Memorandum 1984 (June 2012). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Nonparametric control charts for bivariate high-quality processes

Willem Albers

Department of Applied Mathematics

University of Twente

P.O. Box 217, 7500 AE Enschede

The Netherlands

Abstract. For attribute data with (very) low rates of defectives, attractive control charts can be based on the maximum of subsequent groups of r failure times, for some suitable $r \geq 1$, like $r = 5$. Such charts combine good performance with often highly needed robustness, as they allow a nonparametric adaptation already for Phase I samples of ordinary size. In the present paper we address the problem of extending this approach to the situation where two characteristics have to be monitored simultaneously. Generalization to the multivariate case is straightforward.

Keywords and phrases: Statistical Process Control, health care monitoring, average run length, estimated parameters, order statistics

2000 Mathematics Subject Classification: 62P10, 62H10, 62C05

1 Introduction and motivation

Over the years a lot of effort has been devoted to increase industrial production standards with respect to quality. As a consequence, nowadays many processes exhibit only a tiny fraction of defectives. In the quite different application area of health care monitoring, this situation is in fact the norm: 'bad' events like birth defects or surgical errors should really be rare. In either field, control charts play an important role in maintaining and even further improving such high quality (see e.g. Sonesson and Bock (2003), Thor et al. (2007) and Shaha (1995) for some review papers). The fact that the (average) probability p of a defective during the in-control (IC) phase is typically very small, makes it attractive to base the control charts in question on the waiting times between failures. After accumulating say r (for some $r \geq 1$) of these, a signal follows if the total waiting time involved is considered uncharacteristically small. In this case, the process may well have gone out-of-control (OoC), in the sense that p has increased, and action should thus be taken. Of course, a small outcome can also be due to natural variation during IC. Hence the boundary value should be selected with care, e.g. by specifying which false alarm rate (FAR) is acceptable.

The simplest case is that of homogeneity, i.e. where all items or patients have the same p and the total waiting time thus is negative binomially distributed. This situation is considered in Albers (2010). One question covered there is how the best value of r might be chosen in relation to the underlying parameters, while yet another question concerns the estimation aspect, in view of the fact that p is typically unknown. Simply replacing p by an estimate from a Phase I sample sounds attractive and is common practice. However, as the FAR's involved are typically quite small as well, the impact of this estimation error is not at all negligible. Considerable attention is needed to adequately control, and preferably correct for, these errors.

e-mail: w.albers@utwente.nl;
running title: Nonparametric control charts for bivariate high-quality processes.

Actually, once the estimation problem has been recognized, it often turns out not to stop at just a single parameter. If homogeneity fails, e.g. due to strong variation between patients, clinging to the assumption of a negative binomial distribution will lead to a model error, in addition to the already present estimation error. The latter may become negligible if the size of the Phase I sample can be increased at will (which will not often be the case in practice), but the former will even then remain as it is.

The obvious way to address this serious problem is to move on to a nonparametric version of the chart. This indeed offers a nice improvement, but only after removing some obstacles. First of all, estimation errors were already considerable in the parametric case, but in the nonparametric setup they tend to become really huge. The reason is the involvement of very small quantiles (e.g. $FAR = 0.001$), for which the customary Phase I sample sizes (e.g. 100) are simply far too small. Removing the model error at the cost of an extreme estimation error clearly is not a satisfactory solution. The way around this obstacle is to go from using the sum of r individual waiting times to considering their maximum. The probability that such a maximum falls below the individual q-quantile is q^r, which implies that low values such as 0.001 can actually be attained for not really small, and thus nicely estimable, values of q (e.g. $q = 0.1$ for $r = 3$). The resulting empirical MAX-chart is both easy to understand and to apply.

However, note that here a second obstacle arises: the maximum might turn out to perform poorly compared to the sum, which latter choice actually is the optimal solution if homogeneity happens to be true after all. Fortunately, in Albers (2011a) it is shown that the loss in going from sum to maximum is in fact remarkably small, even under such favorable conditions as pure homogeneity. Hence it can be viewed as a small insurance premium to be paid to ensure robustness against departures from (often overly optimistic) model assumptions such as homogeneity. In fact, similar remarks can be made when comparing this nonparametric proposal to $CUSUM$-type methods, which may even offer some further efficiency gain. However, once more such superiority is somewhat dubious as the optimal conditions required will rarely hold to sufficient precision. Deviations from model assumptions and estimation effects will again have serious effects, which for the rather complicated $CUSUM$ procedures are difficult or even impossible to analyze. Consequently, from a robustness point of view, the simple nonparametric MAX-approach is a serious competitor.

After this brief outline, we now arrive at the topic of the present paper: how to extend the approach above to simultaneously monitoring two (or more) high-quality processes involving related characteristics. First we provide some background on multivariate attribute control charts. An example in health care monitoring was already given by Lu and Bhattacharyya (1991) concerning paired relief time data for the same headache patient. The need for studying the subject has quite recently been stressed by several authors, e.g. Xie et al. (2010) and Xie et al. (2009). One reason mentioned is that with correlated characteristics the results of using independent univariate charts can be very misleading. Another interesting aspect is mentioned by Chiu and Kuo (2010): most of the multivariate control charts (Hotelling T^2, $CUSUM$, $EWMA$) are designed for variable data and only comparatively few are meant for attribute data. These authors briefly discuss some of the methods in this latter category and conclude that most are not very satisfactory. Drawbacks mentioned are bad performance, complexity and reliance on normality. Their own subsequent proposal requires simulations in order to evaluate its performance. None of the papers mentioned are aimed especially at the high-quality case. An exception in this respect seems to be Steiner et al. (1999), who use monitoring an arterial switch operation on newborn babies as their motivating example. In this case the binary surgical outcomes recorded are 'death' and 'near miss' and it is proposed to apply simultaneous $CUSUM$ charts with secondary control limits. In view of our remarks above for the univariate case on robustness (MAX-chart) versus efficiency ($CUSUM$-chart), it definitely makes sense

extend our approach as well to the bivariate case.

In section 2 we provide the material we shall need from the univariate case. Next, in section 3, we deal with the case of two (essentially) independent processes. Even in this simple situation, considerable care is needed to combine the information in a sensible manner. In fact, several options exist here. After this, the really dependent case is considered in section 4. 'Really' here means that the probability of both events occurring at the same time is of the same order of magnitude as that of having precisely one of the events occurring. Only in this way we have a non-negligible correlation and thus a real difference compared to the situation from section 3. As explained above, our starting point will always be the basic homogeneous case, i.e. where the underlying distributions are simply geometric. The empirical nonparametric versions of the charts are subsequently discussed in section 5.

2 The univariate case

Before considering the bivariate situation, it is useful to briefly summarize the ideas and the notation for the univariate case from Albers (2011a). As mentioned, the nonparametric version will follow later; our starting point always is the simple homogeneous case. Let $D_1, D_2, \ldots$ be a sequence of independent identically distributed random variables (r.v.'s), with failure probability $P(D_1 = 1) = 1 - P(D_1 = 0) = p$ during IC. Once the process goes OoC, this p becomes θp, for some $\theta > 1$, and a signal should follow as soon as possible. This basic sequence of D_i's generates a new sequence of r.v.'s X_j, $j = 1, 2, \ldots$, defined as the waiting times from the $(j-1)^{th}$ up to and including the j^{th} failure. Hence the X_j are $G(p)$, where we let 'the r.v. W is $G(q)$' mean that W is geometric with parameter q, i.e. $P(W = k) = q(1-q)^{k-1}$, $k = 1, 2, \ldots,$

For some sensibly chosen $r \geq 1$, the MAX-chart from Albers (2011a) now signals if $\max(X_1, \ldots, X_r)$ is too small; if this does not happen, a new group of size r is considered, etc.. To ensure that charts for different r behave similarly during IC, the boundary value n should satisfy

$$FAR = P(\max(X_1, \ldots, X_r) \leq n) = r\alpha, \tag{2.1}$$

for some selected $\alpha > 0$ small. In that case, for the run length RL we obviously have that RL/r is $G(r\alpha)$, and thus the average run length (ARL) during IC will nicely equal $r/(r\alpha) = 1/\alpha$ for all r. Since $P(X_1 \leq n) = 1 - (1-p)^n$, from (2.1) we note that n should equal $\log(1 - \{r\alpha\}^{1/r})/\log(1-p)$ (either rounding or interpolation can be used; differences involved are negligible in practice). Next we move on to the OoC-phase and observe that now the (no longer false) alarm rate has become $\{1 - (1-\theta p)^n\}^r$. By plugging in the result for n, we arrive at

$$ARL = \frac{r}{\{1 - (1 - \{r\alpha\}^{1/r})^{g(\theta,p)}\}^r}, \tag{2.2}$$

where $g(\theta, p) = \log(1-\theta p)/\log(1-p) \approx \theta\{1 + (\theta - 1)p/2\}$. As we are only interested in (very) small p, it follows that to high precision

$$ARL = ARL_{r,\theta} = \frac{r}{\{1 - (1 - \{r\alpha\}^{1/r})^{\theta}\}^r}. \tag{2.3}$$

Before moving on, we consider the step from (2.2) to (2.3) from yet another angle, as this will be useful in dealing with the bivariate case. Let $Y = \Sigma_{j=1}^{n} D_j$, then Y is $\mathrm{Bin}(n, p)$ and thus for the first waiting time X_1 we can write $P(X_1 > n) = P(Y = 0) = (1-p)^n$, once again showing that X_1 is $G(p)$. Since n is large and p is small, a Poisson approximation is highly

accurate here (also cf. Albers (2010)). Let Y^* be $P(\lambda t)$, with $\lambda t = np$, then we obviously arrive at $P(X_1^* > t) = P(Y^* = 0) = \exp(-\lambda t)$ and hence at X_1^* being $\mathrm{Exp}(\lambda)$. Use of this continuous version, with exponential rather than geometric r.v.'s, leads through the bound $t = \log(1 - \{r\alpha\}^{1/r})/\lambda$ and the alarm rate $\{1 - \exp(-\theta\lambda t)\}^r$ indeed exactly to (2.3).

As concerns the behavior of $ARL_{r,\theta}$ as a function of θ, we observe that it decreases from $1/\alpha$ at $\theta = 1$ to a lower limit r as θ becomes very large. In particular, for $r = 1$ (the simple geometric chart), we have $ARL_{1,\theta} \approx 1/\{1-(1-\alpha)^\theta\} \approx 1/(\theta\alpha)$, which is indeed not impressive, unless θ is really large. Hence, as demonstrated in Albers (2010), the smaller the supposed excess of θ over 1, the larger r should be. This qualitative statement is made more explicit in the rule of thumb (derived for the sum-based negative binomial charts, but also working well for the closely related MAX-charts):

$$r^{\mathrm{opt}} = \frac{1}{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)}, \tag{2.4}$$

for values of α in (0.001,0.01) and θ in (3/2,4). As follows immediately from (2.4) (also see Table 3 in Albers (2010)), quite a variety of values for r^{opt} can occur, including rather large ones. Since the use of too large r may feel awkward in practice and moreover most of the improvement over $r = 1$ is already achieved in the first steps, a truncated version like $\min(5, r^{\mathrm{opt}})$ seems a useful compromise.

A final remark about the monitoring of a single characteristic is that so far we have worked with ARL's on the scale of the numbers of failures involved. Of course, we may also prefer to use instead the time elapsed. In the geometric case this then means multiplying ARL by a factor $1/p$, and in the exponential one by $1/\lambda$. Usually it is quite immaterial which of these two scales is chosen. If we nevertheless want to make a distinction, we should argue as follows. A requirement like $ARL = 1/\alpha$ for some small prescribed α in the first case means that $FAR = r\alpha$ should hold, while in the second one $FAR = r\alpha/\lambda$ is required. Hence, if e.g. the intensity of the process increases, the intended FAR should decrease in the latter case in order to balance this effect, while nothing has to change in the former situation. Which of the two is the proper choice is a matter of taste, i.e. depends on the intended application. Anyhow, when comparing competing proposals, the same scale obviously should be used for each of the candidates.

3 The independent case

After the brief review of the univariate case in the previous section, we are now in a position to address the bivariate situation. As announced in the introduction, we shall study the (essentially) independent case in the present section and postpone treatment of the dependent situation till the next one. Hence here we consider two types of (very) rare failures, both of which can occur, but not simultaneously. E.g. redefine the D_j by letting $P(D_1 = i) = p_i$ for $i = 1, 2$ and $P(D_1 = 0) = 1 - p_1 - p_2$ during IC, and subsequently define X_{ij}, $j = 1, 2, \ldots$ as the j^{th} waiting time for a failure of type i, $i = 1, 2$. Since both p_i are (very) small, these waiting times X_{ij} are indeed almost independent. To begin with, a straightforward computation shows that the correlation of X_{11} and X_{21} equals

$$\rho = \frac{-p_1 p_2}{(p_1 + + p_2)[(1 - p_1)(1 - p_2)]^{1/2}}, \tag{3.1}$$

which is very small as well for such p_i. This suggests that in finding n_i such that $P(X_{i_1} \leq n_i)$ has some prescribed small value, the dependence effect is negligible. In fact, the easiest way to make this negligibility completely transparent is to apply the Poisson step after (2.3) here as well. Let Y_i^* be $P(\lambda_i t)$, with $\lambda_i t = np_i$, then $P(X_{i_1}^* > t) = P(Y_i^* = 0) = \exp(-\lambda_i t)$, and thus

the $X^*_{i_1}$ are Exp(λ_i), but moreover these r.v.'s are clearly also independent. In other words, applying the univariate approximation step from (2.2) towards (2.3) in the present bivariate setting in addition also removes the minor dependence.

Consequently, for ease of presentation, we shall from now on simply use the continuous framework: we have two independent Poisson processes with intensities λ_i, and thus a joint Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$, and also corresponding exponential waiting times X^*_{ij}, $i = 1, 2$ and $j = 1, 2, \ldots$ (If desired, translation back to the discrete case always is immediate.) Generalization of the univariate MAX-chart then entails that boundary values t_i should be set such that if r of the X^*_{ij} fall below these lower limits, a signal will follow. However, note that this description is not yet complete and that various possibilities exist to make it precise. Below we shall discuss three of these.

Method 1. Here we mix the two univariate approaches. First we let

$$t_i = \frac{\log(a)}{\lambda_i}, \text{ with } a = 1 - (r\alpha)^{1/r}, \tag{3.2}$$

ensuring that $P(X^*_{ij} < t_i)$ equals $(r\alpha)^{1/r}$ during IC for $i = 1$ as well as for $i = 2$. Then we wait till r failures of the same type have occurred and signal if all of these fall below the relevant t_i. If not, we try a next group, obviously with the type we just checked in the previous group starting at 0 again, but continuing with the other type from the point at which we were. Clearly, in this way $FAR = r\alpha$ will hold again. Next suppose that during OoC the λ_i have become $\theta_i\lambda_i$, with $\min(\theta_1, \theta_2) > 1$. The probability that an arbitrary failure is of type i obviously equals $\pi_i = \lambda_i/\lambda (= p_i/p$, with $p = p_1 + p_2)$. But then the same will hold for a group of size r: its probability of being of type i is π_i as well. Consequently, during OoC the alarm rate turns into $P_1 = \pi_1(1 - a^{\theta_1})^r + \pi_2(1 - a^{\theta_2})^r$, which in analogy to (2.3) suggests that the average run length equals

$$ARL_1 = \frac{r}{\pi_1(1 - a^{\theta_1})^r + \pi_2(1 - a^{\theta_2})^r}. \tag{3.3}$$

This is indeed correct, but observe that some additional argument is needed, since we no longer simply have that RL/r is $G(P_1)$. In fact, note that for the two independent sub-processes there is in the first place an intensity λ_i for the occurrence of events, and then moreover at each r^{th} event a probability $(1 - a^{\theta_i})^r$ of giving an alarm, thus resulting in an effective alarm frequency $\lambda_i(1 - a^{\theta_i})^r/r$. For the joint process, these two intensities are simply added and their inverse already almost provides the desired ARL in (3.3). The only remaining difference is the presence of λ_i rather than of $\pi_i = \lambda_i/\lambda$, but this is easily explained: here we have worked on the time scale. Going back to the scale based on the numbers of events (as in (2.2) and (2.3)) indeed requires a factor λ (also cf. the remarks above for the univariate case).

Note that the choice for t_i from (3.2) obviously ensures that in each of the two separate processes we have $ARL = 1/\alpha$ during IC as well. The fact that after combining the two, we still wind up with $1/\alpha$ in (3.3) during IC, is due to our working on the scale of the numbers of events. Once we shift to the time scale, the individual ARL's become $1/(\alpha\lambda_i)$ during IC, while the joint process produces the smaller value $1/(\alpha\lambda)$ in that case. Hence in this sense the higher intensity of the combined process does lead to stopping more quickly on average (cf. the discussion at the end of section 2).

Method 2. An alternative is to just consider the joint process with intensity λ. Hence we are in fact back at the univariate case, with a single sequence of waiting times X^*_j. Thus we apply $t = \log(a)/\lambda$ (cf. (3.2)) to each group of r consecutive failures, no matter of which of the two

types these are, which again produces $FAR = r\alpha$. Straightforward application of (2.3) leads to

$$ARL_2 = \frac{r}{\{1-a^{\theta}\}^r}, \text{ with } \theta = \pi_1\theta_1 + \pi_2\theta_2, \tag{3.4}$$

where the last step follows from observing that during OoC λ changes into $\theta_1\lambda_1 + \theta_2\lambda_2 = \theta\lambda$.

Method 3. Yet another possibility is to apply a combination of Methods 1 and 2. As in Method 2, use groups of r failures in the joint process. But then shift to Method 1: for each of these r failures, consider the time elapsed, not since the previous failure, but since the previous failure of the same type. In other words, use the separate waiting times X^*_{ij}, rather than the X^*_j. Judge these by means of the t_i from (3.2) again in order to ensure that here as well $FAR = r\alpha$ and arguing as above obtain that now

$$ARL_3 = \frac{r}{\{\pi_1(1-a^{\theta_1}) + \pi_2(1-a^{\theta_2})\}^r}. \tag{3.5}$$

The obvious question is which of the three Methods is best. Before attempting to answer it, first note that, although we have restricted attention to the bivariate case for simplicity, generalization to $k > 2$ is immediate. In fact, let Q be a r.v. such that $P(Q = \theta_i) = \pi_i$ for $i = 1, \ldots, k$, where $\min(\theta_1, \ldots, \theta_k) > 1$ and $\Sigma_{i=1}^k \pi_i = 1$, then

$$ARL_1 = \frac{r}{E(1-a^Q)^r},\ ARL_2 = \frac{r}{(1-a^{EQ})^r} \text{ and } ARL_3 = \frac{r}{\{E(1-a^Q)\}^r}. \tag{3.6}$$

As $g(y) = y^r$ is convex, Jensen's inequality implies that $E(1-a^Q)^r \geq \{E(1-a^Q)\}^r$, and thus that $ARL_1 \leq ARL_3$. Likewise, $g(y) = 1 - a^y$ is concave and hence $1 - a^{EQ} \geq E(1-a^Q)$, meaning that $ARL_2 \leq ARL_3$. Hence part of the answer to our question is simple: Method 3 is inferior to the other two and will not be considered any further.

The relation between Methods 1 and 2 is more complicated, however. We have:

Lemma 3.1. *Let* $b = b(r,\alpha) = \log(r)/\log(1/a)$, *with* $a = 1 - (r\alpha)^{1/r}$, *then* $ARL_1 \geq ARL_2$ *if* $\min(\theta_i) \geq b$ *and* $ARL_1 \leq ARL_2$ *if* $\max(\theta_i) \leq b$.

Proof. Let $g(y) = (1-a^y)^r$, then $g"(y) \leq 0$ iff $a^y - 1/r \leq 0$, i.e. iff $y \geq b$. If $\min(\theta_i) \geq b$ then all values of Q are $\geq b$. Hence in that situation we may conclude that $Eg(Q) = g(EQ) + E\{(Q-EQ)g'(EQ)\} + \frac{1}{2}E\{(Q-EQ)^2 g"(Q^*)\} \leq g(EQ)$, as the r.v. Q^* also only attains values $\geq b$. In view of (3.6), this result indeed implies that $ARL_1 \geq ARL_2$. The second result follows in exactly the same manner. □

Clearly, if the values of Q are not all on the same side of b, the matter remains undecided. As concerns the behavior of $b = b(r,\alpha)$, we note that it decreases in α and, for all practical purposes, in r as well. This second remark requires some clarification, since at first sight it looks a bit weird, because $b(1,\alpha) = 0$. (Hence the lemma implies $ARL_1 \geq ARL_2$ for $r = 1$; this is indeed correct, as we in fact already concluded after (2.10) by simply using the concavity of $g(y) = 1 - a^y$.) But how can $b(r,\alpha)$ decrease in r if it starts at 0 for $r = 1$? The explanation is that $b(y,\alpha)$ rises steeply in y immediately after 1. In fact, writing $b(y) = \log(y)/h(y)$, with $h(y) = \text{-log}(1-(y\alpha)^{1/y})$, leads to $b'(y) = 1/(yh(y)) - \log(y)h'(y)/h^2(y)$. Hence $b'(1) = 1/h(1) = -1/\log(1-\alpha) \approx 1/\alpha$, which is indeed very large. Using e.g. Maple it can be verified that over the region of interest for α (typically (0.001, 0.01)), a maximum between 7 and 50 is reached well before $y = 1\frac{1}{2}$, after which the decrease mentioned above begins. In Table 3.1 we present some values of b for various r and α of interest.

Table 3.1 Values of $b = b(r, \alpha)$ for various r and α

$\alpha \backslash r$	1	3	5	7
0.001	0	7.05	3.78	2.87
0.01	0	2.93	2.02	1.69

Hence from Table 3.1 we e.g. can conclude that in case of $r = 3$ and $r = 5$ the minimal value will be ARL_1 for $k = 2$, $(\theta_1, \theta_2) = (1, 3)$ and $\alpha = 0.001$, while the winner is ARL_2 for these same values of r and $(\theta_1, \theta_2) = (3, 5)$, $\alpha = 0.01$. To provide more information, in Table 3.2 we have assembled the outcomes for representative values of r, α and (θ_1, θ_2).

Table 3.2. Values ARL_1 (upper) and ARL_2 (lower) from (3.6) for a range of r, α and (θ_1, θ_2), using $\pi_1 = \pi_2 = \frac{1}{2}$. (A '*' means that $ARL_1 > ARL_2$)

r	1	3	5	7	1	3	5	7
(θ_1, θ_2)		$\alpha = 0.001$				$\alpha = 0.01$		
(1,2)	667*	271	150	107	66.9*	34.4	26.9	25.5
	667	332	214	162	66.8	38.3	30.3	27.5
(1,3)	500*	109	50.0	36.7	50.4*	18.0	15.0	16.1
	500	156	80.9	56.4	50.2	20.7	15.6	14.6
(1,5)	334*	37.2	18.7	17.6	33.9*	9.49	10.4*	13.4*
	334	57.7	25.7	18.7	33.7	9.86	8.08	8.76
(2,4)	334*	50.4	23.4	18.9*	33.7*	9.86	8.84*	9.93*
	334	57.7	25.7	18.7	33.7	9.86	8.08	8.76
(1.7)	250*	20.3	12.9	14.7*	25.7*	7.28*	9.70*	13.1*
	250	30.1	13.7	11.3	25.4	6.46	6.17	7.51
(3.5)	250*	28.5	13.8*	12.1*	25.4*	6.62*	6.54*	7.88*
	250	30.1	13.7	11.3	25.4	6.46	6.17	7.51

Quite a few observations can be made from Table 3.2. To begin with, some are already known from the univariate case, which is represented here by ARL_2, where $\theta = (\theta_1 + \theta_2)/2$ (hence the same results on the second rows for (1,5) and (2,4) and for (1,7) and (3,5)). Indeed $ARL \approx 1/(\theta\alpha)$ for $r = 1$, which allows much improvement by letting $r > 1$. Moreover, indeed r^{opt} from (2.4) decreases in θ: for the largest θ from Table 3.2 we see that $r = 7$ is already too large, in the sense that $r = 5$ is better. However, more interesting of course are the observations about the new aspects, concerning the bivariate behavior. The pattern predicted by Lemma 3.1 and Table 3.1 becomes nicely visible in Table 3.2 through the added '*' whenever $ARL_1 > ARL_2$ (sometimes not visible from the values themselves, as these are only given to a precision which makes sense in practice). Indeed the first column ($r = 1$) is starred, as well as a lower right-hand triangle, which is considerably larger for $\alpha = 0.01$ than for $\alpha = 0.001$. Also interesting to note is that for Method 1 slightly smaller r are optimal than for Method 2. This is quite understandable: e.g. a $\theta = 4$ for Method 2 can still use a rather large r. But if the corresponding (θ_1, θ_2) are (1,7), it means that Method 1 focusses on $\theta_2 = 7$ and thus a smaller r is better.

Summarizing, we can say that over the region of interest considered here, small α combined with small θ_i suggests to use Method 1, while a combination of large α and large θ_i favors Method 2. If one of the two is large and the other small, there is not much difference between the two Methods. A distinct overall advantage of Method 1 is that if the θ_i are quite different (e.g. $(\theta_1, \theta_2) = (1, 7)$), this approach is likely to pinpoint the type of failure which has caused the *OoC* signal. Method 2 works in a purely univariate way (e.g. with $\theta = 4$), and thus the occurrence of a signal merely tells that something has happened to the joint process.

4 The dependent case

In the previous section we singled out the essentially independent case of mutually exclusive types of failures and demonstrated that even in this simple situation not all is obvious. Now we address the general case: let (D_{1j}, D_{2j}), $j = 1, 2, \ldots$ be a sequence of independent identically distributed pairs of r.v.'s with $P(D_{i1} = 1) = 1 - P(D_{i1} = 0) = p_i, i = 1, 2$ and $P(D_{11} = 1, D_{21} = 1) = q$ during IC. Clearly, for $q = 0$ we are back in section 3, while for $q = p_1 p_2$ we have exact independence of the D_i's. As in our high-quality setup both p_i are (very) small, in this latter situation as well the interaction is negligible. In fact, even for $q > p_1 p_2$, this remains the case as long as q is of a smaller order than p_1 and p_2. For the interaction to have real impact, we need $P(D_{21} = 1 | D_{11} = 1)$ to be of order 1. In other words, $p_{10} = P(D_{11} = 1,$ $D_{21} = 0)$, $p_{01} = P(D_{11} = 0,\ D_{21} = 1)$ and $p_{11} = q = P(D_{11} = 1,\ D_{21} = 1)$ have to be of the same order of magnitude. Then for example the correlation within a pair of D_i's to first order equals $\{(1 + p_{10}/p_{11})(1 + p_{01}/p_{11})\}^{-\frac{1}{2}}$, which is indeed bounded away from 0.

The conclusion thus is that in high-quality applications we can ignore the interaction, unless the frequency of joint failures is of the same order as those of the separate failures. Consequently, from now on we will assume this to be the case. Again adopting the Poisson approach from the previous section, we thus begin by considering two dependent Poisson processes, with intensities given by $\lambda_1 = \lambda_{10} + \lambda_{11}$ and $\lambda_2 = \lambda_{01} + \lambda_{11}$ (with the obvious correspondence of the λ_{ij} to the p_{ij} above, $i, j = 0, 1$). The joint process hence has intensity $\lambda = \lambda_{10} + \lambda_{01} + \lambda_{11} < \lambda_1 + \lambda_2$. Straightforward application of Method 1 becomes awkward here: the sub-processes are no longer independent and a simple result like (3.3) does not seem to be available. On the other hand, Method 2 continues to work without problem. As mentioned before (3.6), generalization to $k > 2$ is straightforward, so let $k = 3$ and e.g. $\lambda_1 = \lambda_{10}$, $\lambda_2 = \lambda_{01}$ and $\lambda_3 = \lambda_{11}$. Indeed, these three Poisson processes are independent again and ARL_2 readily follows from (3.4) and its generalization (3.6) for the r.v. Q with $k = 3$.

However, observe that this step from $k = 2$ to $k = 3$ also immediately shows how to successfully adapt Method 1. Just monitor each of the three sub-processes, using t_j from (3.2) for $i = 1$, 2 and 3, and use ARL_1 as generalized by (3.6) from (3.3). Do note as well that not only its tractability makes this approach more attractive than the dependent one for $k = 2$. Another distinct advantage is (cf. the remark at the end of section 3) that in case of a signal it makes transparent which of the three categories is/are probably responsible for the process going OoC. Quite often, this type of information is considered to be very useful. Rather than merely knowing that e.g. λ_2 has increased (which information might follow from the dependent approach using $k = 2$), it is interesting to see whether this is (mainly) due to the increase of the separate part λ_{01}, or rather to that of the common part λ_{11}.

This in fact is a quite general issue, which for example also occurs in the continuous data case of monitoring bivariate normal pairs (X_1, X_2) with correlation ρ. Here as well it is of interest to know to what extent an increase during OoC in e.g. $\mu_2 = EX_2$ is in fact due to an increase of $\mu_1 = EX_1$. In other words, did $\mu_2 - \rho\mu_1$ increase as well? Returning to the attribute data case, yet another remark on this aspect concerns the relation to risk adjustment (see e.g. Albers (2011b) and the further references contained in that paper). In this situation, the issue of whether a certain risk has increased (e.g. surgical failures) is related to a possible increase of risk in some underlying factor (e.g. degree of illness of patients treated). Hence although this concerns a univariate case in the sense that a single characteristic is monitored, the simultaneous consideration of such a characteristic together with an underlying risk factor, does give it a distinct bivariate flavor as well.

Summarizing, as long as the probability of a joint failure is of a lower order of magnitude than that of the individual failures, the dependence can be ignored and Methods 1 and 2 from the previous section can be applied without any change. If this is not the case, the solution is

to consider the three separate categories 'only first', 'only second' and 'both' for the failures. Once again Methods 1 and 2 can be used, but now for the generalized version with $k = 3$ rather than $k = 2$. No new technicalities are involved: Lemma 3.1 and Tables 3.1 and 3.2 still give a good impression of what to expect and explicit results can be evaluated through (3.6). Hence for briefness' sake we shall refrain from adding tables involving configurations $(\theta_1, \theta_2, \theta_3)$.

5 The nonparametric version

In sections 2-4 we used the basic homogeneous case as our starting point. It remains to remove this often too optimistic assumption and to present the empirical nonparametric versions of Methods 1 and 2. These will offer the robust alternative to existing methods, as advocated in section 1. Similar steps have been performed in our earlier papers dealing with nonparametric proposals for this area (Albers and Kallenberg (2009) and Albers (2011a)). Hence we shall be quite brief here, referring to these papers for further details and derivations.

Starting with the univariate case from section 2, we hence drop the assumption that the X_j are $G(p)$. Instead, we just have some unknown underlying distribution function (df) F. Consequently, we need some Phase I sample $X_1, \ldots, X_m$, providing us with the corresponding empirical df $F_m(x) = m^{-1}\#\{X_i \leq x\}$ and the quantile function $F_m^{-1}(t) = \inf\{x|F_m(x) \geq t\}$. Note that $F_m^{-1}(t)$ equals $X_{(i)}$ for $(i-1)/m < t \leq i/m$, with $X_{(1)} < \ldots < X_{(m)}$ the order statistics for the sample. Hence any q-quantile $F^{-1}(q)$ can be estimated by $F_m^{-1}(q) = X_{(s)}$, with $s = \{mq]$ and $\{z]$ denoting the smallest integer $\geq z$. For the MAX-chart, (2.1) at present implies that $FAR = F^r(n) = r\alpha$, and thus that the lower bound n is given by $F^{-1}((r\alpha)^{1/r})$. As a result, we obtain for the estimated bound

$$\hat{n} = X_{(s)}, \text{ with } s = \{m(r\alpha)^{1/r}]. \tag{5.1}$$

The empirical nonparametric version of the MAX-chart now starts monitoring at X_{m+1}, X_{m+2}, ...: if $\max(X_{m+1}, \ldots, X_{m+r}) \leq X_{(s)}$, an alarm will follow; if not, we consider the next batch of size r. Note that (5.1) readily shows that the choice $r = 1$ leads to extreme order statistics, as s will typically equal 1 for sample sizes m encountered in practice. Choosing $r > 1$ indeed solves this problem and thus effectively avoids the aforementioned huge estimation errors (e.g. if $r = 3$, customary values like $m = 100$ and $\alpha = 0.001$ produce a not at all extreme value $\{14.4] = 15$ for s). A final remark about the univariate case is that using X_j^* which are Exp(λ) rather than X_j which are $G(p)$ as a starting point, obviously makes no difference whatsoever in going to a nonparametric version. We just translate (5.1) into $\hat{t} = X_{(s)}^*$, with the same $s = \{m(r\alpha)^{1/r}]$.

Next we move on to the extension to the bivariate case. Actually, this is rather straightforward, so we will cover the extension to $k > 2$ in this same step. Hence we consider k independent Poisson processes with intensities λ_i, a joint Poisson process with intensity $\lambda = \Sigma_{i=1}^k \lambda_i$, and exponential waiting times X_{ij}^*, $i = 1, \ldots, k$ and $j = 1, 2, \ldots$. Here we will need Phase I samples of size m_i from the respective sub-processes. Clearly, these m_i should all be of the same order of magnitude, in order to ensure that this will hold for the estimation errors for the various i as well. An alternative approach is to use a single Phase I sample of size m from the joint process. In the latter case, the sample sizes of for the sub-processes will be r.v.'s M_i having expectation $m\pi_i$. According to the discussion in section 4, the various λ_i are supposed to be of the same order of magnitude. This means that the π_i will be bounded away from 0 and thus the M_i will be of the same order of magnitude as well. Incidentally, note that this strengthens the conclusion from section 4 that joint failures which have a probability of lower order compared to the individual failures can be ignored. That is not just a possibility, but in fact more of a

necessity: such rare cases would require far too much time to collect a sample of sufficiently large size and thus lead to relatively very large contributions to the resulting total estimation error.

To be specific, suppose we use fixed m_i and let $m = \Sigma_{i=1}^{k} m_i$. Let $X_{i(j)}$ be the j^{th} order statistic from the i^{th} sample, $i = 1, \ldots, k$ and $j = 1, \ldots, m_i$. Then for Method 1 we replace the t_i from (3.2) by

$$\hat{t}_i = X^*_{i_{(s_i)}}, \text{ with } s_i = \big\{m_i(r\alpha)^{1/r}\big]. \tag{5.2}$$

For the still univariate Method 2, in fact nothing has changed: $\hat{t} = X^*_{(s)}$, with $s = \{m(r\alpha)^{1/r}]$. and $X^*_{(j)}$ the j^{th} order statistic in the total sample of size m. Indeed, in this sense Method 2 has an advantage over Method 1: the former requires estimation of just one quantile, whereas the latter involves estimation of k quantiles. Clearly, this is the price to be paid for the ability of Method 1 to pinpoint the likely cause of going *OoC* (cf. the remark at the end of section 3).

Note that the above provides all we need for the implementation of the chart. It remains as simple as before, with only a replacement of the t_i or t by $\hat{t}_i$ or $\hat{t}$, respectively. For the in fact univariate Method 2 we readily may conclude from Albers (2011a) that its *ARL*, conditional on the Phase I sample $X_1, \ldots, X_m$, is stochastic and distributed as $r/U^r_{(s)}$. Here $U_{(1)}, \ldots, U_{(m)}$ are order statistics from a sample of size m from the uniform distribution on $(0, 1)$. Hence, as always happens once estimation is involved, performance characteristics like *ARL* are now no longer fixed at a given value such as $1/\alpha$, but instead have become random. However, do note that the fact that *ARL* is distributed as $r/U^r_{(s)}$ indeed shows the chart to be nonparametric. Next, it is straightforward to extend this type of result to the situation of Method 1. There we arrive at an *ARL* distributed as $r/\{\Sigma_{i=1}^{k} \pi_i U^r_{i_{(s_i)}}\}$, with the $U_{i_{(s_i)}}$ once more uniform $(0, 1)$ order statistics, but now from k independent samples of sizes m_i, $i = 1, \ldots, k$, respectively. Obviously, in either case $ARL \xrightarrow{P} 1/\alpha$, implying that both charts perform as they should, which concludes our treatment as far as the straightforward application of the nonparametric versions is concerned.

What remains is the optional discussion of how to manage the effects of the estimation error, which were seen to be nonnegligible. In fact, for the nonparametric case these were even huge and the step from sum to maximum was required to bring such errors back to acceptable proportions. An attractive way to control these effects is by monitoring left exceedance probabilities like $P_{Exc} = P(ARL < 1/\{\alpha(1 + \varepsilon)\})$ for some small $\varepsilon > 0$, like $\varepsilon = 0.25$. For example, we could stipulate that $P_{Exc} \leq \beta$ for some small $\beta > 0$, e.g. $\beta = 0.2$. Questions to be answered are what values of m_i and m are minimally required to achieve this for given ε and β. Or, if the m_i or m are simply given, how can small corrections be made to the $\hat{t}_i$ or $\hat{t}$ in order to ensure that the upper bound β is respected after all. For the univariate Method 2, this material can be taken from Albers (2011a) without change. For Method 1, some adaptation is required, but this is quite straightforward. Hence for briefness' sake, we shall not present it here.

References

Albers, W. (2010). The optimal choice of negative binomial charts for monitoring high-quality processes. *J. Statist. Planning & Inference* **140**, 214-225.

Albers, W. (2011a). Empirical nonparametric control charts for high-quality processes. *J. Statist. Planning & Inference* **141**, 3151-3159.

Albers, W. (2011b). Risk adjusted control charts for health care monitoring. *Int. J. of Math. and Math. Sciences*, 16 pages, doi: 10.1155/2011/895273.

Albers, W. and Kallenberg, W.C.M. (2009). *CUMIN* charts. *Metrika* **70**, 111-130.

Chiu, J.-E. and Kuo, T. (2010). Control charts for fraction nonconforming in a bivariate binomial process. *J. Appl. Statist.* **37**, 1717-1728.

Lu, J.-C. and Bhattacharyya, G.K. (1991). Inference procedures for bivariate exponential model of Gumbel. *Statist. & Prob. Letters* **12**, 37-50.

Shaha, S H. (1995). Acuity systems and control charting. *Qual. Manag. Health Care* **3**, 22-30.

Sonesson, C. and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *J. R. Statist. Soc. A* **166**, 5-21.

Steiner, S.H., Cook, R.J. and Farewell, V.T. (1999). Monitoring paired binary surgical outcomes using cumulative sum charts. *Statist. Med.* **18**, 69-86.

Thor, J., Lundberg, J., Ask, J., Olsson, J., Carli, C., Härenstam, K.P. and Brommels, M. (2007). Application of statistical process control in healthcare improvement: systematic review", *Qual. & Safety in Health Care* **16**, 387-399.

Xie, Y.J., Tsui, K.L., Xie, M. and Goh, T.N. (2010). Monitoring time-between-events for health management. *Prognostics & System Health Man. Conf.*, IEEE, 1-8.

Xie, Y.J., Xie, M. and Goh, T.N. (2009). A MEWMA chart for a bivariate exponential distribution. *Proceedings of the 2009 IEEE IEEM*, 424-428.