

# Non-parametric Quality Control Statistics

David McDonald, Jun Mao, Mahmoud Zarepour

November 21, 2008

# On-line Quality Control

- ▶ Quality measurements  $X_1, X_2, \dots$  are made sequentially.
- ▶ We want to signal an alarm quickly if there is an out of control situation; i.e. when the distribution changes.
- ▶ On the other hand we don't want false alarms too often.

# Cusum: Optimal Parametric Procedure

- ▶ For univariate standard normal data a shift in mean to  $\mu$  is best detected by the Cusum.
- ▶ Sequentially calculate  $C_{n+1} = \max\{0, C_n + (X_{n+1} - k)\}$
- ▶ Signal an alarm if  $C_{n+1} > h$ .
- ▶  $k$  is an anchor value,  $\mu/2$  is optimal.
- ▶  $h$  is the threshold.

# Performance of the Cusum

- ▶ If  $k = .5$  and  $h = 4.38$
- ▶ then the on-target average run length is around 500 observations
- ▶ If the off-target shift is  $\mu = 1$  standard deviations
- ▶ then on off-target average run length is around 9.2 observations.

# A nonparametric Cusum

- ▶ Naval Research Logistics, Vol. 37, 627-646 (1990).
- ▶ Calculate the sequential rank  $R_{n+1}$  of observation  $X_{n+1}$  among  $X_1, \dots, X_n$ .
- ▶ Calculate  $U_{n+1} = R_{n+1}/(n+1)$
- ▶  $C_{n+1} = \max\{0, C_n + (U_{n+1} - k)\}$
- ▶ Signal if  $C_{n+1} > h$ .

# Performance of the non-parametric Cusum

- ▶ For  $k = .643$  and  $h = 1.20$
- ▶ The on-target average run length is around 500 observations
- ▶ If the off-target shift is  $\mu = 1$  standard deviations
- ▶ then on off-target average run length is around 10 observations.

# Multivariate Quality Data

- ▶ Now suppose the  $X_1, X_2 \dots$  are vectors.
- ▶ Even if the data is multivariate normal one probably doesn't know the covariance.
- ▶ Some of the components could even be discrete.
- ▶ **It would nice to have a non-parametric procedure**

# Voronoi diagrams

- ▶ The Voronoi tessellation of  $n$  points  $X_1, X_2, \dots, X_n$  in  $\mathbf{R}^d$  is a set of  $n$  regions  $G_1, G_2, \dots, G_n$
- ▶ where region  $G_j$  denotes the set points closest to arrival  $j$ ; i.e. to  $X_j$ .
- ▶ If a new point  $X_{n+1}$  is observed and falls into region  $G_j$  then define  $R_{n+1} = j$  and  $U_{n+1} = R_{n+1}/(n+1)$ .
- ▶ If the points  $X$  are chosen in an i.i.d. manner then  $R_{n+1}$  is uniformly distributed among the values  $\{1, 2, \dots, n\}$  and  $U_{n+1}$  is approximately uniform.

# A Multivariate Cusum

- ▶ Let  $M_{n+1}^c$  be the average of  $\{U_i : X_i \text{ is among the } c \text{ closest to } X_{n+1}\}$  (including  $X_{n+1}$ ).
- ▶ Typically we take  $c = 9$ .
- ▶ Then add the new point to the tessellation at the center of the new region  $G_{n+1}$ .
- ▶ If the sequence of points  $X_1, X_2, \dots, X_n, X_{n+1}$  is i.i.d. then the  $M_{n+1}^c$  is approximately the average of  $c$  i.i.d. uniforms and hence approximately normally distributed with mean 0.5 and variance  $(12c)^{-1}$ .
- ▶ We can use the  $M$ 's to construct a standard normal Cusum:  
$$C_{n+1} = \max\{0, C_n + \sqrt{12c}(M_{n+1}^c - 0.5) - k\}.$$
- ▶ Again the procedure will signal an alarm if  $C_{n+1} \geq h$ .

# Performance of the Multivariate Cusum

- ▶ The value of the anchor  $k = .5$  and the signal level  $h = 4.38$  are chosen to set the on-target run length of 500.
- ▶ The on-target average run lengths are predicted very well.
- ▶ If we shift from a bivariate normal with mean  $(0, 0)$  to mean  $(1, 1)$  the off-target run length is about 20.
- ▶ Note that to achieve this univariately would require two Cusums with average run lengths of 1000 ( $h \approx 5$ ).
- ▶ But then each would have an off-target ARL of about 12 so the ARL of two charts would be about just under 12.
- ▶ The parametric procedure is better of course but not vastly better.