应用概率统计 第 XX 卷
第 X 期 XXXX 年 XX 月

Chinese Journal of Applied Probability
and Statistics, Vol.XX No.X, Mon Year

文章编号: 2013/011

A Distribution-Free Change-Point Model for Monitoring Dispersion*

ZHOU MAOYUAN

(*Sience College, Civil Aviation University of China, Tianjin 300300*)

GENG WEI, LI ZHONGHUA†

(*Institute of Statistics and LPMC, Nankai University, Tianjin 300071*)

Abstract

Most distribution-free control charts in the literature are used to monitor process location parameters, such as mean or median, rather than process dispersion parameters. This paper develops a new distribution-free control chart by integrating a two-sample nonparametric test into the effective change-point model. Our proposed chart is easy in computation, convenient to use, and very powerful in detecting process dispersion shifts. As it avoids the need for a lengthy data-gathering step before charting and it does not require knowledge of the underlying distribution, the proposed chart is particularly useful in start-up or short-run situations.

Keywords: Change-Point, Dispersion, Distribution-Free, Statistical Process Control.

AMS Subject Classification: 62N10.

§1. Introduction

Statistical process control (SPC) has been widely used in various industrial processes. Most SPC applications assume that the quality of a process can be

* The project supported by the NNSF of China Grants (11401573,11201246,11101198), the CAUC Science College Fundamental Research Funds for the Central Universities (3122014K007), and the CAUC Research Funds (2013QD25X).

†**作者信息**: 周茂袁(出生时间-1980), 男, 讲师, 主要研究方向: 统计质量控制, E-mail:yitangzhang@163.com ; 耿薇(出生时间-1967), 女, 副教授, 主要研究方向: 统计质量控制, E-mail:gengwei@nankai.edu.cn ; 李忠华(通讯作者)(出生时间-1982): 男, 讲师, 主要研究方向: 统计质量控制, E-mail:zli@nankai.edu.cn .

本文XXXX 年 XX 月 XX 日收到, XXXX 年 XX 月 XX 日收到修改稿.

adequately represented by the distribution of a quality characteristic and the in-control (IC) and out-of-control (OC) distributions are the same with only differing parameters.

While parametric methods are only useful in certain applications, there is often a lack of enough knowledge about the process distribution. For example, univariate process data are often assumed to have normal distributions, although it is well recognized that, in many applications, particularly in start-up situations, the underlying process distribution is unknown and not normal, so that statistical properties of commonly used charts, designed to perform best under the normal distribution, could potentially be (highly) affected. So distribution-free charts are needed in such situations. A chart is called distribution-free if its IC run-length distribution is nearly the same for every continuous distribution (Chakraborti et al. (2001)).

In the last several years, distribution-free control charts have attracted much attention. For example, Bakir and Reynolds (1979) proposed a cumulative sum (CUSUM) chart for group observations based on the Wilcoxon signed-rank statistic. McDonald (1990) considered a CUSUM procedure for individual observations based on the statistics called "sequential ranks". An exponentially weighted moving average (EWMA) chart for individual observations proposed by Hackl and Ledolter (1991) is constructed by the "standardized ranks" of observations, which is determined by IC distributions. If the distribution is not available, they recommended using the ranks in collected reference data instead. The distribution-free charts considered by Chakraborti et al. (2004, 2009) are based on the precedence test. Recently, a Shewhart-type chart and a scheme using change-point formulation based on the Mann-Whitney test statistic were investigated by Chakraborti and Van deWiel (2008), Zhou et al. (2009) and Hawkins and Deng (2010). Jones et al. (2009) developed a rank-based distribution-free Phase I control scheme for subgroup location. Other developments include Albers and Kallenberg (2004) and Bakir (2004, 2006). Zhou et al. (2008) proposed a robust control chart based on wavelets for preliminary analysis of individual observations. Wu et al. (2009) proposed a synthetic control chart based on Cornish-Fisher expansion. A nice overview on the topic of univariate distribution-free control charts was presented by Chakraborti et al. (2001). In addition, distribution-free control charts in multivariate cases have been discussed

by Liu (1995), Qiu and Hawkins (2001), and Qiu (2008).

Most of distribution-free charts mentioned above focus on monitoring process median, but monitoring the process dispersion is also highly desirable. However there are far fewer distribution-free control charts which can monitor process dispersion. Zou and Tsung(2010) proposed a chart which incorporates a powerful goodness-of-fit (GOF) test (Zhang (2002)) using the nonparametric likelihood ratio into a EWMA chart. It can detect more general changes than location shifts, and is also very easy in computation, but leaves a tuning parameter λ to choose.

This paper develops a new distribution-free control chart by integrating a two-sample nonparametric test (Mood 1954) into the effective change-point model. Simulation studies show that the proposed method is superior to other distribution-free schemes in monitoring dispersion. As it avoids the need for a lengthy data-gathering step before charting (although it is generally necessary and advisable to have about at least 19 warm-up samples) and it does not require knowledge of the underlying distribution, so the proposed chart is particularly useful in start-up or short-run situations.

The rest of this paper is organized as follows. The test for fixed-size sample is given in Section 2. Sequentially monitoring for Phase II is derived in Section 3. The performance comparisons with two other distribution-free control charts are discussed in Section 4. And the conclusion is given in Section 5.

§2. Test for fixed-size sample

We begin by considering the Phase I problem of detecting a change-point in a fixed-size sequence of observations. We denote the observations by $\{X_1, \cdots, X_t\}$, and the goal is to test whether they have all been generated by the same probability distribution. We assume that no prior knowledge is available regarding this distribution, other than that it is continuous. Using the language of statistical hypothesis testing, the null hypothesis is that there is no change-point and all the observations come from the same distribution, while the alternative hypothesis is that there exists a single change-point τ in the sequence which partitions them into two sets, with $X_1, \cdots, X_\tau$ coming from the pre-change distribution F_0, and $X_{\tau+1}, \cdots, X_t$ coming

from a different post-change distribution F_1:

$$H_0 : X_i \sim F_0, \text{ for } i = 1, \cdots, t$$

$$H_1 : X_1, \cdots, X_\tau \sim F_0, \ X_{\tau+1}, \cdots, X_t \sim F_1.$$

We can test for a change-point immediately following any observation X_k by partitioning the observations into two samples $S_1 = \{X_1, \cdots, X_k\}$ and $S_2 = \{X_{k+1}, \cdots, X_t\}$ of sizes $n_1 = k$ and $n_2 = t - k$ respectively, and then performing an appropriate two sample hypothesis test. For example, to detect a change in location parameter without making assumptions about the distribution, Mann-Whitney statistic would be an proper test statistic (Hawkins and Deng (2010)). In order to monitor the process dispersion, we will consider the Mood test.

The Mood test uses a statistic like this:

$$M'_{k,t} = \sum_{j=1}^{n_1} (R_{1j} - \frac{n_1 + n_2 + 1}{2})^2$$

where R_{1j} is the rank of the jth observation X_j in the pooled sample.

R_{1j} could be computed as: $R_{1j} = \sum_{i=1}^{n_1+n_2} I(X_i \le X_j)$, where $I(X_i < x)$ is the indicator function:

$$I(X_i < x) = \begin{cases} 1, & \text{if } X_i < x, \\ 0, & \text{otherwise.} \end{cases}$$

The mean and variance of the Mood test statistic are

$$E_{H_0}(M'_{k,t}) = n_1((n_1 + n_2)^2 - 1)/12$$

and

$$Var_{H_0}(M'_{k,t}) = n_1 n_2 (n_1 + n_2 + 1)((n_1 + n_2)^2 - 4)/180,$$

respectively.

In fact, we use the absolute value of the standardized Mood test statistic

$$M_{k,t} = |(M'_{k,t} - E_{H_0}(M'_{k,t}))/\sqrt{Var_{H_0}(M'_{k,t})}|.$$

We reject the null hypothesis that no change occurs at k if $M_{k,t} > h_{k,t}$ for some appropriately chosen value of $h_{k,t}$.

The statistic can be integrated into the change-point model, and is easy to compute. Now, since we do not know in advance where the change-point is located, we do not know which value of k to use for partitioning. We therefore specify a more general null hypothesis, that there is no change at any point in the sequence. The alternative hypothesis is then that there exists a change-point for some unspecified value of k. We can perform this test by computing $M_{k,t}$ at every value $0 < k < t$, and taking the maximum value. This leads to the maximized test statistic:

$$M_t = \max_k M_{k,t}, 0 < k < t.$$

If $M_t > h_t$ for some suitably chosen threshold h_t, then the null hypothesis is rejected, and we conclude that a change occurred at some point in the data. In this case, the best estimate $\hat{\tau}$ of the location of the change-point is at the value of k which maximized M_t. If $M_t \leq h_t$, then we do not reject the null hypothesis, and hence conclude that no change has occurred. The choice of this threshold will be discussed further in the following section.

§3. Sequentially monitoring

Having considered the problem of detecting changes in a fixed-size sample, we now turn to the task of sequentially Phase II monitoring where new observations are being received over time. Let X_t denote the tth observation where t is increasing over time.

Once a new observation X_t is received, we then regard $\{X_1, \cdots, X_t\}$ to be a fixed-size sample, and use the method from the above Section to test if a change-point has occurred. The problem of sequentially monitoring is then reduced to performing a sequence of fixed-size tests. Suppose it is desired to have an IC average run length (ARL_0) of γ. This can be achieved if we choose the h_t values so that the probability of incurring a false alarm at the tth observation equals to $1/\gamma$. We hence require that for all t:

$$\begin{aligned} &P(M_1 > h_1) = \alpha \\ &P(M_t > h_t | M_{t-1} \leq h_{t-1}, \cdots, M_1 \leq h_1) = \alpha, t > 1. \end{aligned} \tag{3.1}$$

It is not trivial to find a sequence of h_t values which satisfy this property. The approach in Hawkins and Deng (2010), is to use Monte-Carlo simulation. We will

do in the same way. One million realizations of the sequence $\{X_1, \cdots, X_{1000}\}$ were generated. Because the distribution of M_t is independent of the distribution of the X_i observations, these X_i values can be sampled from any continuous distribution so long as they are independent and identically distributed. Then for each value of t, M_t is computed for each of the million realizations. The values for h_t corresponding to the desired ARL_0 can then be read off from them.

Note that since there are only a finite number of ways to assign ranks to a set of t points, the M_t statistic can only take a discrete set of values. This creates a problem for threshold choice when t is small, since it may not be possible to find a value for h_t which gives the exact ARL_0 required, which is a general problem when dealing with discrete valued test statistics. Therefore we recommend that Phase II monitoring only begins after the first 19 observations have been received, which gives sufficient possibilities for rank assignments to make most ARL_0s achievable. This seems a reasonable compromise, since in practice it would be very difficult to detect a change that occurred during the first 19 observations. Now we denote our chart by CPMD, implying change-point model for monitoring process dispersion.

§4. Performance comparisons

We now evaluate the performance of our chart. As is standard in the quality control literature, we measure performance as the average time taken to detect a change of magnitude δ, which we denote by $\mathrm{ARL}_1(\delta)$. We consider changes which affect the process dispersion. Three different process distributions are considered: the standard Normal distribution $N(0, 1)$, the Student-t distribution with 3 degrees of freedom $t(3)$, and the chi-square distribution with 3 degrees of freedom χ^2_3. The latter two correspond to the heavy tailed and skewed distributions respectively.

Because our chart can be treated as a self-starting chart, the number of observations available before the change may have a large impact on its performance. We will consider changes which occur after both 50 and 100 observations, i.e. $\tau \in [50, 100]$. We compare our CPMD chart to two other change-point detection algorithms. The first is the method described in Hawkins and Deng (2010) for location shifts, which we will denote by CPML. It uses a similar change-point model

Table 1 $ARL_1(\delta)$ for dispersion shifts in the $N(0,1), t(3)$ and χ^2_3 distributions, for several values of the change time τ.

		$N(0,\delta^2)$			$t(3)/\sqrt{3}\times\delta$			$(\chi^2_3-3)/\sqrt{6}\times\delta$		
τ	δ	CPML	cpmd	EWMAZ	CPML	cpmd	EWMAZ	CPML	cpmd	EWMAZ
50	2.0	228	18.3	25.8	278	50.7	80.2	149	14.4	19.7
	3.0	141	7.9	12.6	179	12.5	25.0	78.7	6.9	11.6
	0.5	707	38.8	408.9	707	79.8	386.8	624	29.1	303.2
	0.33	769	17.1	223.7	769	22.6	312.8	595	14.6	110.0
300	2.0	73.0	10.1	10.6	104	18.6	25.3	43.3	8.3	7.6
	3.0	37.3	5.0	5.4	52.3	8.4	10.5	24.8	5.1	5.0
	0.5	1285	22.8	213.8	1144	32.1	307.9	476	21.0	63.7
	0.33	728	15.2	36.9	728	19.0	72.7	229	14.3	30.0

to ours, but there test statistic is the Mann-Whitney statistic. Second, we compare our CPMD chart to Zou and Tsung (2010), which integrates the nonparametric likelihood ratio test framework into the change-point model. We notice that their chart contains a tuning parameter λ used in the EWMA scheme. Large values of λ produce a chart which is more efficient to large changes, while small values of λ produce a chart which is sensitive to small changes. We choose to use $\lambda = 0.1$ which is a value considered in their paper, and we denote their chart by EWMAZ. To allow fair comparisons we set the ARL_0 of every chart to 500. Similar results hold for other values of ARL_0, but we omit them for space reasons.

For each of the three distributions, 10000 sequences were generated, and the change consists of multiplying δ to all post-change observations respectively. The average time taken to detect the change is then recorded for each chart.

Table 1 shows the average time required to detect shifts in dispersion, from which we can get the following conclusions.

- Our chart is much better than the CPML at all cases of dispersion shifts.
- Our chart is much better than the EWMAZ at most cases of dispersion shifts.

So we can conclude that: when we want to monitor dispersion shifts, our chart is the best choice since it gives excellent performance across all magnitudes of shifts

considered based on comparisons above.

§5. Conclusions

We proposed a new distribution-free and self-starting control chart to detect dispersion shifts by integrating a two-sample nonparametric test (Mood 1954) into the effective change-point model.

Our chart is much better than some other nonparametric methods at most cases for shifts in dispersion. As it avoids the need for a lengthy data-gathering step before charting (although it is generally necessary and advisable to have several warm-up samples) and it does not require knowledge of the underlying distribution, so the proposed chart is particularly useful in start-up or short-run situations.

References

[1] Albers, W. and Kallenberg, W. C. M. (2004). "Empirical Nonparametric Control Charts: Estimation Effects and Corrections". *Journal of Applied Statistics* 31, pp. 345-360.

[2] Bakir, S. T. (2004). "A Distribution-Free Shewhart Quality Control Chart Based on Signed-Ranks". *Quality Engineering* 16, pp. 611-623.

[3] Bakir, S. T. (2006). "Distribution-Free Quality Control Charts Based on Signed-Rank-Like Statistics". *Communications in Statistics: Theory and Methods* 35, pp. 743-757.

[4] Bakir, S. T. and Reynolds, Jr., M. R. (1979). "A Nonparametric Procedure for Process Control Based on Within-Group Ranking". *Technometrics* 21, pp. 175-183.

[5] Chakraborti, S.; Eryilmaz, S.; and Human, S. W. (2009). "A Phase II Nonparametric Control Chart Based on Precedence Statistics with Runs-Type Signaling Rules". *Computational Statistics and Data Analysis* 53, pp. 1054-1065.

[6] Chakraborti, S.; van der Laan, P.; and van de Wiel, M. A. (2004). "A Class of Distribution-Free Control Charts". *Journal of the Royal Statistical Society*, Series C 53, pp. 443-462.

[7] Chakraborti, S.; van der Laan, P.; and Bakir, S. T. (2001). "Nonparametric Control Charts: An Overview and Some Results". *Journal of the Royal Statistical Society* 33, pp. 304-315.

[8] Chakraborti, S. and van deWiel, M. A. (2008). "A Nonparametric Control Chart Based on the Mann-Whitney Statistic". *IMS Collections. Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* 1, pp. 156-172.

[9] Hackl, P. and Ledolter, J. (1991). "A Control Chart Based on Ranks". *Journal of Quality Technology* 23, pp. 117-124.

[10] Hawkins, D. M. and Deng, Q. (2010). "A Nonparametric Change Point Control Chart". *Journal of Quality Technology* 42, pp. 165-173.

[11] Jones-Farmer, L. A.; Jordan, V.; and Champ, C. W. (2009). “Distribution-Free Phase I Control Charts for Subgroup Location”. *Journal of Quality Technology* 41, pp. 304-316.

[12] Liu, R. (1995). “Control Charts for Multivariate Processes”. *Journal of the American Statistical Association* 90, pp. 1380-1387.

[13] McDonald, D. (1990). “A CUSUM Procedure Based on Sequential Ranks”. *Naval Logistic Research* 37, pp. 627-646.

[14] Mood, A. (1954). “On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests”. *Annals of Mathematical Statistics* 25, pp. 514 - 533.

[15] Qiu, P. (2008). “Distribution-Free Multivariate Process Control Based on Log-Linear Modeling”. *IIE Transactions* 40, pp. 664-677.

[16] Qiu, P. and Hawkins, D. M. (2001). A Rank-Based Multivariate CUSUM Procedure. *Technometrics* 43, pp. 120-132.

[17] Wu, c. and Wang, Z. (2009). A Synthetic Control Chart Based on Cornish-Fisher Expansion. *Chinese Journal of Applied Probability and Statistics* 25(3), pp. 258-265.

[18] Zhang, J. (2002). “Powerful Goodness-of-Fit Tests Based on Likelihood Ratio”. *Journal of the Royal Statistical Society, Series B* 64, pp. 281-294.

[19] Zhou, C.; Zou, C.; and Wang, Z. (2008). A Robust Control Chart Based on Wavelets for Preliminary Analysis of Individual Observations. *Chinese Journal of Applied Probability and Statistics* 24(3), pp. 274-288.

[20] Zhou, C.; Zou, C.; Zhang, Y.; and Wang, Z. (2009). Nonparametric Control Chart Based on Change-Point Model. *Statistical Papers* 50, pp. 13-28.

[21] Zou, C., Tsung, F., Likelihood ratio based distribution-free EWMA schemes. *Journal of Quality Technology* 2010. 42, 174-196.

一个监控方差的与分布无关的变点模型

周茂袁

(理学院, 中国民航大学, 天津市 *300300*)

耿薇，李忠华

(统计研究院, 南开大学,天津市 *300071*)

文献中绝大部分与分布无关的控制图用于监控过程位置参数，如均值或中位数，而非过程方差。该文开发了一个新的与分布无关的控制图，通过整合一个两样本非参数检验和有效的变点模型。所提出的控制图容易计算，方便应用，并且对于探测过程方差的漂移非常有效。因为它避免了在监控之前的一个很长时间的收集数据的阶段，并且它不需要潜在的过程分布的知识，因此，所提出的控制图在开始阶段或者短程运行情况下特别有用。

关键词: 变点, 方差, 与分布无关, 统计过程控制.

学科分类号: O213.1