SUPPORT VECTOR REGRESSION AS CONDITIONAL VALUE-AT-RISK MINIMIZATION WITH APPLICATION TO FINANCIAL TIME-SERIES ANALYSIS

Akiko Takeda[1], *Jun-ya Gotoh*[2], *and Masashi Sugiama*[3]

[1] Department of Administration Engineering, Keio University, Japan.
[2] Department of Industrial and Systems Engineering, Chuo University, Japan.
[3] Department of Computer Science, Tokyo Institute of Technology, Japan.

ABSTRACT

Support vector regression (SVR) is a popular regression algorithm in machine learning and signal processing. In this paper, we first prove that the SVR algorithm is equivalent to minimizing the *conditional value-at-risk* (CVaR) of the distribution of the ℓ_1-loss residuals, which is a popular risk measure in finance. The equivalence between SVR and CVaR minimization allows us to derive a new upper bound on the ℓ_1-loss generalization error of SVR. Then we show that SVR actually minimizes the upper bound under some condition, implying its optimality. We finally apply the SVR method to an index tracking problem in finance, and develop a new portfolio selection method. Experiments show that the proposed method compares favorably with alternative approaches.

1. INTRODUCTION

Support vector classification (SVC) is a useful classification algorithm in machine learning [1, 2, 3]. SVC separates training samples in different classes by the hyperplane with maximum margin. The maximum margin hyperplane was shown to minimize the Vapnik-Chervonenkis bound, an upper bound of the generalization error [4]. Thus, the generalization performance of SVC is theoretically guaranteed.

Following the success of the SV method in classification, it was extended to be able to handle real-valued outputs, i.e., regression scenarios [5, 2]. The SV regression (SVR) method was shown to perform well, and thus becoming one of the popular data analysis tools in machine learning and signal processing. However, the superior performance of SVR was not completely understood beyond its experimental success. A primal goal of this paper is to provide a novel theoretical insight into the SVR algorithm.

We first prove that SVR is equivalent to minimizing the *conditional value-at-risk* (CVaR) [6, 7] of the distribution of the ℓ_1-loss residuals. The β-CVaR is defined as the mean of the largest $100\beta\%$ residuals (see Figure 1), and is a popular risk measure in risk-sensitive learning, e.g., in the context of financial data analysis. Similar equivalence has been obtained for SVC [8], so the present result can be regarded as generalization of the previous work to regression scenarios.

We then give a novel upper bound of the generalization error measured by the ℓ_1-loss function, where the equivalence between SVR and CVaR minimization plays an important role in its derivation. Under some condition, we can show that the SVR method minimizes the upper bound. This would partially explain the reason why SVR is a superior regression approach.

As an application of the CVaR minimization approach, we consider an index tracking problem in finance and develop a new portfolio selection method. Our SVR-based portfolio selection method is shown to perform well in experiments.

2. SVR AND CVAR MINIMIZATION

Let us consider a regression problem of obtaining a linear function approximator $y = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ from m training samples, $(\boldsymbol{x}_i, y_i), i \in M := \{1, \ldots, m\}$, where $\boldsymbol{x}_i \in \mathbb{R}^n$ is an input point, $y_i \in \mathbb{R}$ is an output value, and $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are parameters to be learned.

In this section, we briefly review the definition of SVR and CVaR minimization methods.

2.1. SVR

In the ϵ-SVR framework [5], $\boldsymbol{w}$ and b are determined so that the following regularized empirical risk is minimized:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + \frac{C'}{m} \sum_{i \in M} [|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| - \epsilon]^+,$$

where ϵ is a positive constant and $[X]^+ := \max\{X, 0\}$. $[|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b| - \epsilon]^+$ is called the Vapnik's ϵ-insensitive loss function [4]. $C' > 0$ is a regularization constant controlling the trade-off between the goodness-of-fit and the complexity of the model. The positive parameter ϵ controls the sensitivity to noise in the training data. A potential

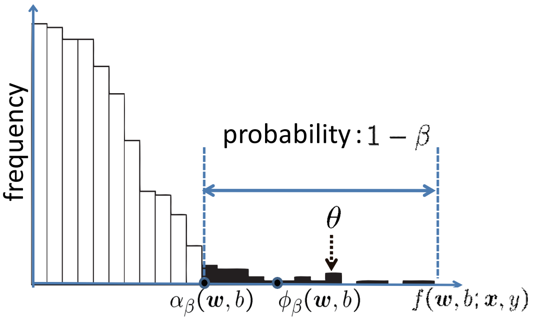
frequency
probability : $1-\beta$
θ
$\alpha_\beta(\boldsymbol{w},b)$
$\phi_\beta(\boldsymbol{w},b)$
$f(\boldsymbol{w},b;\boldsymbol{x},y)$

Fig. 1. Illustration of CVaR $\phi_\beta(\boldsymbol{w},b)$ and VaR $\alpha_\beta(\boldsymbol{w},b)$.

weakness of the ϵ-SVR formulation is that the choice of ϵ is not intuitive beyond the role as the insensitive zone.

Another formulation of SVR called ν-SVR was proposed in [2], which automatically determines ϵ based on another tuning parameter ν. In ν-SVR, $\boldsymbol{w}$ and b are learned as follows:

$$\begin{array}{ll}\min\limits_{\boldsymbol{w},b,\alpha,\boldsymbol{z}} & \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}+C'\left(\nu\alpha+\frac{1}{m}\sum_{i\in M} z_i\right) \\ \text{subject to} & z_i-|y_i-\langle\boldsymbol{w},\boldsymbol{x}_i\rangle-b|+\alpha\geq 0, \\ & z_i\geq 0,\quad i\in M.\end{array} \tag{1}$$

By setting $\nu=0$ and $\alpha=\epsilon$, the formulation (1) is reduced to the ϵ-SVR formulation.

An advantage of the ν-SVR formulation is that the meaning of ν is intuitive [2]: ν is an upper bound of the fraction of margin errors (i.e., the samples incurring positive z_i^*, $i\in M$), and is a lower bound of the fraction of SVs (i.e., the non-zero elements of the dual solution).

2.2. CVaR Minimization

Let $f(\boldsymbol{w},b;\boldsymbol{x},y)$ be some loss function. Figure 1 illustrates the distribution of loss $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$. Let $\Phi(\alpha|\boldsymbol{w},b)$ be the cumulative distribution function of $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$, $\forall i\in M$:

$$\Phi(\alpha|\boldsymbol{w},b):=\frac{1}{m}|\{i\in M: f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)\leq\alpha\}|.$$

Its 100β-percentile ($\beta\in(0,1)$), denoted by $\alpha_\beta(\boldsymbol{w},b)$, is called the *$\beta$-value-at-risk* (VaR) (see Figure 1 again) [9, 6]:

$$\alpha_\beta(\boldsymbol{w},b):=\min\{\,\alpha:\Phi(\alpha\,|\boldsymbol{w},b)\geq\beta\,\}.$$

The conditional value-at-risk (CVaR), denoted by $\phi_\beta(\boldsymbol{w},b)$, is defined as the average loss exceeding the VaR $\alpha_\beta(\boldsymbol{w},b)$, i.e., the average of $f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)$ in the shaded region in Figure 1.

Now let us consider minimizing the CVaR $\phi_\beta(\boldsymbol{w},b)$ with respect to $(\boldsymbol{w},b)$. According to [6], $\min_{\boldsymbol{w},b}\phi_\beta(\boldsymbol{w},b)$ is equivalently expressed as

$$\min_{\boldsymbol{w},b,\alpha}\ \alpha+\frac{1}{(1-\beta)m}\sum_{i=1}^{m}[f(\boldsymbol{w},b;\boldsymbol{x}_i,y_i)-\alpha]^+. \tag{2}$$

More precisely, the solutions $(\boldsymbol{w}^*,b^*)$ of these two problems are the same, and their optimal values $\phi_\beta(\boldsymbol{w}^*,b^*)$ are also the same. The solution α^* of (2), obtained as a by-product, is equal to the β-VaR $\alpha_\beta(\boldsymbol{w}^*,b^*)$ under some condition [6].

3. SVR AS CVAR MINIMIZATION

In this section, we first show that SVR can be interpreted as minimizing CVaR. Then, based on this interpretation, we derive an upper bound of the generalization error of SVR.

3.1. Equivalence between SVR and CVaR Minimization

For some $C>0$, let us consider the following loss function:

$$f(\boldsymbol{w},b;\boldsymbol{x},y)=\left|y-C\frac{\langle\boldsymbol{w},\boldsymbol{x}\rangle+b}{\|\boldsymbol{w}\|}\right|. \tag{3}$$

Let $\boldsymbol{w}^*$ and b^* be the solution of CVaR minimization (2) for the loss (3), with which we construct a regressor as

$$h(\boldsymbol{x})=\langle\frac{C}{\|\boldsymbol{w}^*\|}\boldsymbol{w}^*,\boldsymbol{x}\rangle+\frac{C}{\|\boldsymbol{w}^*\|}b^*. \tag{4}$$

Then we have the following lemma.

Lemma 1 *Let $C_\beta:=\|\widetilde{\boldsymbol{w}}_\beta\|$, where $\widetilde{\boldsymbol{w}}_\beta$ is the solution of*

$$\min_{\boldsymbol{w},b,\alpha}\ \alpha+\frac{1}{(1-\beta)m}\sum_{i=1}^{m}[|y_i-(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)|-\alpha]^+. \tag{5}$$

When $\beta\in(0,1)$ and $C\leq C_\beta$, the CVaR minimization problem (2) for the loss (3) *is equivalent to*[1]

$$\begin{array}{ll}\min\limits_{\boldsymbol{w},b,\alpha,\boldsymbol{z}} & \alpha+\frac{1}{(1-\beta)m}\sum_{i\in M} z_i \\ \text{subject to} & z_i-|y_i-\langle\boldsymbol{w},\boldsymbol{x}_i\rangle-b|+\alpha\geq 0, \\ & z_i\geq 0,\quad i\in M,\quad \boldsymbol{w}^\top\boldsymbol{w}\leq C^2.\end{array} \tag{6}$$

A sketch of proof of the above lemma is given in Appendix A. This lemma implies that the problem (2) for the loss (3) can be interpreted as the CVaR minimization problem with the loss $f(\boldsymbol{w},b;\boldsymbol{x},y)=|y-(\langle\boldsymbol{w},\boldsymbol{x}\rangle+b)|$ and the region of $\boldsymbol{w}$ restricted to $\boldsymbol{w}^\top\boldsymbol{w}\leq C^2$.

Based on the above lemma, we can easily show that the problem has the following properties (proofs are omitted due to lack of space).

Lemma 2 *The optimal value of (6), $\phi_\beta(\boldsymbol{w}^*,b^*)$, is strictly decreasing as C increases or as β decreases.*

[1] When $C>C_\beta$, (2) for the loss (3) is equivalent to (6) with the constraint $\boldsymbol{w}^\top\boldsymbol{w}\leq C^2$ replaced by $\boldsymbol{w}^\top\boldsymbol{w}=C^2$, which is not convex.

Theorem 3 *The CVaR minimization problem (2), equivalently (6), is the same as ν-SVR (1) under $\beta = 1 - \nu$ and*

$$C = (C'\nu)\sqrt{\sum_{i,j \in M} (\bar{\lambda}_i^{(1)} - \bar{\lambda}_i^{(2)})(\bar{\lambda}_j^{(1)} - \bar{\lambda}_j^{(2)})\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle} = \|\bar{\boldsymbol{w}}\|,$$

where $\bar{\boldsymbol{\lambda}}^{(1)}$ and $\bar{\boldsymbol{\lambda}}^{(2)}$ are the solutions of the dual ν-SVR. $\bar{\lambda}_i^{(1)}$ and $\bar{\lambda}_i^{(2)}$ correspond to the constraints $z_i - (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) + \alpha \geq 0$ and $z_i + (y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) + \alpha \geq 0$, respectively, and $\bar{\boldsymbol{w}}$ is the solution of the primal ν-SVR.

By comparing the dual problem of (6) and that of ν-SVR (1), we find that both dual problems have the same solution under the above-mentioned parameter transformation, although there is a gap between their optimal values.

3.2. Generalization Performance of ν-SVR

Here, we derive a new bound of the generalization error of ν-SVR based on the notion of CVaR (a sketch of proof of the following theorem is given in Appendix B).

Theorem 4 *Suppose that the training samples $(\boldsymbol{x}_i, y_i), i \in M$, are drawn independently from a distribution P, whose support lives in the ball of radius B centered at the origin. Then, for any $\nu \in (0,1)$ and any $(\boldsymbol{w}, b)$ such that $\|\boldsymbol{w}\| \leq C > 0$, with probability at least $1 - \delta$ over the training set, the probability that $f(\boldsymbol{w}, b; \boldsymbol{x}, y) = |y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)|$ is larger than a threshold θ is bounded as*

$$\mathbb{P}\{f(\boldsymbol{w}, b; \boldsymbol{x}, y) > \theta\} \leq \nu + G(\theta - \phi_{1-\nu}(\boldsymbol{w}, b)) \text{ if } \phi_{1-\nu}(\boldsymbol{w}, b) < \theta, \quad (7)$$

$$\mathbb{P}\{f(\boldsymbol{w}, b; \boldsymbol{x}, y) > \theta\} \leq \nu + G(\theta - \alpha_{1-\nu}(\boldsymbol{w}, b)) \text{ if } \alpha_{1-\nu}(\boldsymbol{w}, b) < \theta, \quad (8)$$

$$\mathbb{P}\{f(\boldsymbol{w}, b; \boldsymbol{x}, y) > \theta\} \geq \nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b) - \theta) \text{ if } \alpha_{1-\nu}(\boldsymbol{w}, b) > \theta, \quad (9)$$

where $\mathbb{P}$ denotes the probability with respect to $(\boldsymbol{x}, y)$ over the distribution P, and

$$G(\gamma) := 2\sqrt{\frac{2}{m}\left(\frac{\kappa}{\gamma^2}\log_2(2m) - 1 + \ln\left(\frac{2}{\delta}\right)\right)},$$
$$\kappa := 4c^2(B + \theta + 1)^2\left\{C^2 + B^2(C^2 + 1) + 1\right\}.$$

When $\phi_{1-\nu}(\boldsymbol{w}^*, b^*) < \theta$, the solution $(\boldsymbol{w}^*, b^*)$ of ν-SVR minimizes the upper bound (7) because of the following facts: (a) $G(\gamma)$ is strictly decreasing as $|\gamma|$ increases. (b) The upper bound $\nu + G(\theta - \phi_{1-\nu}(\boldsymbol{w}, b))$ is lowered if $\phi_{1-\nu}(\boldsymbol{w}, b)$ is reduced. (c) The solution $(\boldsymbol{w}^*, b^*)$ of ν-SVR minimizes $\phi_{1-\nu}(\boldsymbol{w}, b)$ under the constraint $\|\boldsymbol{w}\| \leq C$.

Similarly, when $\alpha_{1-\nu}(\boldsymbol{w}, b) < \theta$ or $\alpha_{1-\nu}(\boldsymbol{w}, b) > \theta$, the upper bound (8) or the lower bound (9) is lowered if $\alpha_{1-\nu}(\boldsymbol{w}, b)$ is reduced. Since $\alpha_{1-\nu}(\boldsymbol{w}, b) \leq \phi_{1-\nu}(\boldsymbol{w}, b)$ always holds (see Figure 1), minimizing $\phi_{1-\nu}(\boldsymbol{w}, b)$ by ν-SVR may lower the bounds (8) and (9).

4. APPLICATION: PORTFOLIO SELECTION BY INDEX TRACKING

We have shown that SVR is equivalent to CVaR minimization, and SVR minimizes a bound of the generalization error under some condition. In this section, we apply the CVaR minimization idea to financial data analysis, and develop a new algorithm.

4.1. Background

The problem of allocating funds to a given set of investable assets is known as *portfolio selection* in finance. In his seminal paper [10], Markowitz assumed that only the expectation and variability of return (i.e., mean and variance) matter to investors, where the variance plays a role as a measure of the risk.

Although the variance would be the most fundamental risk measure to be minimized, it has several drawbacks. A critical one is that the variance is affected by the deviation in the direction of both profit and loss—in reality, we only want to suppress the deviation in the direction of loss. A partial risk measure such as VaR, which gained popularity in the 1990s in finance, can account for a large loss with a small probability. More recently, CVaR has been growing in popularity, as it can overcome critical shortcomings of VaR such as non-convexity.

In financial markets, investment strategies can be divided into *passive* investment strategies and *active* investment strategies. Investors who adopt active investment strategies aggressively exchange assets so that they can constantly find profit opportunities. Active investors take it for granted that they can beat markets continuously. On the other hand, investors who adopt passive investment strategies conservatively consider that they cannot continuously go beyond the average level of market.

Index tracking investment is a kind of passive investment strategy: investors purchase all or some of the assets contained in a market index, and construct a portfolio that tracks the market index. Since the market index is considered as a benchmark, the investors expect to obtain a similar return to that of the benchmark through the index tracking investment.

In this section, we propose a new index tracking portfolio selection model based on SVR. An investor's wealth is allocated among n risky assets which are component stocks contained in the asset market index.

We will use the following notations: Let $\boldsymbol{R}_t := (R_{1,t}, \ldots, R_{n,t})^\top$ be an observed historical return vector of the n assets at time t ($t = 1, \ldots, T$), $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top$ be a decision vector which shows the proportion of the total amount of money devoted to asset i ($i = 1, \ldots, n$), and I_t be observed market index return at time t ($t = 1, \ldots, T$).

4.2. Traditional Models

The problem of portfolio selection by index tracking has been formulated as a linear regression problem, where a linear model $I = \boldsymbol{R}^\top \boldsymbol{\pi}$ (with parameter $\boldsymbol{\pi}$) is estimated based on a given set of observed data $(I_t, \boldsymbol{R}_t)$, $t = 1, \ldots, T$. For some $\delta > 0$, the *tracking error* is defined as

$$g_\delta(\boldsymbol{\pi}) := \frac{1}{T}\left[\sum_{t=1}^{T} |I_t - \boldsymbol{R}_t^\top \boldsymbol{\pi}|^\delta\right]^{(1/\delta)}.$$

Roll [11] used $g_2(\boldsymbol{\pi})$ for portfolio selection. More precisely, the squared error $g_2(\boldsymbol{\pi})$ is minimized with respect to $\boldsymbol{\pi}$ subject to $\boldsymbol{\pi} \in \Pi$, where $\Pi = \{\boldsymbol{\pi} : \sum_{i=1}^{n} \pi_i = 1,\ \ \pi_i \geq 0,\ \ i = 1, \ldots, n\}$. The non-negativity constraint is called the *short sale constraint*, meaning that the investor cannot sell an asset that the investor does not own. The squared-error formulation results in a quadratic program, which is the same as the standard Lasso regression subject to $\boldsymbol{\pi} \in \Pi$, and thus the solution can be obtained by a standard optimization software. We refer to this method as ‘*Sqr*’.

Instead of the squared deviations $g_2(\boldsymbol{\pi})$, the absolute error $g_1(\boldsymbol{\pi})$ is also a popular choice (e.g., [12, 13, 14]). This is advantageous in that the optimization problem is reduced to a linear program, which can be more efficiently solved than quadratic programs. Furthermore, the absolute error formulation is more robust against outliers than the squared error formulation [15]. We refer to this method as ‘*Abs*’.

Alternatively, the ∞-norm was used in [14], which results in the minmax measure: $\max_t |I_t - \boldsymbol{R}_t^\top \boldsymbol{\pi}|$. The optimization problem still yields a linear program, but the solution may be too conservative due to the minmax nature. We refer to this method as ‘*Minmax*’.

The β-CVaR formulation would be a useful alternative to the minmax formulation since it bridges the absolute error and the minmax approaches by the parameter β:

$$\min_{\boldsymbol{\pi}\in\Pi,\alpha} \alpha + \frac{1}{(1-\beta)T}\sum_{t=1}^{T}[|I_t - \boldsymbol{R}_t^\top \boldsymbol{\pi}| - \alpha]^+. \qquad (10)$$

Indeed, if the parameter β is sufficiently close to 1, the problem (10) is reduced to the minmax formulation, and if β goes to 0, it agrees with the absolute error formulation (see also Figure 1). Furthermore, the CVaR minimization problem is still a linear program, so it can be solved efficiently. However, naively minimizing CVaR may result in overfitting if the number of training samples is small. We refer to this method as ‘*CVaR*’.

4.3. Proposed Model

Here, we propose to solve the following SVR-type problem:

$$\min_{\boldsymbol{\pi}\in\Pi,\alpha} \alpha + \frac{1}{(1-\beta)T}\sum_{t=1}^{T}\left[\left|I_t - C\frac{\boldsymbol{R}_t^\top \boldsymbol{\pi}}{\|\boldsymbol{\pi}\|}\right| - \alpha\right]^+. \quad (11)$$

As shown in Lemma 1, this is equivalent to

$$\begin{array}{ll} \displaystyle\min_{\boldsymbol{\pi}\in\Pi,\alpha,\boldsymbol{z}} & \alpha + \frac{1}{(1-\beta)T}\sum_{t=1}^{T} z_t \\ \text{subject to} & z_t - |I_t - \boldsymbol{R}_t^\top \boldsymbol{\pi}| + \alpha \geq 0,\ \ z_t \geq 0,\ \forall t, \\ & \|\boldsymbol{\pi}\| \leq C. \end{array}$$

We refer to this formulation as the *norm-constrained CVaR (NCCVaR) deviation model*.

Let $\mathcal{I}$ be the return of a target asset to be mimicked, and let each component of $\boldsymbol{\mathcal{R}}$ represent the rate of return of each asset. We regard $(\mathcal{I}, \boldsymbol{\mathcal{R}})$ as a random variable. Suppose that $(\mathcal{I}, \boldsymbol{\mathcal{R}})$ has a bounded support in a ball of radius B centered at the origin, and that $(I_1, \boldsymbol{R}_1), \ldots, (I_T, \boldsymbol{R}_T)$ are independently drawn from a distribution P. We then immediately have the following corollary from Theorem 4.

Corollary 5 *For any feasible portfolio $\boldsymbol{\pi}$ satisfying $\|\boldsymbol{\pi}\| \leq C$, the probability of the tracking error being greater than a threshold θ, $\mathbb{P}\{|\mathcal{I} - \boldsymbol{\mathcal{R}}^\top \boldsymbol{\pi}| > \theta\}$, is bounded as (7)–(9) with probability at least $1 - \delta$, where*

$$G(\gamma) := 2\sqrt{\frac{2}{T}\left\{\frac{4c^2(B{+}\theta)^2(C^2{+}1)}{\gamma^2}\log_2(2T){+}\ln\frac{2}{\delta \mathrm{e}}\right\}}.$$

At a glance, the difference between the plain CVaR model (10) and the proposed NCCVaR model (11) seems rather minor—we just included the additional norm constraint $\|\boldsymbol{\pi}\| \leq C$ in NCCVaR. However, this small difference is highly fruitful in two respects. One is that the NCCVaR formulation has the theoretical guarantee as shown in the above corollary, while no theoretical generalization bound exists for the plain CVaR, to the best of our knowledge. Another is the experimental performance. Since the norm constraint essentially works as a regularizer, NCCVaR works better than plain CVaR, as shown next.

4.4. Experiments

First, we report experimental results on an artificial data set. We randomly generated $\boldsymbol{R}_t \in \mathbb{R}^n$, $t = 1, \ldots, T$ ($n = 100$ and $T = 120$), following a *regime-switching model* [16] with a low-return low-volatility regime and a high-return high-volatility regime. More specifically, each element of $\boldsymbol{R}_t$ were independently drawn from $N(\mu_{\text{low}}, \sigma_{\text{low}}^2)$ or $N(\mu_{\text{high}}, \sigma_{\text{high}}^2)$, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. For each element of $\boldsymbol{R}_t$, the parameters μ_{low}, μ_{high}, σ_{low}, and σ_{high} were drawn from $U(0, 2)$, $U(3, 5)$, $U(7, 10)$, and $U(10, 13)$, respectively, where $U(a, b)$ denotes the uniform distribution on $[a, b]$. The true index I_t at time t was set to the mean of the elements of $\boldsymbol{R}_t$. The initial regime was chosen randomly. The regimes change from time t to time $t + 1$ with probability 0.05; otherwise the regime stays unchanged.

Let $\widehat{\boldsymbol{\pi}}$ be a decision vector learned from the T training samples. In NCCVaR (11), the parameters (β, C) were systematically tuned as follows. Using the first $\frac{5}{6}T$-period of

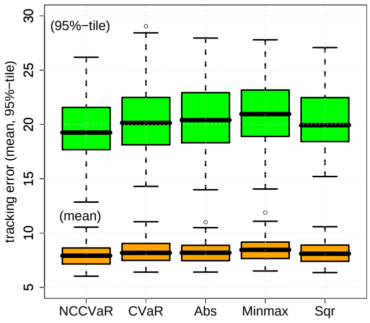

Fig. 2. Test error (artificial data).

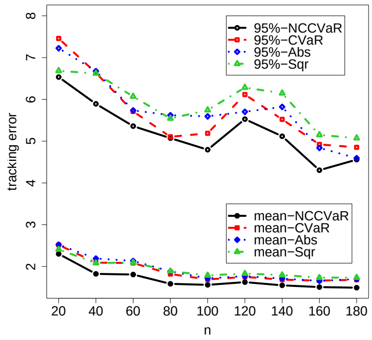

Fig. 3. Test error (Nikkei monthly).

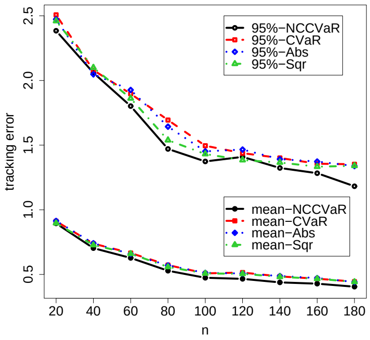

Fig. 4. Test error (Nikkei weekly).

the training samples, we obtained the pair (β, C) that gave the best prediction for the remaining $\frac{1}{6}T$-period. β was chosen from $\{0.1, 0.3, \ldots, 0.9\}$, and C was chosen from $1/\sqrt{n} + k(1 - 1/\sqrt{n})/(10\sqrt{n})$, $k = 1, \ldots, 5$.

The prediction performance of the learned vector $\widehat{\boldsymbol{\pi}}$ was evaluated for the $(T+1)$-th sample (generated in the same way as the training samples) as $|I_{T+1} - \boldsymbol{R}_{T+1}^{\top}\widehat{\boldsymbol{\pi}}|$. Changing the random seed, we generated the $(T+1)$-th sample 100 times and evaluated the mean and the 95-percentile of the above test error. We repeated this experimental procedure 100 times, and evaluated the distributions of the mean and the 95-percentile of the test error over 100 runs. Figure 2 depicts the box plots of the experimental results, showing that NCCVaR compares favorably with other methods in particular in the 95-percentile.

Next, we report experimental results on real financial market data: *monthly and weekly return data of stocks listed in the Nikkei 225 index*. The monthly data set consists of returns of 182 companies during the 270 consecutive months between May 1987 and October 2009, whereas the weekly data set consists of returns of the 182 companies during the 1178 consecutive weeks from April 12, 1987 to November 1, 2009. These monthly and weekly returns of the Nikkei 225 index are our target to be tracked.

We randomly chose n $(= 20, 40, \ldots, 180)$ assets from the 182 assets, and designed a portfolio $\widehat{\boldsymbol{\pi}}_t$ using historical data $\boldsymbol{R}_t, \ldots, \boldsymbol{R}_{t+T-1}$ for $T = 120$ (10 years) consecutive periods from the monthly data set or for $T = 150$ (almost 3 years) consecutive periods from the weekly data set. We evaluated the test error $|I_{t+T} - \boldsymbol{R}_{t+T}^{\top}\widehat{\boldsymbol{\pi}}_t|$ for the next-step sample $(I_{t+T}, \boldsymbol{R}_{t+T})$. This procedure was repeated for $t = 1, \ldots, \overline{T}$ ($\overline{T} = 150$ for the monthly data set and $\overline{T} = 1028$ for the weekly data set). The test error $\frac{1}{\overline{T}} \sum_{t=1}^{\overline{T}} |I_{t+T} - \boldsymbol{R}_{t+T}^{\top}\widehat{\boldsymbol{\pi}}_t|$ was employed as a performance measure.

Figure 3 depicts the mean and the 95-percentile of the test error for the monthly data, showing that NCCVaR gave the lowest test error for almost all cases (Minmax was omitted since it performed very poorly). NCCVaR dominated over the other methods also for the weekly data (see Figure 4). From these results, we conclude that the proposed NCCVaR is a useful alternative to the existing methods.

5. CONCLUSIONS

In this paper, we showed that the popular SVR algorithm is equivalent to minimizing the *conditional value-at-risk* (CVaR). This finding allowed us to derive a new upper bound of the generalization error. We showed that SVR actually minimizes the upper bound under some condition, implying its optimality. We then applied the SVR method to portfolio selection based on index tracking, and showed the proposed method has a theoretical performance guarantee and also performs well in experiments.

As shown in the proof of Theorem 4, a sharper generalization error bound can be obtained in terms of the VaR $\alpha_{1-\nu}(\boldsymbol{w}, b)$. Thus, directly minimizing the VaR, instead of the CVaR, would be theoretically more favorable. However, minimizing the VaR is known to be hard since the resulting optimization problem is non-convex. Thus, developing a powerful optimization algorithm for better solving the VaR minimization problem would be a challenging and promising future direction to be pursued.

6. REFERENCES

[1] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*. 1992, pp. 144–152, ACM Press.

[2] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett, “New support vector algorithms,” *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[3] F. Perez-Cruz, J. Weston, D. J. L. Hermann, and B. Schölkopf, “Extension of the ν-SVM range for classification,” in *Advances in Learning Theory: Methods, Models and Applications 190*, pp. 179–196. IOS Press, Amsterdam, 2003.

[4] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 1995.

[5] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*. 1997, pp. 155–161, The MIT Press.

[6] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *Journal of Banking & Finance*, vol. 26, no. 7, pp. 1443–1472, 2002.

[7] H. Kashima, "Risk-sensitive learning via minimization of empirical conditional value-at-risk," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 12, pp. 2043–2052, 2007.

[8] A. Takeda and M. Sugiyama, "Nu-support vector machine as conditional value-at-risk minimization," in *Proceedings of 25th Annual International Conference on Machine Learning*, 2008, pp. 1056–1063.

[9] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[10] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[11] R. Roll, "A mean/variance analysis of tracking error," *Journal of Portfolio Management*, vol. 18, no. 4, pp. 13–22, 1992.

[12] M. Gilli and E. Këllezi, "The threshold accepting heuristic for index tracking," in *Financial Engineering, E-Commerce, and Supply Chain*, pp. 1–18. Kluwer Academic, Dordrecht, 2002.

[13] J.-L. Prigent, *Portfolio Optimization and Performance Analysis*, Chapman & Hall/CRC, Boca Raton, 2007.

[14] M. Rudolf, H. J. Wolter, and H. Zimmermann, "A linear model for tracking error minimization," *Journal of Banking & Finance*, vol. 23, no. 1, pp. 85–103, 1999.

[15] M. Sugiyama, H. Hachiya, H. Kashima, and T. Morimura, "Least absolute policy iteration for robust value function approximation," in *Proceedings of 2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 2904–2909.

[16] M. R. Hardy, "A regime-switching model of long-term stock returns," *North American Actuarial Journal*, vol. 5, no. 2, pp. 41–53, 2001.

[17] A. Takeda, "Support vector machine based on conditional value-at-risk minimization," Tech. Rep. B-439, Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, 2007.

A. SKETCH OF PROOF OF LEMMA 1

Suppose $C \leq C_\beta$. If $\boldsymbol{w}^\top \boldsymbol{w} < C^2$ holds at optimality, we can construct a feasible solution with a smaller objective value than the optimal value (see [17] for details). However, this contradicts the optimality of (6). Therefore, by contradiction, the solution of (6) satisfies $\boldsymbol{w}^\top \boldsymbol{w} = C^2$. The variable z_i of (6) corresponds to $[f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i) - \alpha]^+$ of (2) with the loss (3). The scaling of the solutions is different, but their (normalized) regressors $h(\boldsymbol{x})$ defined by (4) are the same. □

B. SKETCH OF PROOF OF THEOREM 4

In this proof, we use the following lemma, which holds when $y = \pm 1$ (i.e., classification) and $b = 0$ (i.e., homogeneous):

Lemma 6 *[2] Suppose that the margin $\widetilde{\gamma} > 0$ and the support $\mathcal{X}$ is in a centered ball of radius $\widetilde{R}$. Then, for all $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| \leq 1$, there exists a positive constant c such that the following bound holds with probability at least $1 - \delta$:*

$$\mathbb{P}\{y(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) < 0\} \leq \frac{|\{i : y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle < \widetilde{\gamma}\}|}{m} + \sqrt{\frac{2}{m}\left(\frac{4c^2\widetilde{R}^2}{\widetilde{\gamma}^2}\log_2(2m) - 1 + \ln\frac{2}{\delta}\right)}.$$

In order to fit our CVaR minimization (6) to the setup of the above lemma, let us express

$$\begin{aligned}&\mathbb{P}\{\theta - f(\boldsymbol{w}, b; \boldsymbol{x}, y) < 0\}\\ &= \mathbb{P}\{\theta - y + \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b < 0\} + \mathbb{P}\{\theta + y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle - b < 0\}.\end{aligned}$$

Then, the function $\theta - f(\boldsymbol{w}, b; \boldsymbol{x}, y)$ can be expressed as $\langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}}^{(1)} \rangle$ or $\langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{x}}^{(2)} \rangle$, where

$$\begin{aligned}&\widetilde{\boldsymbol{w}} = \begin{pmatrix}\boldsymbol{w}^\top & b & 1\end{pmatrix}^\top, \ \widetilde{\boldsymbol{x}}^{(1)} = \begin{pmatrix}\boldsymbol{x}^\top & 1 & \theta - y\end{pmatrix}^\top,\\ &\widetilde{\boldsymbol{x}}^{(2)} = \begin{pmatrix}-\boldsymbol{x}^\top & -1 & \theta + y\end{pmatrix}^\top.\end{aligned}$$

We begin with the case where $\alpha_{1-\nu}(\boldsymbol{w}, b) < \theta$. Let us consider the distribution of $(\theta - f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i))/\|\widetilde{\boldsymbol{w}}\|$, $i \in M$ against the "margin" $\gamma := (\theta - \alpha_{1-\nu}(\boldsymbol{w}, b))/\|\widetilde{\boldsymbol{w}}\|$. Then the property of $(1 - \nu)$-CVaR (see [6]) provides

$$\begin{aligned}&\frac{1}{m}\left|\left\{ i \ : \frac{\theta - f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)}{\|\widetilde{\boldsymbol{w}}\|} < \gamma \right\}\right|\\ &= \frac{1}{m}|\{ i \ : \alpha_{1-\nu}(\boldsymbol{w}, b) < f(\boldsymbol{w}, b; \boldsymbol{x}_i, y_i)\}| \leq \nu.\end{aligned}$$

Since $\|\widetilde{\boldsymbol{w}}\| \leq \sqrt{C^2 + B^2(C^2 + 1) + 1}$ and $\|\widetilde{\boldsymbol{x}}^{(i)}\| \leq B + \theta + 1$, Lemma 6 yields that $\mathbb{P}\{f(\boldsymbol{w}, b; \boldsymbol{x}, y) > \theta\}$ is upper-bounded by $\nu + G(\theta - \alpha_{1-\nu}(\boldsymbol{w}, b))$.

In the same way, the upper bound $\nu + G(\theta - \phi_{1-\nu}(\boldsymbol{w}, b))$ when $\phi_{1-\nu}(\boldsymbol{w}, b) < \theta$ can be obtained by using $\alpha_{1-\nu}(\boldsymbol{w}, b) \leq \phi_{1-\nu}(\boldsymbol{w}, b)$.

Finally, we consider the case where $\alpha_{1-\nu}(\boldsymbol{w}, b) > \theta$. In this case, γ is negative although it should be positive in Lemma 6. In order to resolve this issue, let us consider an "inverted" classifier $-\theta + f(\boldsymbol{w}, b; \boldsymbol{x}, y)$, whose sign is opposite of that of $\theta - f(\boldsymbol{w}, b; \boldsymbol{x}, y)$. Applying Lemma 6 to

$$\mathbb{P}\{-\theta + f(\boldsymbol{w}, b; \boldsymbol{x}, y) < 0\} = 1 - \mathbb{P}\{\theta - f(\boldsymbol{w}, b; \boldsymbol{x}, y) < 0\},$$

we obtain an upper bound in the same way as the above case, leading to the lower bound $\nu - G(\alpha_{1-\nu}(\boldsymbol{w}, b) - \theta)$. □