Kernel Density Estimation for An Anomaly Based Intrusion Detection System

Xiaoping Shen* and Sonali Agrawal

April 9, 2006

ABSTRACT

This paper presents a new nonparametric method to simulate probability density functions of some random variables raised in characterizing an anomaly based intrusion detection system (ABIDS). A group of kernel density estimators is constructed and the criterions for bandwidth selection are discussed. In addition, statistical parameters of these distributions are computed, which can be used directly in ABIDS modeling. Statistical experiments and numerical simulations are used to test these estimators.

Index Terms **— Kernel Density Estimation, Intrusion Detection System (IDS), anomaly based IDS (ABIDS), Self-Organizing Maps (SOM).**

I. INTRODUCTION

Probability density functions (pdf) are very important in many engineering applications. Methods used in density estimations are classified as parametric (techniques estimate pdfs with known underlying distributions) and non-parametric (Techniques estimate pdfs directly from the data without known underlying distributions). Smoothing splines and kernel estimators are the most important ones in non-parametric density estimation. In this study, we aim to develop non parametric models for anomaly based intrusion detection systems (IDS).

An intrusion detection system monitors network traffic and monitors for suspicious activity and alerts the system or network administrators. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network. An Anomaly-Based Intrusion Detection System is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either Normal or Anomalous.

The classification is based on heuristics or rules, rather than patterns or signatures, and will detect any type of misuse that falls out with normal system operation. In order to determine attack traffic, the system must be taught to recognize normal system activity. This can be accomplished in several ways, most often with artificial intelligence type techniques such as Self Organized Map (SOM). A crucial assumption has to be made for probability density functions of six random variables which used to describe ABIDS. In practice, these variables are assumed to follow the Gaussian distribution. Then the abnormality will be judged based on properties of the Gaussian distribution. That is, data are drawn from For example, about 68.26% of the values are at within 1 standard deviation away from the mean, about 95.46% of the values are within two standard deviations and about 99.73% lie within 3 standard deviations. This is known as the "68-95-99.7 rule". Unfortunately, statistical experiments show that the assumption for the underline probability density distribution (See Figure 3. and Figure 4 below). That is, detection and update algorithms based on the Gaussian parametrical model can not describe the behavior of the system accurately.

Kernel density estimation is a simple technique for estimating a density function without imposition of parametric model. In this work, we construct kernel density estimators for six random variables raised in characterizing an anomaly based intrusion detection system (ABIDS). A group of kernel density estimators is constructed

X. Shen, is with the Department of Mathematics, Ohio University, Athens, OH, 45701, USA. phone: 740-593-1288; fax: 740-593-9805; E-mail: shen@math.ohiou.edu).

S. Agrawal, is with the School of Electrical Engineering and Computer Science, and the Department of Mathematics, Ohio University, Athens, OH 45701 USA. E-mail: sa237903@ohio.edu. Her contribution to this paper is part of her Math 692 project.

The authors are thankful to Manikantan Ramadas for his help in collecting the data used in the experiments of the paper.

and the criterions for bandwidth selection are discussed. In addition, statistical parameters of these distributions are computed, which can be used directly in SOM algorithm. Statistical experiments and numerical simulations are used to test these estimators.

II. Background

A. Kernel Density Estimation

From what follows, we will assume a random variable x is independent and identical distributed (i.i.d.) and denote its density function by $p(x)$. That is,

$$P(a < x < b) = \int_a^b p(x)dx\,.$$

The density estimation is to estimate the density function by given an observed data set $\{X_i\}_{i=1}^N$. One approach to density estimation is parametric, which assumes the data are drawn from a known parametric family of distributions with statistical parameters. The density function $p(x)$ is then can be estimated by finding estimates of these parameters. Another approach is the nonparametric, which does not assume that the data drawn from a parametric family although it assumes that the distribution possesses a density p.

The conventional kernel used in density estimation. The kernel density estimator $\hat{p}(x)$ for the estimation of the density value $p(x)$ at point x is defined as

$$\hat{p}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right), \qquad (1)$$

where K is called the kernel function, and h is called window width or bandwidth. The kernel K satisfies the following conditions:

C1. $K(u) = K(-u)$;

C2. $\int_{-\infty}^{\infty} K(u)du = 1$;

C3. $\int_{-\infty}^{\infty} u^2 K(u)du = k_2 \neq 0$.

For density estimation, we would like the estimators themselves to be density functions, and hence non-negative. Condition C1 can be removed. Some non symmetric kernels have been used successfully in density estimation (see [8] for example). C2 indicates that positive kernels are probability functions. Most popular kernels are Epanechnikov kernel, Triangular kernel, biweight kernel and Gaussian kernel (See Figure 1 below).

The Epanechnikov kernel was first suggested in density estimation. The analytical formulas for these kernels are,

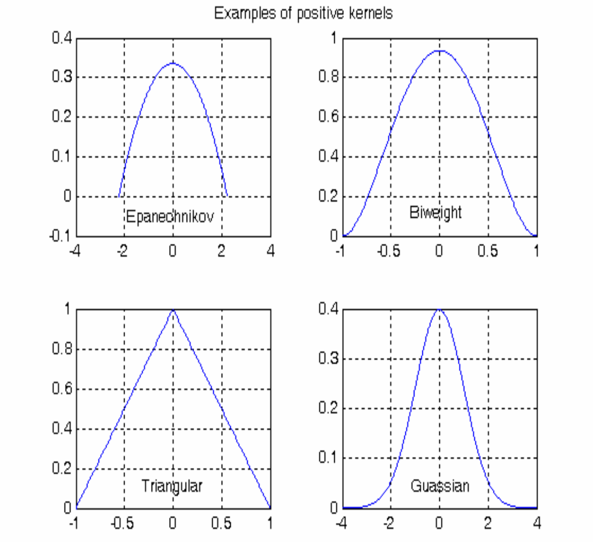

Figure 1. Some popular positive kernels used in density estimation.

1) The Epanechnikov kernel:

$$K_E(x) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}x^2), & |x| < \sqrt{5} \\ 0, & \text{else} \end{cases};$$

2) The Triangular kernel:

$$K_T(x) = \begin{cases} 1 - |x|, & |x| < 1 \\ 0, & \text{else} \end{cases};$$

3) The biweight kernel:

$$K_B(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & |x| < 1 \\ 0, & \text{else} \end{cases};$$

4) The Gaussian kernel:

$$K_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,.$$

Notice that the first 3 kernels are polynomial type with compact support, while the Gaussian kernel has the highest regularity but infinite support. Table 1 shows the second moments k_2 defined in C3 where $R(K)=\int_{-\infty}^{\infty}K^2(x)dx$.

Table 1. Second moments and R(K)

Kernel	k_2	**R(K)**
K_E	1	$3\sqrt{5}/25$
K_T	1/6	$2/3$
K_B	1/7	5/7
K_G	$1/4\pi^{3/2}$	$1/2\sqrt{\pi}$

More discussions about the kernel density estimation method can be found in [2], [3], [4], [11], [12], [13], [14] and [17].

In recent years, some multiscale density estimators were introduced. For instance, the wavelet density estimators (as orthogonal series estimation, see [5], [6], [7], [15], and [16]) have multiscale structure and better convergence rate.

The choice of the kernel is not crucial for density estimation in the case of i.i.d random variables. However, the bias and variance of the density estimators will depend on the kernel. In addition, the smoothness of the kernel will be required in error analysis. A more difficult task in kernel estimations is to select the bandwidth (See [1], [9] and references therein).

B. The Anomaly Based Intrusion Detection Systems

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network- what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other and alert the administrator or user when traffic is detected which is anomalous, or significantly different, than the baseline. Such a system characterizes each network connection by using random variables. For instance, the Integrated Network Based Ohio University's Network Detective Service (INBOUNDS) is an anomaly based intrusion detection system, which is described by six random variables. Namely, interactivity (INTER), average size of question (ASOQ), average size of answer (ASOA), log of question answer idle time (LQAIT), log of answer question idle time (LAQIT) and duration of connection (DOC). For simplicity, we will denote these six variables by $V_i,\ i=1,\ldots,6$. Random samples of size 100 drawn from data set of size 2305 for each of the six random variables are displayed in Figure 1. Histograms of a random sample of size 512 are shown in Figure 2. The sample statistics can be found in Table 2.

The histograms of these random variables suggest these samples do not come from normal distributed populations. To gain more insight, we performed the quantile-quantile (QQ) tests to compare these samples with normal distributed population. QQ plots of the sample data versus standard normal distribution are shown in Figure 3. The results suggest similar properties reflected by the histograms of these data. Before jumping to the final conclusion, we exam skewness and kurtosis of these samples by using the Jarque-Bera test for goodness-of-fit to a normal distribution. The tests were set up as in the following:

H_0^i : The input data V_i has normal distribution

H_A^i : The input data V_i has other distribution

The rejection to the hypothesis is significant at the 5% level. The Results from Jarque-Bera tests reject the hypothesis that V_i has normal distribution at significant level of 5%. Based on the statistical analysis, we conjecture that these random variables do not have normal distribution.

More detailed description of INBOUNDS system can be found in [10] and [12]. The system collects the raw data in the form of network packets from the *Data Source* sub-module. The tool used in this module is *tcpurify*. The tools reads the packets on wire, encrypt the sender and receiver IP address, removes the payload from the packet, and reports only the first 64 bytes of the packet. It normally keeps the data up to the transport layer i.e. includes the TCP, IP, Ethernet, and ICMP headers for further analysis. The encryption details are maintained in a file, which can be used by *Active Response Module* for adding a rule to the firewall or even blocking the traffic. Intrusion Detection module is the heart of detection engine. Real time data from multiple sources is fed into the ANDSOM module. The intrusion detection module decides whether traffic is anomalous.

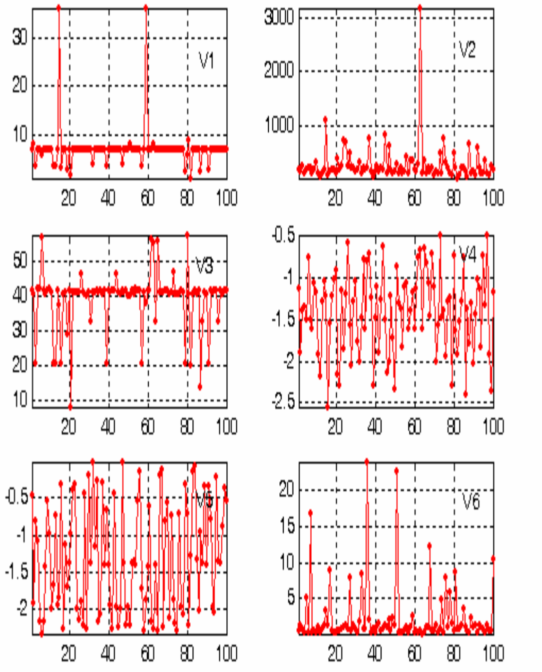

Figure 2. Random samples of size 100 drawn from data set of size 2306.

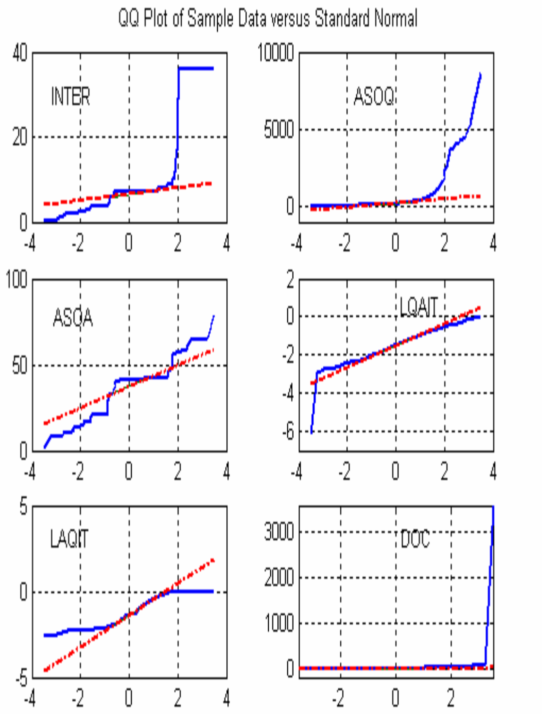

Figure 3. Quantile-Quantile tests for samples of six random variables. The straight lines are joining the first and third quartiles of each distribution.

III. Kernel Density Estimators for Random Variables in Anomaly Based Intrusion Detection Systems

A. The Density Estimators

We assume random variables V$_i$, i=1…,6, has continuous distributions with density functions $p_i(x)$. Let X_i be a random sample with size N drawn from the population. The kernel density estimator of V$_i$, denoted by $\hat{p}_h(x)$, will take the from:

$$\hat{p}_h(x) = \frac{1}{Nh}\sum_{j=1}^{N} K(\frac{x-X_j}{h}),$$

where, h is the window width (or the bandwidth) and K is one of the kernels in II.A, We construct kernel density estimators for each of the random variables by using 4 different kernels defined in II.A. For each random variable, a couple of random samples of size 500 from the original data are used to generate the density estimator. Selected graphical results are displayed in Figure 5 and Figure 6.

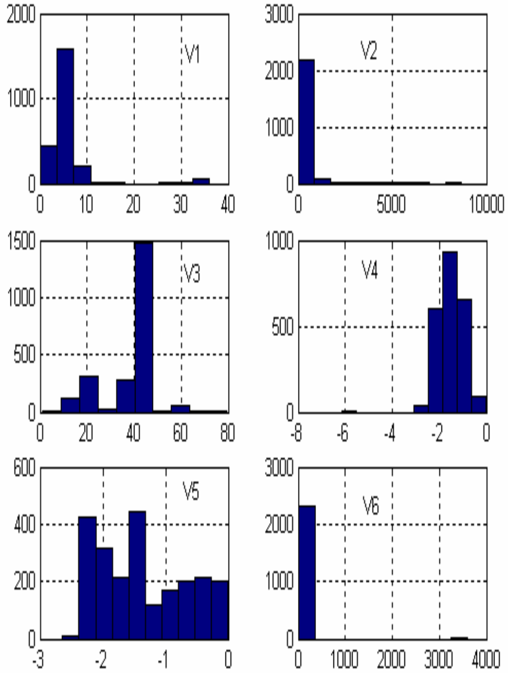

Figure 4. Histogram plots of random sample 2, sample size =512.

B. Bandwidth Selections

The problem of bandwidth selections is a crucial but challenge problem. Figure 5 and Figure 6 show density estimators derived by using different bandwidth selections. We observe that when the bandwidth is too small, the density estimator is very rough, that is, they are "undersmoothed". However, when the bandwidth is too large, the density estimator is too flat, that is, they are "oversmoothed". There are many interesting discussions about bandwidth selections in literature (see for example, [1], [9] and [13]). However, there is no perfect answer in general.

If we denote by h_{opt} the optimal bandwidth subject to the minimization of the approximate mean integrated square error, then we have a rough estimation formula:

$$h_{opt} \approx 1.06\hat{\sigma}(N)^{-1/5},$$

where $\hat{\sigma}$ is sample standard deviation, and N is the sample size. Table 2 show the approximated optimal bandwidth for each of the density estimators where $\hat{\sigma}$ is the average of the std of 2 samples of size 500 (shown in Table 3). Figure 7 shows the "finalists" of density estimators.

Table 2. Estimated optimal bandwidth

	h_{opt}		h_{opt}
V1	1.5125	V4	0.1576
V2	178.084	V5	0.2138
V3	2.9632	V6	1.6004

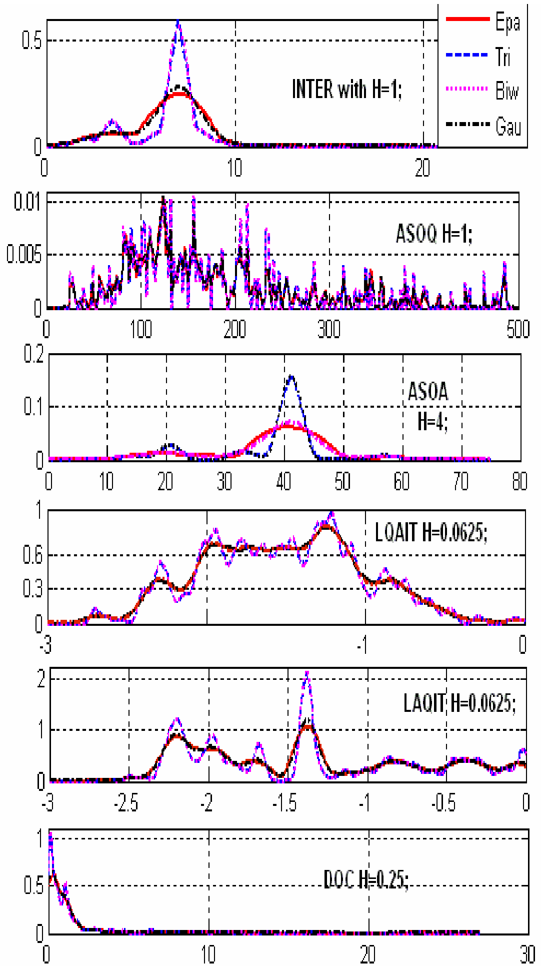

Figure 5. Kernel density estimators for variables V1-V6 using 4 difference kernels.

C. Asymptotic Formula for Density Estimators

For computational purpose, we use cubic spline regressions of the density functions (called cubic spline density curves) other than the density functions themselves. Figure 8 shows cubic spline density for random variables $V_1, \ldots, V_6$. The corresponding statistical parameters for cubic spline density curves can be found in Table 3.

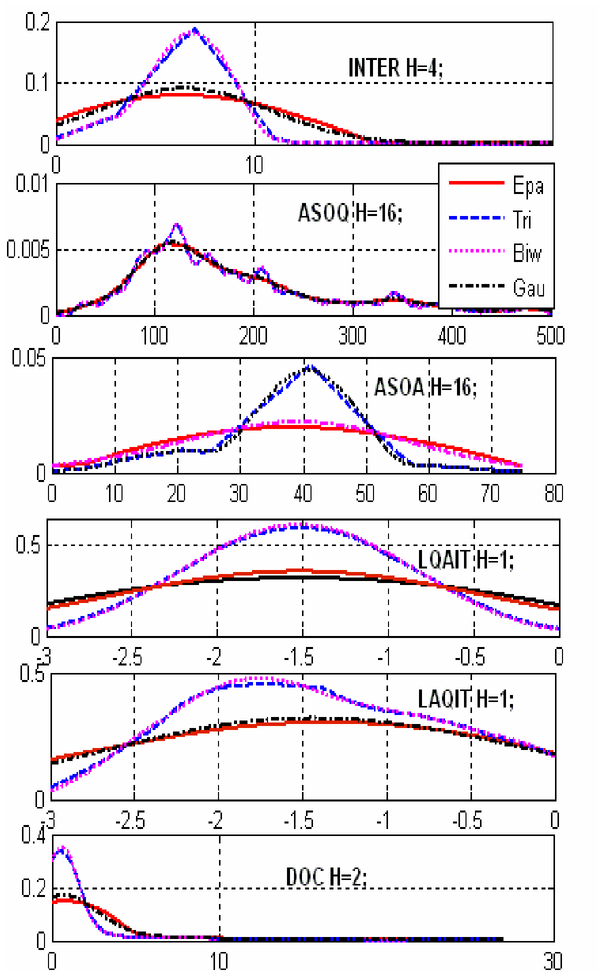

Figure 6. Kernel density estimators for variables V1-V6 using 4 difference kernels.

IV. Properties of Statistical Parameters of Dendisty Estimators

In this section, we discuss properties of the density estimators. In general, the expectation of $\widehat{p}_h(x)$ will equal the expectation of $\widehat{p}(x)$ up to order $O(h^2)$, and the variance will be the order $O(h)$. The asymptotic integrated squared bias and mean integrated squared error (AMISE) are given by (see [13], page 131):

$$ISB = \frac{1}{4} k_2^2 h^4 R(p''(x)) + O(h^6)$$

and

$$AMISE = \frac{R(K)}{nh} + \frac{1}{4} k_2^2 h^4 R(p''(x)) .$$

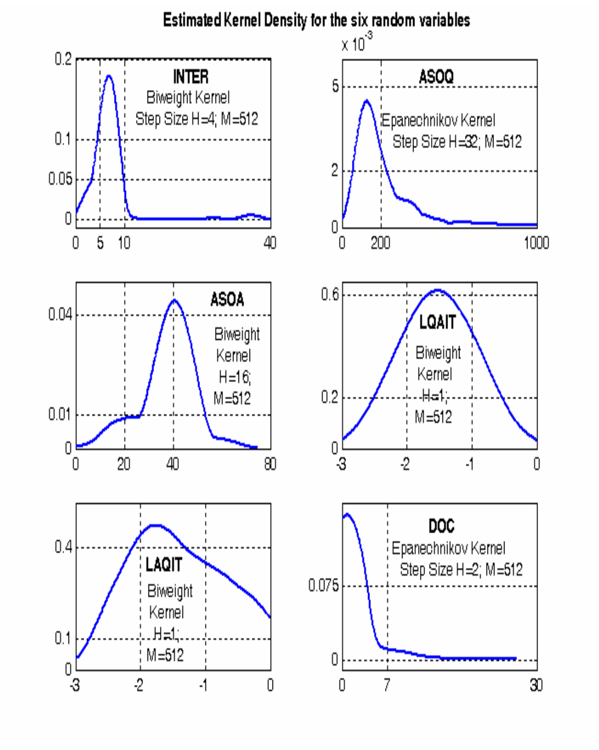

Figure 7. Finalists for kernel density estimators for variables V1-V6.

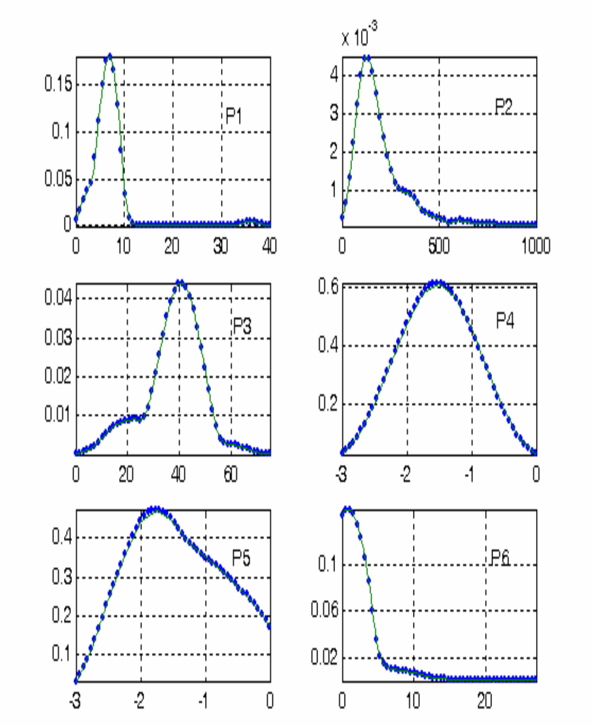

Figure 8. Spline approximation of density functions P1 to P6 of random variables V1 to V6, respectively.

V. Conclusion and discussion

In this article, we introduced density estimators for random variables used to describe an abnormally based IDS, the method is illustrated by data from INBOUNDS system. A group of density estimators are derived and their corresponding statistical parameters are computed. These density estimators can be used to improve the SOM algorithm to get more accurate detection for intrusion. The discussion of bandwidth selection, error analysis and the use of multiscale density estimators will be discussed in future work.

Table 3. Mean and Standard Deviation For Random Variables and Their Samples

	Std. σ	Spline Std $\overline{\sigma}$	Sample Std $\sqrt{S^2}$		Mean μ	Spline Mean $\overline{\mu}$	Sample Mean a	
			Sample 1	Sample2			Sample 1	Sample2
V_1	4.8933	4.8520	4.6868	5.2039	7.1280	7.2056	7.0473	6.8278
V_2	580.77	571.93	688.772	475.74	203.2575	201.4453	327.87	316.22
V_3	11.3071	11.252	9.6597	9.7170	38.0396	37.8865	37.914	36.811
V_4	0.5982	0.5365	0.4980	0.5324	-1.4955	-1.2941	-1.517	-1.540
V_5	0.6998	0.6093	0.7125	0.6854	-1.3565	-1.1661	-1.358	-1.374
V_6	2.8076	2.7915	5.5772	4.8878	1.9927	2.0803	2.1141	2.004

Reference

[1] A. Ahmad and I. S. Ran. Data based bandwidth selection in kernel density estimation with parametric start via kernel contrasts. J. Nonparametr. Stat., Vol. **16** (6), 841--877, 2004.

[2] W. Bowman. *Applied smoothing techniques for Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Oxford University Press, New York, 1997.

[3] W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations*. Oxford University Press, USA, 1997.

[4] L. Devroye, *A course in density estimation*. Birkhauser, Boston, 1987.

[5] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard. Density estimation by wavelet thresholding. Ann. Statist. Vol. **24** (2) 508--539, 1996.

[6] A. Elgammal, R. Duraiswami. Background and foreground modeling using nonparameteric kernel density estimation for visual surveillance. Proceedings of the IEEE, Vol. **90** (7), 2002.

[7] R. A. Kronmal and M. E. Tarter. The Estimation of Probability Densities and Cumulatives by Fourier Series Methods. J. Amer. Statist. Assoc., Vol. **63**, 925-952, 1968.

[8] R. Mugdadi and A. M. Lahrech. The exponential kernel in density estimation. Far East
Far East J. Theor. Stat., Vol. **14** (1), 1--14, 2004.

[9] R. Mugdadi, I. A. Ahmad. A bandwidth selection for kernel density estimation of functions of random variables. Comput. Statist. Data Anal, Vol. **47** (1), 49-62, 2004.

[10] S. Ostermann. Tcpttrace–TCP connection analysis tool. URL: http://www.tcptrace.org.

[11] E. A. Nadaraya. *Nonparametric estimation of probability densities and regression curves*. Kluwer Academic Publishers Group, 1989.

[12] A. Sawant. *Time-based approach to intrusion detection using multiple self-organizing maps*. Master's thesis, Ohio University, 2005.

[13] W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, Boca Raton, FL 1986.

[14] W. Scott. *Multivariate density estimation, theory, practice, and visualization*. John Wiley & Sons, Inc., New York, 1992.

[15] G. G. Walter and X. Shen. *Wavelets and other orthogonal systems. 2^{nd} edition*, Studies in Advanced Mathematics, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[16] G. G. Walter and X. Shen. Continuous non-negative wavelets and their use in density estimation, Comm. Statist. Theory Methods, Vol. **28**, 1-18, 1999.

[17] M. P. Wand, *Kernel smoothing,* Chapman & Hall, New York, 1995.