life datasets with outliers. The computations are performed with Matlab using Portfolio Safeguard (PSG) solver,[3] which uses advanced techniques for optimizing VaR and CVaR, see Section 9.16 in [18]. With PSG, solving of the problems (13) and (5) involves three stages:

1. *Formulating the optimization problem with precoded VaR and CVaR functions.* A typical meta-code uses 5–10 operators (see Appendix A 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.*

5.1 Artificial Dataset

An artificial dataset consists of $ł_1 = 400$ samples (with label $+1$) and $ł_2 = 400$ 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 $\mathbb{R}_{10\times10}$. The outliers are modeled to follow a Gaussian distribution $(\mu_{out}, \Sigma_{out})$ with $\mu_{out} = 100e$, $\Sigma_{out} = 20I$ and having class label $+1$.

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

The percentage of outliers is varied in the range 0%–10% of the original 800 samples. VaR-SVM and ν-SVM are then compared on out-of-sample with the following parameters:

- Fraction of training samples is $2/3$;
- Number of random splitting of the entire dataset into training and testing is 10;
- The parameter ν in (13) is selected from a grid 0:0.0025:1 to maximize out-of-sample performance.

Figure 3 shows out-of-sample (OOS) performance of VaR-SVM (13) and of ν-SVM as a function of the percentage of outliers for the artificial dataset. For a small number of outliers ($< 3\%$), OOS is approximately the same for both classifiers and is close to 1. However, for a larger number of outliers ($> 4\%$), the OOS graphs deviate substantially. The OOS graph for ν-SVM dramatically drops down to almost 0.5 due to sensitivity of ν-SVM to outliers. The OOS graph for VaR-SVM, in contrast, stabilizes at the level of OOS ~ 0.9. The 10% of misclassifications correspond to 10% of outliers. The graph confirms that VaR-SVM is stable to outliers.

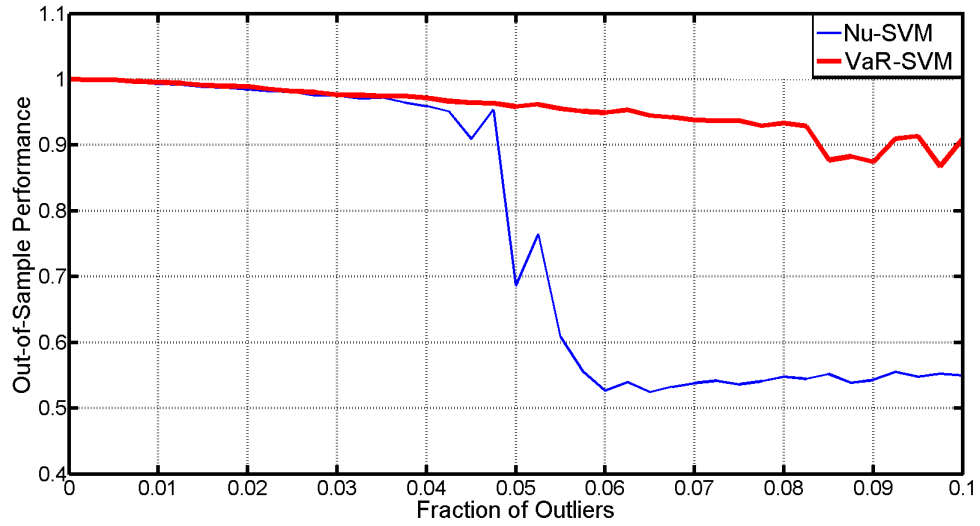

Figure 3: Out-of-Sample (OOS) performance of ν-SVM and of VaR-SVM as a function of the percentage of outliers for the artificial dataset.

5.2 Real-Life Data Sets

The problems (13) and (5) are solved with datasets from UCI Machine Learning Repository[4]: Liver Disorders, Heart Disease, Indian Diabetes, German and Ionosphere. Outliers are generated by artificially multiplying the fraction of 0%, 1%, 5%, and 10% of the original dataset by 1000. Testing accuracy is evaluated with 10-fold cross validation, where 2/3 of the training set is used to solve (13), and the remaining 1/3 is used to fit the parameter ν selected from the grid $0 : 0.05 : 1$.

Tables 2–6 show that as the percentage of outliers increases, the performance of ν-SVM degrades significantly, whereas VaR-SVM's performance is almost unaffected. The original features were normalized (zero mean, unit standard deviation). Table 1 compares running time of VaR-SVM and ν-SVM for different datasets. The running time for VaR-SVM is slightly higher than for ν-SVM, though it is of the same order.

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

Table 1: 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	Solving Time (sec)	
			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	1000	24	2.03	3.07
Ionosphere	796	14	1.76	2.49

Table 2: Experimental results for Liver Disorders Dataset with outliers.

Percent of Ouliers (%)	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	training		testing		training		testing	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
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 3: Experimental results for Heart Disease Dataset with outliers.

Percent of Outliers (%)	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	training		testing		training		testing	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
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

Table 4: Experimental results for German Credit Dataset with outliers.

Percent of Ouliers (%)	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	training		testing		training		testing	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
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

6 Conclusions

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. Compared to ν-SVM, VaR-SVM has a superior out-of-sample performance on artificial and real-life datasets. However, VaR-SVM is non-convex, since VaR is not

Table 5: Experimental results for Indian Diabetes Dataset with outliers.

Percent of Ouliers (%)	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	training		testing		training		testing	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
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 6: Experimental results for Ionosphere Dataset with outliers.

Percent of Ouliers (%)	ν-SVM accuracy (%)				VaR-SVM accuracy (%)			
	training		testing		training		testing	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
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

convex. This calls for the search for convex functionals having similar property of stability to outliers.

Acknowledgments

This research was supported by AFOSR grant FA9550-11-1-0258, New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization.

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,[5] see Problem 1b. For convenience, the PSG meta-code is presented below.

```
1  Problem: problem_var_nu_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 =1000, lowerbounds = −1000
6  Solver: VAN, precision = 6, stages = 6
```

Command **minimize** informs the solver that (13) is a minimization problem, whereas **objective** is a declaration of the objective function defined in lines 3 and 4. The quadratic part of the objective in line 3 is defined by command **quadratic**, and the corresponding data matrix is to be found in file **matrix_quadratic_matrix.txt**.

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

The VaR part of the objective in line 4 is defined by command **var_risk_1**, and the corresponding data matrix is to be found in file **matrix_prior_scenarios.txt**. The coefficients C and α are set to be 0.5.

References

[1] B. Boser, I. Guyon, and V. Vapnik, *A training algorithm for optimal margin classifiers*, in Proceedings of the fifth annual workshop on Computational learning theory, ACM, 1992, pp. 144–152.

[2] D. Burges and Crisp C., *A geometric interpretation of vsvm classifiers*, Advances in Neural Information Processing Systems 12, 12 (2000), p. 244.

[3] O. Chapelle, *Training a support vector machine in the primal*, Neural Computation, 19 (2007), pp. 1155–1178.

[4] C. Cortes and V. Vapnik, *Support-vector networks*, Machine learning, 20 (1995), pp. 273–297.

[5] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.

[6] D. Duffie and J. Pan, *An overview of value at risk*, The Journal of derivatives, 4 (1997), pp. 7–49.

[7] J. Goto and A. Takeda, *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, 2005.

[8] P. Jorion, *Value at risk: the new benchmark for controlling market risk*, vol. 2, McGraw-Hill New York, 1997.

[9] C. Lin and S. Wang, *Fuzzy support vector machines*, Neural Networks, IEEE Transactions on, 13 (2002), pp. 464–471.

[10] P. Lingras and C. Butz, *Rough set based 1-v-1 and 1-vr approaches to support vector machine multi-classification*, Information Sciences, 177 (2007), pp. 3782–3798.

[11] F. Pérez-Cruz, J. Weston, D. Herrmann, and B. Scholkopf, *Extension of the nu-svm range for classification*, NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES, 190 (2003), pp. 179–196.

[12] L. Sakalauskas, A. Tomasgard, and S. Wallace, *Advanced risk measures in estimation and classification*, Proceedings. Vilnius, (2012), pp. 114–118.

[13] B Schölkopf and A. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT press, 2001.

[14] B. Schölkopf, A. Smola, R. Williamson, and P. Barlett, *New support vector algorithms*, Neural computation, 12 (2000), pp. 1207–1245.

[15] Q. Song, W. Hu, and W. Xie, *Robust support vector machine with bullet hole image classification*, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 32 (2002), pp. 440–448.

[16] A. Takeda and M. Sugiyama, *v-support vector machine as conditional value-at-risk minimization*, in Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 1056–1063.

[17] V. Vapnik, *The nature of statistical learning theory*, springer, 1999.

[18] M. Zabarankin and S. Uryasev, *Statistical Decision Problems: Selected Concepts and Portfolio Safeguard Case Studies*, Springer, 2013.

[19] J. Zhang and Y. Wang, *A rough margin based support vector machine*, Information Sciences, 178 (2008), pp. 2204–2214.

[20] X. Zhang, *Using class-center vectors to build support vector machines*, in Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop., IEEE, 1999, pp. 3–11.