基于加权支持向量机的VaR计算方法研究*

胡 莹, 王安民
(西安电子科技大学 经济管理学院, 陕西 西安 710071)

摘 要 针对统计学框架下传统VaR计算方法的不足, 发展了基于加权支持向量机(W—SVM)的VaR计算新方法. 为了在VaR模型中计入金融时间序列的记忆效应, 采用最优市场因子作为支持向量机的加权模型. 对2001—2009年上证综指的实证研究表明, 基于W-SVM的VaR模型优于传统的VaR方法, 在小样本、厚尾、非线性及有异常波动的市场条件下, 各种置信度下的W-SVM方法均能取得较好的性能.

关键词 在险风险值; 支持向量机; 概率密度估计; 上证综指
中图分类号 F830.9 文献标识码: A

1 引 言

现代投资组合风险管理高度依赖于定量技术来描述金融市场的行为. 风险管理者一方面通过20世纪70年代发展起来的金融衍生品进行了许多风险管理的创新; 另一方面也创立了许多用于识别和量化风险的高级风险管理模型. 其中, Value at Risk (VaR)已成为当今最为流行的风险管理技术. 由于VaR方法能简单清晰地表示金融资产的市场风险大小, 又有比较严格系统的统计理论作为基础, 因此得到了国际金融理论和实业界的广泛认可[1].

国内外陆续提出了多种VaR计算方法. 大致上可以分为参数方法(Delta-Gamma法、GARCH系列模型等)、非参数方法(历史模拟法、Monte Carlo法、Bootstep区间估计等)和半参数方法(极值理论法、分位数回归模型、核估计法等)[2-3]. 这些方法绝大部分都是基于统计理论, 需要大量的样本数据, 因而在金融市场预测时总是存在这样和那样的不足.

支持向量机(Support Vector Machine, SVM)是在统计学习理论的基础上发展起来的一种新的学习方法. 由于其完备的理论基础、出色的学习性能及预测性能, 该技术已成为机器学习界的研究热点[4]. 除了回归估计和模式识别领域, SVM在概率密度估计领域也有良好的应用前景. 而VaR计算的关键正是市场因子的概率密度估计. 研究基于SVM的VaR计算方法是一种很诱人的选择. 但是实际金融时间序列通常具有较强的时效, 直接采用标准的SVM方法来进行概率密度估计效果通常都不是十分理想[5]. 本文将研究基于加权支持向量机(W-

* 收稿日期: 2009-09-13
作者简介: 胡 莹(1986—)女, 江西南昌人, 硕士研究生
E-mail: rebecca1511@sina.com

SVM)的一种新的 VaR 建模方法,并以上证综指为对象进行实证研究.

2 VaR 的定义

VaR 指在一定的概率水平(置信度)下,某一金融资产或证券组合在未来特定的一段时间内最大可能损失.用公式可以表示为

$$\mathrm{Prob}(P_\tau - P_0 > \mathrm{VaR}) = 1-\alpha \tag{1}$$

其中,P_0 为投资组合在初始时刻的价格,P_τ 为投资组合在 τ 时刻的价格.设组合回报的概率密度函数为 $f(x)$,组合回报的累计概率密度函数为 $F(x)$,$x_1, x_2, \cdots, x_n$ 为组合回报的 n 个样本观测值.则由式(1)可得:

$$F(\mathrm{VaR}) = \int_{-}^{\mathrm{VaR}} f(x)\mathrm{d}x = \alpha \tag{2}$$

从式(2)可知 VaR 可以看作置信水平 α 和 $f(x)$ 的函数.如果能够根据已有数据估计出 $f(x)$ 的经验分布,那么通过求解式(2)就可以得到在给定置信水平下的 VaR 值.由于通常 α 预先给定,故求解 VaR 的核心在于确定资产组合回报率的概率密度函数.不同的 $f(x)$ 求解方法就构成了不同的 VaR 预测模型.本文将采用加权支持向量机来求解 $f(x)$.

3 基于 W-SVM 的概率密度估计

3.1 概率密度估计问题

所谓概率密度估计就是求解线性算子方程(3)的解.

$$\int_{-} \theta(x-t)p(t)\mathrm{d}t = F(x). \tag{3}$$

其中,$\theta(x)=\begin{cases}1, & x \geqslant 0,\\ 0, & x < 0,\end{cases}$ 同时必须满足条件:$p(x)\geqslant 0$,$\int_{-} p(x)\mathrm{d}x = 1$.在该线性算子方程中,分布函数 $F(x)$ 未知,但通常给出了一些独立同分布样本 $x_1, x_2, \cdots, x_n$.利用这些样本可以构造经验分布函数:

$$F_l(x) = \frac{1}{l}\sum_{i=1}^{l}\theta(x-x_i) \tag{4}$$

3.2 概率密度估计的 W-SVM 方法

采用 SVM 方法进行概率密度估计,是从概率密度的定义出发,求解算子方程的 $Af(t) = F(x)$ 解.其中,算子 A 实现了一个从希尔伯特空间 E_1 到希尔伯特空间 E_2 的一对一映射.在实际金融时间序列中,离现在近的数据其重要性要大于早期数据,也就这些样本数据点要求相对较小的训练误差.因此在优化问题描述时,对每个样本数据点采用不同的惩罚系数 C 或采用不同的训练误差 ε,以得到更准确的回归估计,这就说所谓的加权支持向量机.

假设回归估计函数为

$$f(x) = (\omega \cdot x) + b \tag{5}$$

在约束条件 $\| Af - F \|_{E_2} \leqslant \sigma$ 下最小化泛函 $W_p(f) = \Omega(f)$, 其中 σ 为预先定义的常数. 考虑数据对 $((x_1, F(x_1)), \cdots, (x_n, F(x_n)))$, 寻找 $f(x) = \sum_{i=1}^{n} \beta_i K(x, x_i)$, 其中 $K(x, x_i)$ 为对称正定的 Mercer 核. 加权支持向量机通过对参数 C 加权, 其优化问题为最小化函数:

$$\Phi(\omega \cdot \omega) + C\left(\sum_{i=1}^{l} s_i(\xi_i^* + \xi_i)\right). \tag{6}$$

该问题可以转化为下列 SVM 的解:

$$\min \Phi(W, \xi_i, \xi_i^*) = \frac{1}{2}(W \cdot W) + C\left(\sum_{i=1}^{l} \xi_i^* + \sum_{i=1}^{l} \xi_i\right). \tag{7}$$

约束条件为

$$Y_i - (W \cdot Z_i) \leqslant \sigma_i + \xi_i^* \text{ 或 } (W \cdot Z_i) - Y_i \leqslant \sigma_l + \xi_i. \tag{8}$$

其中,

$$Y_i = F_l(x), Z_j(x) = \int_{-\infty}^{x_j} K(x, x')\mathrm{d}x',$$

$$Z_i(Z_i(x_1), \cdots, Z_i(x_i)), W = (\beta_1, \cdots, \beta_l). \tag{9}$$

此时考虑使用加权支持向量机

$$\min \Phi(W, \xi_i, \xi_i^*) = \frac{1}{2}(W \cdot W) + C\left(\sum_{i=1}^{l} \xi_i^* + \sum_{i=1}^{l} \xi_i\right). \tag{10}$$

其约束条件为

$$\begin{cases} Y_i - (W \cdot Z_i) \leqslant \sigma_i + \xi_i^*, & i = 1, 2, \cdots, l, \\ (W \cdot Z_i) - Y_i \leqslant \sigma_l + \xi_i, & i = 1, 2, \cdots, l, \\ \xi_i^* \geqslant 0, \ \xi_i \geqslant 0, & i = 1, 2, \cdots, l. \end{cases}$$

通过求解式(10)的拉格朗日函数鞍点, 可将其转化为相应的对偶问题, 式(10)的拉格朗日函数为

$$\begin{aligned} L(W, \xi_i, \xi_i^*; \alpha^*, \alpha, \upsilon, \upsilon^*) = {} & \frac{1}{2}(W \cdot W) + C\left(\sum_{i=1}^{l} \xi_i^* + \sum_{i=1}^{l} \xi_i\right) \\ & - \sum_{i=1}^{l} \alpha_i [Y_i - (W \cdot Z_i) + \sigma_l + \xi_i] \\ & - \sum_{i=1}^{l} \alpha^* [(W \cdot Z_i) - Y_i + \sigma_l + \xi_i^*] - \sum_{i=1}^{l} (\upsilon_i \xi_i + \upsilon_i^* \xi_i^*). \end{aligned} \tag{11}$$

将式(11)对 ω, ξ_i, ξ_i^* 求偏导数后代入式(10)可得其对偶问题为

$$\min \frac{1}{2} \sum_{i=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(Z_i \cdot Z_j) + \sigma_l \sum_{i=1}^{l} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} Y_i (\alpha_i^* - \alpha_i). \tag{12}$$

约束条件为

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0, \ 0 \leqslant \alpha_i^*, \alpha_i \leqslant C_i \quad i = 1, 2, \cdots, l$$

最终可得

$$\omega = \sum_{i=1}^{l}\left(a_i^* - a_i\right)x_i,\ b = y_j - \sum_{i=1}^{l}\left(a_i^* - a_i\right)\left(x_i \cdot x_j\right) + \varepsilon \tag{13}$$

将式(13) 代入式(5) 即可求得概率密度函数.

3.3 加权模型

加权系数 C_i 的选择方法对预测效果起着至关重要的作用. 对时间序列预测问题来说, 通常最近出现的数据要比先前的数据对预测结果影响要大, 需要取较高的权值. 通过对历史数据进行加权处理, 除了可以较好地计算金融时间序列的记忆效应, 还可以较好地解决 VaR 建模中样本区间长度选择困难的问题. 对于金融时间序列, 本文选择的加权模型为[6]:

$$C_i = M\vartheta_i;\ \vartheta_i = \frac{1-\lambda}{1-\lambda^l}\lambda^i,\ i = 0, 1, \cdots, l-1;\ \sum_{i=0}^{l-1}\vartheta_i = 1. \tag{14}$$

其中, M 给预先给定的一个大数, 例如 100. λ 为金融市场的最优衰退因子. 衰退因子代表了给予过去预测值与最近一期实际值的分配权数. 其值越小, 代表最近的观察值越能包含更多信息, 衰退因子的衰减速度也越快, 市场的记忆长度会越短. 因此, 从某种意义上来说, 衰退因子过小是一种市场效率低下的表现. 最优衰退因子可以通过专门的实证分析给出. 不同国家的市场衰退因子一般在0.94 ~ 0.98之间. 美国 Riskmetric 集团对于单日的风险测量取统一的衰退因子 0.94, 而对于月度的风险测量则取 0.97.

3.4 投资组合的 VaR 值计算

选择高斯概率密度函数 $N\left(x; u; B\right) = \frac{1}{\sqrt{2\pi u}}\exp\left[-\frac{\left(y-u\right)^2}{2B}\right]$ 作为 SVM 的核函数, 其中 u 为正态分布均值, B 为正态分布的方差. $\Phi\left(x; u; B\right)$ 为正态回报的累计概率密度分布函数. 设通过 W-SVM 求得的概率密度分布函数为

$$f(x) = \sum_{i=1}^{k}\alpha_i N\left(x; x_i; B\right). \tag{15}$$

设每种组合的权重为 $\boldsymbol{W} = \left(\omega_1, \cdots, \omega_n\right)$, $\sum_{i=1}^{n}\omega_i = 1$, 收益率为 $\boldsymbol{R} = \left(r_1, \cdots, r_2\right)$, 则组合收益率可表示为: $r = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{W} = \sum_{i=1}^{n} r_i\omega_i$, 其概率分布密度可以表示为

$$\begin{aligned} f_p\left(r\right) &= \int_{\boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}\leqslant r}\sum_{i=1}^{k}\alpha_i N\left(x; x_i; B\right)\mathrm{d}x = \sum_{i=1}^{k}\alpha_i\int_{\boldsymbol{R}^{\mathrm{T}}\boldsymbol{W}\leqslant r} N\left(x; x_i; B\right)\mathrm{d}x \\ &= \sum_{i=1}^{k}\alpha_i N\left(r; \boldsymbol{W}^{\mathrm{T}}x_i; \boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{W}\right)\mathrm{d}x. \end{aligned} \tag{16}$$

以收益率表示相对 VaR, 根据置信度水平 α 下 VaR 的定义 $P(r - VaR^*) = \alpha$ 有

$$1-\alpha = \sum_{i=1}^{n}\alpha_i\varphi\left(\mathrm{VaR}^*; \boldsymbol{W}^{\mathrm{T}}x_i; \boldsymbol{W}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{W}\right), \tag{17}$$

则未知变量 VaR^* 可以通过各种标准优化算法得到. 以上算法构成了完整的 W-SVM 框架下的 VaR 计算方法.

4　实证分析

4 1　数据选取与预处理

选取样本数据为上海证券交易所 1996 年 1 月 2 日到 2009 年 5 月 1 日综合指数的每日的收盘价. 用对数收益率 $r_t=\ln P(t+1)-\ln P(t)$对原数据进行处理, 从而得到可供 VaR 分析的时间序列(见图 1). 从图 1 中可以看到最近 3 年(2006 ~2009 年)波动相对比较剧烈, 股市风险变大, 而这一阶段正对应着(2005 年 5 月 9 日股改启动、2007 年 10 月 17 日最高点、人民币升值和股权分置改革、金融危机)股权分置改革和金融危机. 将样本数据分为 2001 年 1 月 ~ 2006 年 1 月和 2006 年 1 月到 2009 年 5 月两个阶段, 分别计算其收益分布的统计特征. 从统计特征可以出上证综指收益分布具有"尖峰厚尾"特性(见图 2 ~ 图 3), 第一个阶段稍微左偏, 而第二个阶段稍微右偏. 前一个阶段样本数量多, 样本波动率也对较小; 而后一个阶段样本数量少, 波动率大. 通过对两个统计特性差异较大的样本进行 VaR 的预测, 从而对 W-SVM 的性能进行评价.

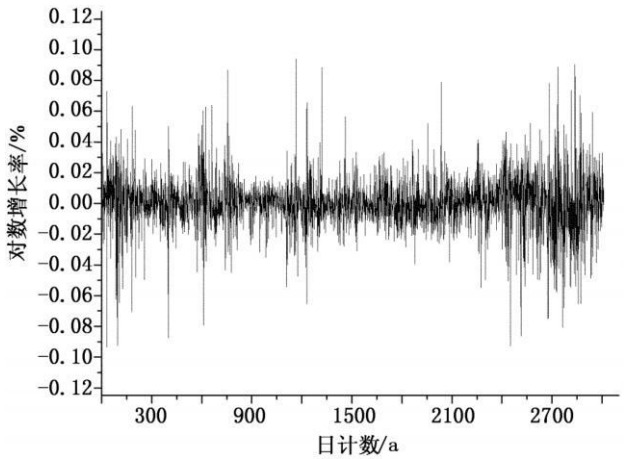

图 1　2001— 2009 年上综指对数增长率

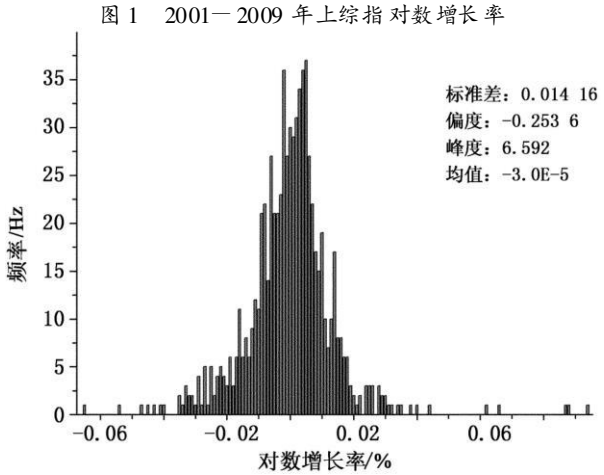

图 2　2001 ~ 2005 年频数分布

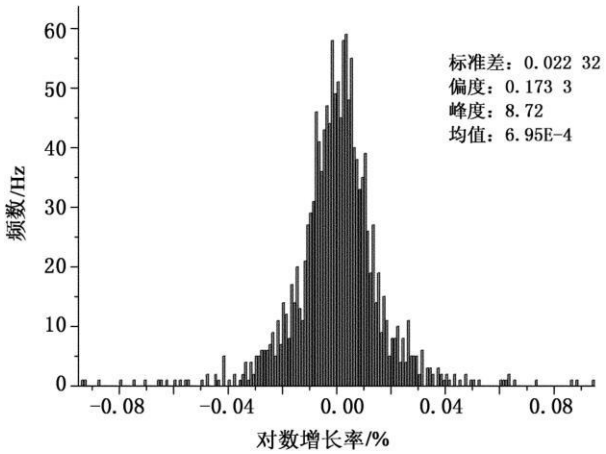

图 3　2006～2009 年频数分布

4.2　W-SVM 模型建立

W-SVM 模型采用高斯径向密度核函数，核函数参数 σ 取 10，误差控制参数 $\varepsilon=0.01$，加权参数模型 $C=M\gamma$ 中，M 取 100.2001～2006 年间样本取 $\gamma=0.967$；2006～2009 年间样本取 $\gamma=0.973$. 采用改进的序贯二次规划方法来对 W-SVM 进行训练. 建立好上证指数的对数增长率概率分布的 W-SVM 模型后，就可以用其通过式(17)来进行 VaR 的预测.

4.3　基于 W-SVM 的 VaR 模型评价

4.3.1　模型评价方法

本文采用 Kuipic 检验法对 VaR 模型进行评价. 记 N 为评价样本中投资组合的损失值大于 VaR 的次数. 基于似然比检验的思想，Kuipic 给出了某一模型是否有效的接收或拒绝的区间(见表 1)[7]. 设评价样本的个数为 T，则 N/T 则称为失败率. 将失败率与相应的左尾概率 c 进行比较，若他们无显著差异，则表示该模型所估计的值是有效的. 若他们相差很大，则说明所采用的模型是不适当的，应予拒绝. 同时，当 N 位于拒绝域左侧时，表明 VaR 模型高估了风险；位于拒绝域右侧时，表明 VaR 模型低估了风险.

表 1　$T=250$ 的 Kuipic 检验非拒绝域

左尾概率 c	5%	2.5%	1%	0.05%	0.01%
非拒绝域	[7, 19]	[3, 11]	[1, 6]	[0, 4]	[0, 1]

4.3.2　测试算例 1

上证综指训练样本取值范围为 2001 年 1 月 2 日至 2004 年 12 月 22 日，共 1 188 个数据；评价样本取值范围为 2004 年 12 月 23 日至 2005 年 12 月 30 日，共 250 个数据. 在此取 VaR 的计算窗口为 1 188 天，计算评价样本第 1 天的 VaR 值，然后将 VaR 的计算窗口和待考察的交易日依次后移 1 日可得下一日的 VaR 值共 242 个数据. 采用正态分布模型、历史模拟模型、蒙特卡罗模型和 W-SVM 模型分别进行计算，其结果见表 2.

正态分布方法在 95%时例外数在非拒绝域的左边，明显高估了风险，模型失败. 基于历史模拟法计算的值在左尾概率大于或等于 2.5%时仍高估了风险，但其在高置信度时表现较好.

对于上证综指来说是一种比较保守的模型, 不适合风险偏好型的投资者. 蒙特卡罗法在较低置信度下效果尚可, 但在较高的置信度下低估了收益率的实际损失值. 特别是当置信度为99. 9% 时, 这种低估倾向十分明显, 相较而言比较适合风险偏好型投资者.

表 **2**　各种 **VaR** 模型性能比较

置信度(1－c)	95%	97. 5%	99%	99. 5%	99. 9%	
正态分布	VaR 均值	－0. 034 82	－0. 038 72	－0. 049 57	－0. 056 35	－0. 065 12
	例外天数	5	5	2	1	1
	失败率	2. 0%	2. 0%	0. 8%	0. 4%	0. 4%
历史分析	VaR 均值	－0. 035 56	－0. 054 87	－0. 076 94	－0. 095 34	－0. 104 1
	例外天数	15	2	1	1	0
	失败率	6. 0%	0. 8%	0. 4%	0. 4%	0. 0%
蒙特卡罗	VaR 均值	－0. 036 98	－0. 049 64	－0. 054 61	－0. 059 79	－0. 072 55
	例外天数	10	5	5	4	2
	失败率	3. 2%	2. 4%	2. 0%	1. 6%	0. 8%
W-SVM	VaR 均值	－0. 031 3	－0. 047 1	－0. 055 6	－0. 064 31	－0. 077 1
	例外天数	12	6	2	1	1
	失败率	4. 8%	2. 4%	0. 8%	0. 4%	0. 4%

在较高的置信度下, SVM 模型估计时可能会低估市场风险, 而用 W-SVM 估计时也可能会轻微的高估股市的真实风险. 但检验样本中实际损失超过 VaR 值的比例与相应的 c 值比较一致, 模型拟合效果较好. 从模型可接受程度及失败率来看, W-SVM 模型在各种置信度情况下都能取得比较好的结果, 同传统方法相比具有较好的改进效果.

4. 3. 3　测试算例 2

2006 年 1 月 4 日到 2008 年 4 月 1 日为样本训练, 样本数据为 543 个. 2008 年 4 月 2 日到 2009 年 4 月 13 日为评价期, 评价样本数据为 250 个. 各种 VaR 模型的性能测试如表 3 所示. 由于这一阶段股市风险和波动幅度明显加大, 且监管层加强政策调控及国际金融环境变坏, 导致异常波动和极端事件偶有发生. 同算例 1 的样本相比, 市场变化的不确定性加剧. 加上这一阶段样本数据相对较少, 正态分布和历史分析都明显低估了市场风险, 且在低置信度时模型还通不过 Kuipic 检验. 而模特卡罗由于可以模拟出多种波动, 在低置信度时相对较好, 但在高置信度依然较大低估了市场风险, 特别是在 99. 9% 置信度时通不过 Kuipic 检验.

表 **3**　各种 **VaR** 模型性能比较

置信度(1－c)	95%	97. 5%	99%	99. 5%	99. 9%	
正态分布法	VaR 均值	－0. 035 9	－0. 037 78	－0. 051 42	－0. 060 13	－0. 068 91
	例外天数	4	4	3	3	2
	失败率	1. 6%	1. 6%	1. 2%	1. 2%	0. 8%
历史分析法	VaR 均值	－0. 036 8	－0. 061 13	－0. 079 83	－0. 096 32	－0. 102 3
	例外天数	15	8	3	2	0
	失败率	6. 0%	3. 2%	1. 2%	0. 8%	0. 0%
蒙特卡罗法	VaR 均值	－0. 037 86	－0. 052 45	－0. 057 27	－0. 063 27	－0. 072 33
	例外天数	10	7	5	4	3
	失败率	4. 0%	2. 8%	2. 0%	1. 6%	1. 2%
W-SVM 法	VaR 均值	－0. 03 43	－0. 05 21	－0. 059 1	－0. 065 54	－0. 081 4
	例外天数	13	6	2	2	1
	失败率	5. 2%	2. 4%	0. 8%	0. 8%	0. 4%

同算例 1 相比, W-SVM 模型仅略微低估了市场风险, 但基本上在各种置信度下都有比较好的表现, 且均通过了 Kuipic 检验, 具备较好的鲁棒性. W-SVM 模型在较少样本依然能够保持较好的性能, 这正是 W-SVM 相对传统统计方法的固有优势之一. 同时由于 W-SVM 模型对时间序列采用了加权技术, 减小了样本内涵信息的记忆长度, 更加符合股市运行的实际情况, 有效弥补了历史模拟法的滞后和需要大容量样本的不足, 改善了在小样本下正态分布和模特卡罗模型假设容易失误的问题.

5 结 论

本文研究了一种基于加权支持向量机的 VaR 计算新方法. 采用最优市场因子作为加权模型, 能够合理地将股市的记忆效应考虑到 VaR 模型中. 加权支持向量机能够在较少样本下比较好地预测股市风险. 对上证综指的实证研究表明, 尽管存在后尾、非线性及异常大幅度波动, W-SVM 模型在各种置信度下依然能够较好的预测 VaR, 适合于作为各种风险偏好投资者采用的 VaR 模型.

参考文献

[1] 乔瑞 F. 风险价值 VaR[M]. 陈跃译. 北京: 中信出版社, 2005.
[2] 龚锐, 陈仲常, 杨栋锐. GARCH 族模型计算中国股市在险价值风险的比较研究与评述[J]. 数量经济技术经济研究, 2005(7): 67—81.
[3] 肖志勇, 宿永铮. VaR 模型在金融风险管理中的应用[J]. 生产力研究, 2008, (24): 44—46.
[4] Vapnik V N. 统计学习理论的本质[M]. 张学工译. 北京: 清华大学出版社, 2000.
[5] 张诏, 张素, 章琛曦, 等. 基于支持向量机的概率密度估计方法[J]. 系统仿真学报, 2005(17): 2355—2357.
[6] MORGAN P. RiskMetrics-techinical document[R]. 4th ed. New-York, http://www.riskmetrics.com, 1996.
[7] KUPIEC P. Techniques for verifying the accuracy of risk measurement models[J]. Journal of Derivatives, 1995, (3): 73—84.

A New Var Model Based on Weighted Support Vector Machine

HU Ying, WANG An-min
(*School of Economics and Management, Xidian University, Xi'an, Shanxi* 710049, *China*)

Abstract According to the defects of the traditional VaR computation methods in the statistics framework, a new VaR model based on weighted support vector machine (W-SVM) was investigated. In order to deal with the memory effect in the financial time serials, a weighted model based on optimal market factor was introduced. The Shanghai composite index from the year 2001 to 2009 was modeled, and the simulation results indicated that the new VAR method based W-SVM was better than traditional methods. Even for small sample, abnormal fluctuations and heavy tails in nonlinear market, W-SVM model can obtain good performance at different confidence intervals.

Keywords value at risk; support vector machine; probability density estimation; Shanghai Composite index