Mathematical Problems of Computer Science 44, 145--153, 2015.

Application of Multivariate Statistical Analysis in Process Control

Tigran Z. Khachikyan and Sahak M. Narimanyan

Yerevan State University
Department of Probability Theory and Mathematical Statistics
e-mail: tkhach@inbox.ru, sahakn@yandex.ru

Abstract

Due to significant increase of information systems and its intensive usage in our everyday life, several problems like automatic identification of system faults, finding times of drastic change in stochastic characteristics as well as locating those characteristics, which “went out of control” need to be addressed. To solve these problems, we propose an algorithm based on multivariate statistical analysis. The algorithm is implemented with the R software environment and tested on custom metrics for Vesta server and other groups of random metrics.

Keywords: Principal component analysis, level of significance, T^2 statistics, Q statistics, Loading matrix, Score matrix, Eigenvalues, Covariance matrix, Upper critical value.

1. Introduction

Due to significant increase of information systems and its intensive usage in our everyday life, several problems like automatic identification of system faults, finding times of drastic change in stochastic characteristics as well as locating those characteristics, which “went out of control” need to be addressed. The normal process is usually conditioned by some characteristics, which may correlate with each other. In that case, analysis of individual characteristics may lead to significant errors due to different confidence intervals as well as impossibility of identification of joint level of significance. We used principal component analysis (PCA), Hotelling’s criteria based on T^2 statistics, which is known to be uniformly the most powerful test (the null hypothesis for a vector of average values) in the class of all randomized tests invariant to transformations of similarity, to solve this problem. We also used Q-statistics for the residual matrix of dataset after PCA prediction.

2. Method

Let $X = (X_1, X_2, ..., X_m)$ be a multivariate random variable, individual components of which characterize the state of a system having joint normal probability density function

$$f(x) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-(x-\mu)^T \Sigma^{-1}(x-\mu)\right\},$$

$x = (x_1, x_2, \dots, x_m)$, where μ a vector of average values, Σ is a covariance matrix. To test the hypothesis $H_0 : \mu = \mu_0$ in multivariate case, let us use the Hotelling's T^2 statistics ([1])

$$T^2 = n\left(\bar{X} - \mu_0\right)^T S^{-1}\left(\bar{X} - \mu_0\right),$$

where n, $n > m$, is a sample size, S is a sample estimate of covariance matrix Σ, $\bar{X}$ is a mean vector of X. We will find values of T^2 statistics for each instant time t ($t = 1, 2, \dots$).

For this, we take historical data of fixed sample size n before time t, and shifting time t forward, while keeping the sample size unchanged. We will denote these values of T^2 statistics as T_t^2. If the covariance matrix Σ is known, then the T^2 statistics has a distribution χ^2 and $T_{cr}^2 = q\chi_m{}^2(1-\alpha)$, where $q\chi_m{}^2(1-\alpha)$ is the quantile of $\chi_m{}^2$ distribution with significance level α. And when the covariance matrix Σ is unknown, the upper critical value of T^2 statistics is calculated as

$$T_{cr}^2 = \frac{m(n-1)}{n-m} qf(1-\alpha, m, n-m),$$

where qf is the quantile of Fisher's distribution with parameters $(m, n-m)$, α is a significance level.

The T^2 statistics has a probability density function

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right) x^{\frac{m}{2}-1}\left(1+\frac{x}{n}\right)^{-\frac{n+1}{2}}}{\Gamma\left(\frac{n-m+1}{2}\right)\Gamma\left(\frac{m}{2}\right) n^{\frac{m}{2}}}, \quad x > 0.$$

We will consider that the studied system operates normally, when $T_t^2 < T_{cr}^2$.

The method of PC ([2]) is one of the ways to reduce the size of datasets, while losing minimum amount of information. For our system and in case of multivariate random variable, we need to construct such an orthogonal coordinate system to transform correlated variables into new uncorrelated variables. Sample divergence in relation to principal components is organized in decreasing order. We take so many principal components that the summarized sample divergence of PC is comprising 95% of the total divergence. Using this method, we get a loading matrix and a score matrix. The values of these new variables form the factor scores, and these scores can be interpreted geometrically as the projections of the observations onto the principal components. The loadings are simply the coordinates of original variables in the principal components space. The loading matrix P has dimensions $(m \times k)$, where m is a space dimension and k is the quantity of principal components. The scoring matrix T has dimensions $(n \times k)$, where n is the sample size, k - the number of principal components. The residual matrix is $R = X - TP^t$, where X is the dataset matrix, $X = (X_1, X_2, \dots, X_m)$. Using the method of principal components we decrease dataset space dimension and hence, lose some information. In order to estimate the influence of other parameters on our system, let us consider Q -statistics for the residual matrix introduced by Jackson ([2]).

The Q –statistics is the following $Q = r^t r$, where r is the vector-column of the residual matrix R. The upper critical value of Q statistics is

$$Q_{cr} = \theta_1 \left[\frac{C_\alpha h_0 \sqrt{2\theta_2}}{\theta_1} + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{\frac{1}{h_0}},$$

where

$$\theta_i = \sum_{j=k+1}^{m} \lambda_j^i, \quad i = 1, 2, 3, \; h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}.$$

Here C_α is a quantile of standard normal distribution with significance level $1-\alpha$ and λ_j are eigenvalues of sample covariance matrix S.

3. Implementation

1) Let $t_0, t_1, t_2, \ldots$ be the arrival times of dataset, and $t_i - t_{i-1} = const$. Let Δ be the length of the interval for historical data. We will investigate the data matrix in the time interval $\left[t_k - \Delta, t_k\right]$, $k = 0, 1, \ldots$
2) At time t_0, we remove those variables from the data matrix, which are constant (do not change over that time interval). The resulting matrix represents a multivariate data sample at time t_0. We can normalize this matrix then.
3) We can employ PC method and as a result can find those components, which provide 95% of total divergence, score matrix, loading matrix, and residual matrix.
4) We can calculate the T^2 statistics at time t_0 and T_{cr}^2 at time t_0. Then we calculate both Q statistics and Q_{cr} at time t_0.If $T^2 < T_{cr}^2$ and $Q < Q_{cr}$, then our system functions normally. Otherwise, if $T^2 \geq T_{cr}^2$ or $Q \geq Q_{cr}$ the null-hypothesis is declined and it is assumed that the system “went out of control”, i.e., malfunctioning occurred. An automatic alert messaging to system administrators can be organized to take measures.
5) Calculation of weights for individual parameters in T^2 and Q statistics takes place.

Those parameters, which have significant weight, can be considered as cause for both the T^2 and Q statistics to go out of control. For the next time t_1, we go back to point 1) and start over and again.

4. Computerized and Real-life Example

In the real-life example below, the system monitors 16 different parameters of VESTA system working on real multivariate data. Implementation of the suggested algorithm and using PC method, the following results are obtained.

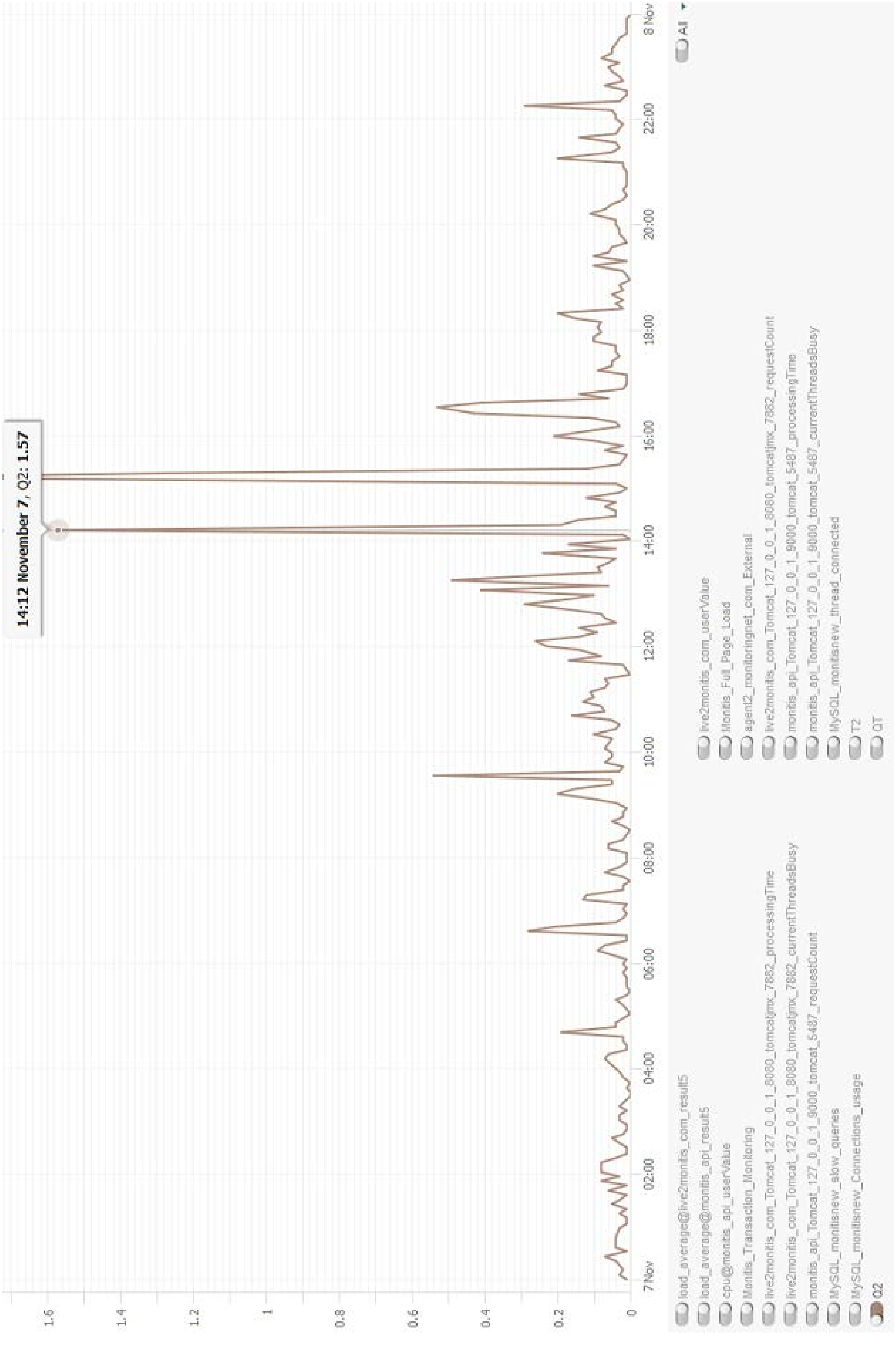

Fig. 1. Q statistics transformed to unity Q/Q_{cr} . At time 14:12 the value of Q statistics is greater than 1, so the system “went out of control ”.

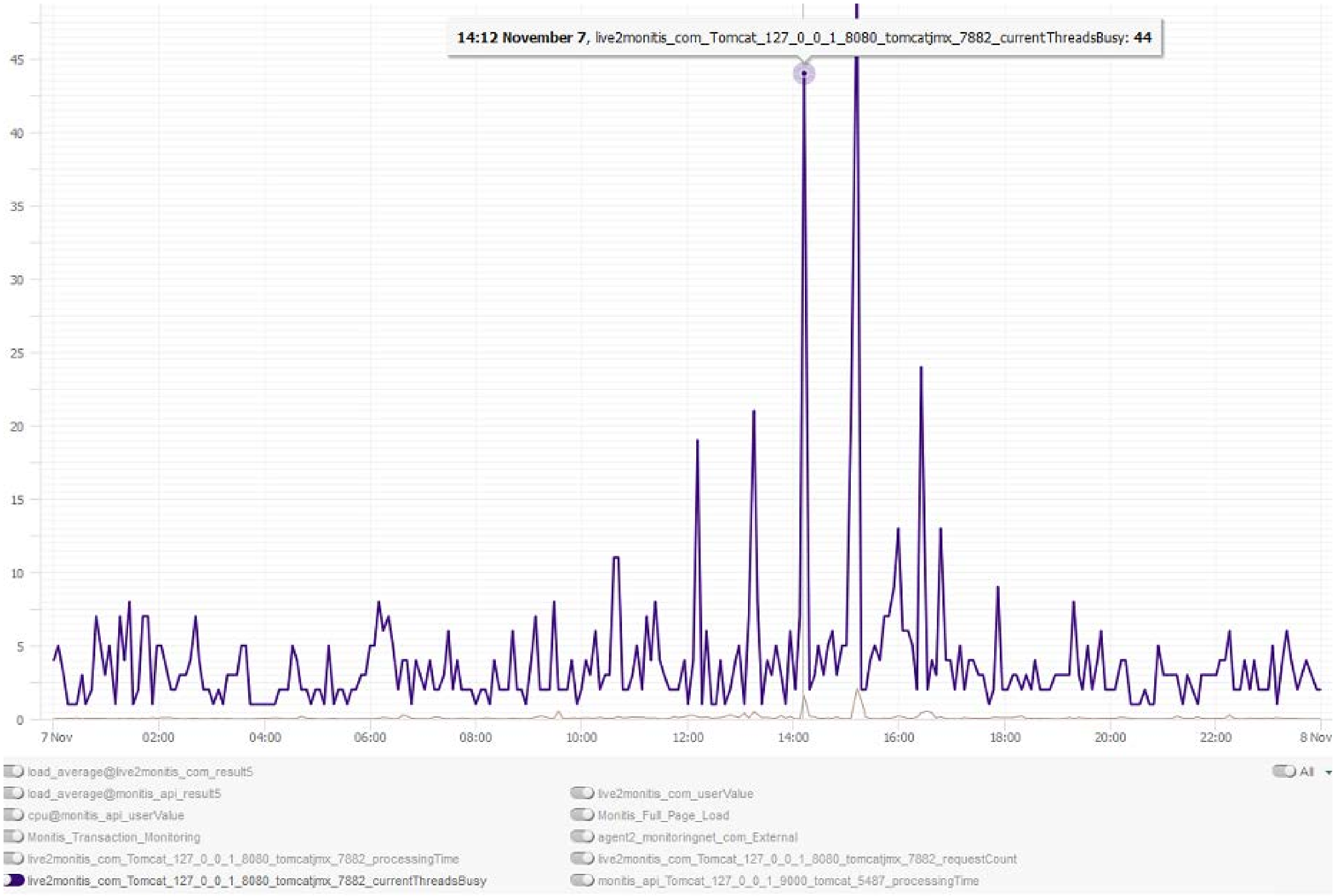

Fig. 2. At time 14:12 the graph depicts the metrics, weight of which is the highest, resulting the value of Q statistics is greater than 1.

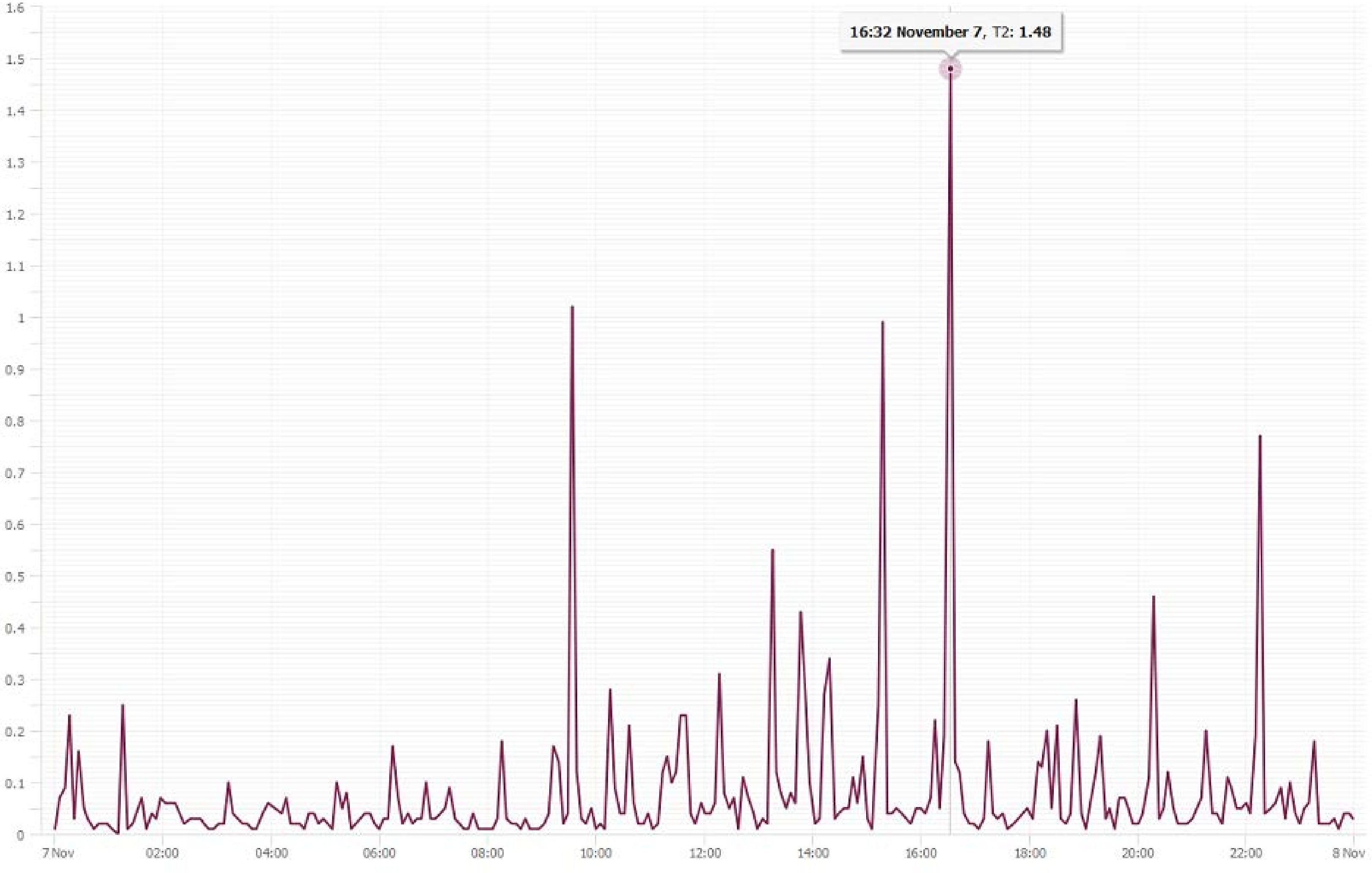

Fig. 3. The graph of T^2 transformed to unity T^2/T_{cr}^2. At time 16:32 the value of T^2statistics is greater than 1, so the system “went out of control”.

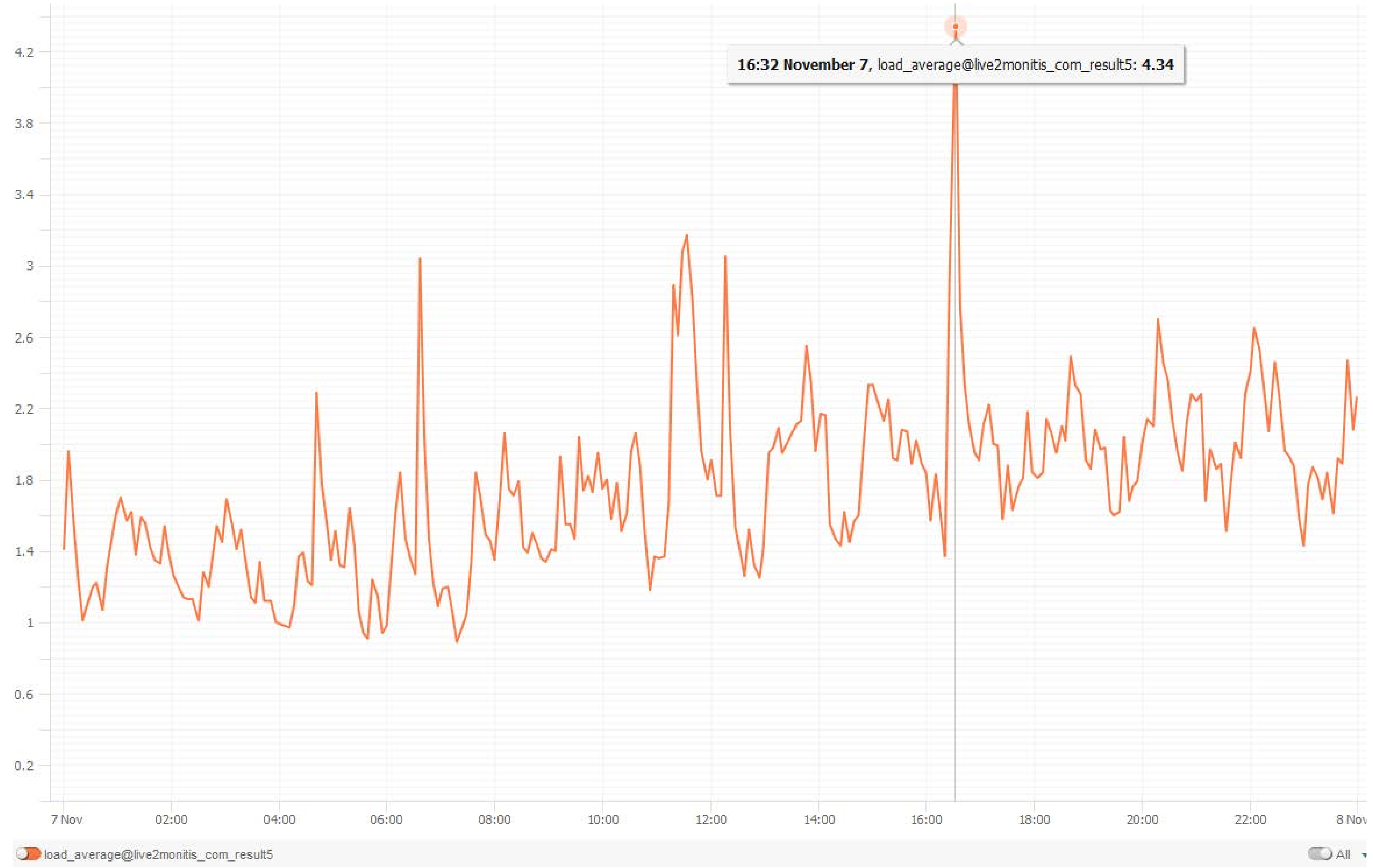

Fig. 4. At time 16:32 the graph depicts the metrics, weight of which is the highest, resulting the value of T^2 statistics to be greater than 1.

5. Conclusion

Principal component analysis (PCA) is the most popular multivariate statistical technique and it is used by almost all scientific disciplines. PCA method can be successfully applied to provide solutions for many IT-related problems like automatic identification of system faults, finding times of drastic change in stochastic characteristics as well as locating those characteristics, which “went out of control”. A particular problem for system administrators is to monitor performance of custom metrics due to both absence of relevant thresholds and quantity of such metrics, which in some situations can be significant. The suggested algorithm used principal component analysis (PCA), Hotelling’s criteria based on T^2 statistics, which is known to be uniformly the most powerful test (the null hypothesis for a vector of average values) in the class of all randomized tests invariant to transformations of similarity, to solve this problem. Q -statistics is also used for the residual matrix of dataset after PCA prediction. The algorithm is applied to real-life situation with VESTA system monitoring 16 different parameters on real multivariate data. The algorithm enables system administrators to identify event times when the system “went out of control” as well as to locate the “problematic” parameters causing such problems. Automatic alert messaging and control mechanism can be organized to warn system administrators to take measures.

References

[1] T. W. Anderson, *An Inroduction to Multivariate Statistical Analysis*, 3rd ed., Wiley series in Probability and Mathematical Statistics, 2003.

[2] J. E. Jackson, *A User’s Guide to Principal Components*, Wiley series in Probability and Mathematical Statistics, 1991.

Submitted 23.07.2015, accepted 27.11.2015

Բազմաչափ վիճակագրական վերլուծության կիրառություն գործընթացների կառավարման ոլորտում

Տ. Խաչիկյան և Ս. Նարիմանյան

Ամփոփում

Մեր առօրյա կյանքում ինֆորմացիոն համակարգերի զգալի աճի և նրանց ինտենսիվ օգտագործման պատճառով, կարիք է առաջանում բազմաթիվ խնդիրների հետ առնչվել, ինչպիսիք են՝ համակարգի անսարքությունների ավտոմատ հայտնաբերումը, ստոխաստիկ բնութագրիչների կտրուկ փոփոխությունների պահերի որոշումը, ինչպես նաև՝ այն բնութագրիչների վերհանումը, որոնք «դուրս են եկել կառավարումից»: Այդպիսի խնդիրների լուծման նպատակով առաջարկում ենք ալգորիթմ՝ հիմնված բազմաչափ վիճակագրական վերլուծության վրա։ Ալգորիթմը իրականացված է R ծրագրավորման լեզվով և փորձարկված է Վեստա սերվերի պատահական բնութագրիչների համար, ինչպես նաև այլ պատահական մետրիկների խմբերի համար։

Применение многомерного статистического анализа для контроля процессов

Т. Хачикян и С. Нариманян

Аннотация

В связи со значительным увеличением информационных систем и их интенсивным использованим в нашей повседневной жизни, возникают проблемы, такие как автоматическая идентификация неисправностей системы, нахождение времен резкого изменения стохастических характеристик, а также определение тех характеристик, которые "вышли из-под контроля". Для решения таких задач, мы предлагаем алгоритм основанный на многомерном статистическом анализе. Алгоритм реализован в среде программирования R и тестирован на случайных метриках сервера Веста и в других группах случайных метрик.