Publication information

Title	Directional Control Schemes for Multivariate Categorical Processes
Author(s)	Li, Jian; Tsung, Fugee; Zou, Changliang
Source	Journal of Quality Technology , v. 44, (2), April 2012, p. 136-154
Version	Published version
DOI	Nil
Publisher	American Society for Quality (ASQ)

Copyright information

Notice

Directional Control Schemes for Multivariate Categorical Processes

JIAN LI and FUGEE TSUNG

Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

CHANGLIANG ZOU

Nankai University, Tianjin, China

We consider statistical process control of multivariate categorical processes and propose a Phase II log-linear directional control chart that exploits directional shift information and integrates the monitoring of multivariate categorical processes into the unified framework of multivariate binomial and multivariate multinomial distributions. We also suggest a diagnostic scheme for identifying the shift direction. Both the control chart and the diagnostic approach are simple and easily computed. Numerical simulations and real applications are presented to demonstrate their effectiveness.

Key Words: Contingency Table; EWMA; Generalized Likelihood-Ratio Test; Multivariate Multinomial Distribution; Statistical Process Control.

Introduction

THE OVERALL quality of most modern processes is best described in terms of multiple characteristics. Multivariate statistical process control (SPC) is applied in practical situations where several quality characteristics must be monitored simultaneously. Much effort has been devoted to the monitoring problem in settings where all the observed variables are numerical and continuous. For instance, tensile strength and diameter, which have been assumed to follow a bivariate normal distribution, are two important quality characteristics of a textile fiber that must be jointly controlled. Refer to Lowry and Montgomery (1995) and Bersimis et al. (2007) for thorough reviews of monitoring multivariate continuous processes.

In manufacturing and especially in service industries, however, it is increasingly common to have quality characteristics that cannot be measured numerically. Although obtaining their precise continuous values may be expensive, unnecessary, or even impossible, collecting some attribute data related to them may be quite feasible. Such rough classification levels do not require precise measurements. Examples include items on a production line of which several quality characteristics are each evaluated as conforming or nonconforming according to predefined specifications and multiple indexes in a service flow that may be assessed as excellent, acceptable, or unacceptable. The characteristics all have two or more attribute levels, so that the data are multivariate categorical. Note that this is similar to having multiple factors in design of experiments (DOE), where each factor has several specific levels. Here, for simplicity, we use "factor" to designate a categorical characteristic.

The p-chart and the np-chart for binomial distributed variables, together with the c-chart and the u-chart for Poisson processes, are typical statistical process control (SPC) tools used with univariate categorical processes. To monitor multiple factors, we may employ multiple univariate categorical control charts as a multichart (see Woodall and Ncube

Mr. Li is a Doctoral Student in the Department of Industrial Engineering and Logistics Management. He is a student member of ASQ. His email is jianli@ust.hk.

Mr. Tsung is a Professor in the Department of Industrial Engineering and Logistics Management. He is a Fellow of ASQ. His email is season@ust.hk.

Dr. Zou is an Assistant Professor in LPMC and Department of Statistics, School of Mathematical Sciences. His email is chlzou@yahoo.com.cn. Dr. Zou is the corresponding author.

(1985)). However, this may be an unattractive approach for two reasons: First, it is difficult to determine the control limit to achieve a desired average run length (ARL), which is defined as the average number of samples until the control chart signals. The control limits of the separate univariate charts must be set such that each chart achieves a specific individual in-control (IC) ARL, so that the overall IC ARL of the multichart is the desired value. Determining these control limits is nontrivial even for the low-dimensional case of a small number of categorical factors, let alone for a general multivariate categorical distribution. This complexity also increases dramatically when the marginal distributions of the categorical factors are not identical, because the individual charts have different run-length distributions. Also, a multichart considers individual categorical factors in parallel and therefore is unable to account for any correlations among them. Naturally, it is desirable to introduce multivariate categorical control charts that can appropriately describe and exploit the relationships among multiple categorical factors.

Woodall (1997) summarized many aspects of the control charts for attribute data, but mostly considered univariate categorical charts. In the literature, some efforts have been devoted to monitoring multinomial and multiattribute processes. See Topalidou and Psarakis (2009) for a good overview. Among others, Patel (1973) suggested a χ^2-chart for multivariate binomial or multivariate Poisson populations, which is based on the assumption that the joint distribution of the correlated binomial or Poisson variables can be approximated by a multivariate normal distribution given a sufficiently large sample size. More recent developments include the Shewhart-type *mnp*-chart proposed by Lu et al. (1998), which uses the weighted sum of the number of nonconforming units for each quality characteristic. There is also the *mp*-chart designed by Chiu and Kuo (2008) for multivariate Poisson count data. However, as with Patel's χ^2-chart, these two charts can only deal with factors having two levels. On the other hand, some methods focus on monitoring multinomial processes that have only one factor having more than two levels, such as the generalized *p*-chart developed by Marcucci (1985). This chart extends the traditional *p*-chart by adopting the Pearson chi-square statistic. The multinomial cumulative sum (CUSUM) chart proposed by Ryan et al. (2011) can also deal with this problem and is based on the likelihood-ratio statistic equipped with a CUSUM scheme. Note that the prior approaches for monitoring multinomial processes are actually univariate charts because only one factor is involved. With several factors, at least one of which has more than two levels, there are no appropriate approaches available. In addition to this deficiency, most of the existing methods focus entirely on the marginal sums with respect to each categorical factor, neglecting the cross-classifications between factors. Because of this, if some cross-classification probabilities shift to out-of-control (OC) states, these charts may not detect them quickly. In summary, a general monitoring methodology for multivariate categorical processes is needed.

Multi-way analysis of variance (ANOVA) models in DOE may provide some assistance if the observations are assumed to be linearly dependent on the levels of several factors (i.e., if they depend on main-factor effects and interaction effects). As with the observations in an ANOVA model, the logarithms of the cross-classification probabilities may also depend linearly on the levels of these multiple factors, as in a log-linear model (see Bishop et al. (2007)). The log-linear model effectively characterizes the association and interaction patterns among the categorical factors and therefore can be used to develop multivariate categorical control charts. Our paper is not the first application of log-linear models in SPC. Qiu (2008) dealt with distribution-free monitoring schemes for multivariate continuous processes by dichotomizing continuous multivariate data into categorical data. He then estimated the IC factor distribution using log-linear models in Phase I SPC and proposed a Phase II multivariate CUSUM chart by employing the Pearson chi-square statistic.

In our study, a systematic directional monitoring mechanism and a diagnostic scheme for multivariate categorical processes were developed based on log-linear models. By analogy with multiway ANOVA, the cross-classification probabilities are expressed in terms of main-factor effects and factor-interaction effects. The log-linear model can then be equivalently rewritten as a regression model, leading to a one-to-one correspondence between factor effects and coefficient subvectors. Potential shifts to OC states in the factor effects therefore appear in their corresponding coefficient subvectors, which constitutes prior information on the potential shift directions. To monitor a process as efficiently as possible, such practical information on the shift directions should be exploited. For these purposes, we develop a Phase II control chart based on the log-likelihood function of the log-linear model and an exponentially-weighted moving average (EWMA) scheme. The proposed control chart provides a unified framework for handling multiple factors when at least one has more than two levels. Using certain approximations, the resulting chart is easy to construct and convenient to implement. Furthermore, we present a diagnostic scheme to identify the shift direction following an OC signal. Some implementation guidelines are provided and illustrated using a practical example. Monte Carlo simulations were performed to demonstrate the effectiveness of the proposed chart and the diagnostic approach.

Multivariate Categorical Processes and Conventional Monitoring Approaches

Multivariate Categorical Processes

We first illustrate the multivariate categorical process with two motivating examples. Aluminium electrolytic capacitor (AEC) manufacturing is a multistage process, and the quality of the semi-finished AECs is inspected immediately after each stage. Here, for illustration, we concentrate on the quality after the aging stage, which is assessed mainly in terms of leakage current (LC), dissipation factor (DF), and capacity (CAP). Each characteristic is classified automatically as conforming or nonconforming to the specifications by an electronic device at a very high speed, and engineers are reluctant to obtain their precise continuous or numerical values (which is costly but not impossible). Consequently, this is a multivariate categorical process with three factors (LC, DF ,and CAP), each with two levels and therefore $2^3 = 8$ level combinations. Without loss of generality, for each factor, -1 represents "conforming" and 1 "nonconforming". For example, the combination $(-1, 1, -1)$ means an AEC with conforming LC and CAP and nonconforming DF.

Another example is the quality control of welding rods. One of the key aspects of welding-rod inspection is their appearance, which directly reflects the integrated level of welding-rod manufacturing and influences the welding performance. The welding rod is composed of a cylindrical metallic core wire and a coating composition (flux) covering the circumference of the metallic core wire. Its appearance has some important characteristics, such as eccentricity of the core wire, moisture resistance of the coating, strength of the coating, and rod bend. During testing, each is simply evaluated as either conforming or nonconforming, and their (latent) continuous values are not considered. With the four factors (eccentricity, moisture resistance, strength, and bend), each with two attribute levels, this is also a multivariate categorical process with $2^4 = 16$ cross-classification level combinations.

Now we turn to a general multivariate categorical process. Suppose that there are p factors, $C_1, \ldots, C_p$, and that each classification factor C_i has h_i possible levels. The overall cross-classifications among all the level combinations of these factors form a p-way $h_1 \times \ldots \times h_p$ contingency table with $h = \prod_{i=1}^{p} h_i$ cells. Each cell corresponds to one-level combination of the p factors and stores the count under this level combination.

For a simple $h_1 \times h_2 \times h_3$ three-way table, denote the observed count by n_{ijk} in cell(i, j, k) $(i = 1, \ldots, h_1;\ j = 1, \ldots, h_2;\ k = 1, \ldots, h_3)$ and its expectation by m_{ijk}. If the observations are made over a period of time, it is reasonable to assume that each cell count follows an independent Poisson distribution (see Bishop et al. (2007)). During the Phase II SPC monitoring process, the total sum of observations is usually fixed. Conditional on the total sum of cell counts, a series of independent Poisson distributions result in a multinomial distribution. So in one sample of size N, the cell counts in the three-way contingency table therefore follow the multinomial distribution $\text{MN}(N; p_{ijk})$ $(i = 1, \ldots, h_1;\ j = 1, \ldots, h_2;\ k = 1, \ldots, h_3)$. Here $p_{ijk} = m_{ijk}/N$ is the probability of an observation falling into cell(i, j, k), and these probabilities must sum to 1.

To generalize the three-way table to a general p-way $h_1 \times \ldots \times h_p$ contingency table (see Johnson et al. (1997)), let the probability of obtaining the combination of factor levels $a_1, \ldots, a_p$ be $p_{a_1 \ldots a_p}$ $(a_i = 1, \ldots, h_i$ and $i = 1, \ldots, p)$. Furthermore, denote the count of observations among a sample of size N with the combination $a_1 \ldots a_p$ by $n_{a_1 \ldots a_p}$. Clearly, the cell counts $n_{a_1 \ldots a_p}$ jointly follow an $\text{MN}(N; p_{a_1 \ldots a_p})$ distribution. Furthermore, consider, for example, the group of marginal counts $n_{(i)v} : v = 1, \ldots, h_i$ of the factor C_i, where $n_{(i)v}$ is defined as the sum of the cell counts $n_{a_1 \ldots a_p}$ over all levels of all factors other than C_i and for C_i at level v. It follows that this group of marginal sums follows the multinomial distribution $\text{MN}(N; p_{(i)1}, \ldots, p_{(i)h_i})$. Here $p_{(i)1}, \ldots, p_{(i)h_i}$ are the marginal probabilities of the factor C_i, which can be calculated in a similar way to $n_{(i)1}, \ldots, n_{(i)h_i}$ based on the cell probabilities $p_{a_1 \ldots a_p}$. The joint distribution of the p groups of variables $n_{(i)1}, \ldots, n_{(i)h_i}$ $(i = 1, \ldots, p)$, each being a multinomial distribution, is defined to be a multivariate multinomial distribution (see Johnson et al. (1997)). When each factor has two levels, this reduces naturally to the multivariate binomial distribution. So by applying multivariate binomial or multivariate multinomial distributions and a cross-classified contingency table, multivariate categorical processes can be studied.

Conventional Monitoring Approaches

Generally, statistical process control involves two phases (Montgomery (2009)). In Phase I, a set of process data are collected and examined. Any unusual patterns in the data are identified and, based on this, the data and the process may be adjusted, resulting in a clean dataset collected from stable process conditions. This dataset is called the IC dataset and used for estimating the IC parameters representative of the IC operating conditions. Phase I analysis primarily assists in bringing the data into a state of statistical control. Given the IC parameters, in Phase II, the process is monitored with control charts to detect and diagnose any deviations from the IC state. We review some typical methods for monitoring multivariate binomial and multivariate multinomial processes. Hereafter, we use the superscript "(0)" and "(1)" to denote the IC and OC states, respectively.

In a multivariate binomial process, each factor has two levels, giving rise to a p-way contingency table with 2^p cells for p factors. Denote the two levels of each factor by 1 and 0. Given a sample size N in Phase II, if the process is IC, the level 1 count of the factor C_i ($i = 1, \dots, p$) is binomially distributed with the total size N and its IC level 1 probability. So the IC mean of this count is known, and the IC covariance between any two factors C_i and C_j ($i, j = 1, \dots, p$ and $i \neq j$) can also be calculated, which gives the IC mean vector and the IC covariance matrix of the level 1 counts of the p factors. Based on these, Patel (1973) constructed the χ^2 charting statistic of the Hotelling's T^2 form, the expression for which can be found in Appendix A in the supplemental file (available at http://www.asq.org/pub/jqt/).

Factors with three or more levels are also common in production and service applications and they can be treated as multivariate multinomial processes. Consider customer attitudes toward a service, for instance. Suppose that there are four indexes, each of which may take the values of excellent, acceptable, or unacceptable. This forms a four-way contingency table with 3^4 cells. No appropriate monitoring methods exist that incorporate the cross-classifications among the p factors when at least one of them has more than two attribute levels.

The only feasible existing approach of monitoring multivariate multinomial processes, albeit a naive one, might be to monitor the p groups of marginal sums of each factor using p individual charts. If we consider only the group of marginal sums of the factor C_i ($i = 1, \dots, p$), it is a multinomial process, which could be handled by applying Marcucci's (1985) generalized p-chart. For the kth sample, the generalized p-chart employs the Pearson chi-square statistic as the charting statistic for the factor C_i, and its expression is also listed in Appendix A in the supplemental file. Finally, there would be p separate charts, which jointly form a multichart. Such a multichart would signal whenever at least one of the p individual charts signals.

Clearly, the χ^2-chart applies to only multivariate binomial processes and the multichart developed for multivariate multinomial processes is problematic. Moreover, both charts focus on the one-way marginal sums of each factor and almost entirely neglect the cross-classification interactions among factors.

New Methodologies for Monitoring and Diagnosis

Log-Linear Models

There is a clear need to model the relationship between each cell count and factor levels associated with it. The cell counts are stored in a multiway contingency table, as in standard multiway ANOVA, where responses to all factor level combinations are also placed in a multiway table. Standard multiway ANOVA is based on the assumption that the responses are normally distributed, and it aims to quantify how the responses are influenced by the main-factor effects and the factor-interaction effects. Consider for illustration a three-way ANOVA model, where the three factors take h_1, h_2, and h_3 levels, respectively. Denote the expected response with the first factor at its ith level, the second factor at its jth level, and the third factor at its kth level as y_{ijk} ($i = 1, \dots, h_1$; $j = 1, \dots, h_2$; $k = 1, \dots, h_3$). The three-way ANOVA model is

$$y_{ijk} = u^{(0)} + u_i^{(1)} + u_j^{(2)} + u_k^{(3)} + u_{i,j}^{(1,2)} + u_{i,k}^{(1,3)} + u_{j,k}^{(2,3)} + u_{i,j,k}^{(1,2,3)},$$

where $u^{(0)}$ is the overall mean; $u^{(1)}, u^{(2)}, u^{(3)}$ are the main effects; $u^{(1,2)}, u^{(1,3)}, u^{(2,3)}$ are the two-factor-interaction effects; and $u^{(1,2,3)}$ is the three-factor-interaction effect. Furthermore, identifiability requires constraints such as

$$\sum_i u_i^{(1)} = \sum_i u_{i,j}^{(1,2)} = \sum_i u_{i,k}^{(1,3)} = \sum_i u_{i,j,k}^{(1,2,3)} = 0$$

for the first factor along its index i. Similar equations describe the second and third factors along with their indexes j and k, respectively. Therefore, the ANOVA can be represented as a linear regression model. From the generalized linear model (GLM) point of view, a linear regression model where the response is normally distributed has a canonical link function of unity (see McCullagh and Nelder (1989)).

By analogy with the preceding standard ANOVA, it is possible to build a similar regression model relating the cell counts and their corresponding factor levels in the multiway contingency table. Note that, in our case, the response (the cell count) is not normally distributed. With no restriction on the total sample size, we have assumed that each cell count is independently Poisson distributed. A GLM where the response follows a Poisson distribution has a canonical link function that is a logarithm (see McCullagh and Nelder (1989)). Therefore, we assume a log-linear model. For a three-way contingency table of size $h_1 \times h_2 \times h_3$, the log-linear model characterizing the relationship between the expectation m_{ijk} ($i = 1, \ldots, h_1$; $j = 1, \ldots, h_2$; $k = 1, \ldots, h_3$) of the count in cell(i, j, k) and the factor levels indexed with i, j, k is

$$\ln m_{ijk} = u^{(0)} + u_i^{(1)} + u_j^{(2)} + u_k^{(3)} + u_{i,j}^{(1,2)} + u_{i,k}^{(1,3)} + u_{j,k}^{(2,3)} + u_{i,j,k}^{(1,2,3)},$$

where the u-terms are the main or factor-interaction effects defined as in the ANOVA model (see Bishop et al. (2007)). They also satisfy the identifiability constraints. When the total sample size N is fixed, the cell counts jointly follow a multinomial distribution, and it is more convenient to focus on the probability p_{ijk} instead of the expectation $m_{ijk} = Np_{ijk}$. In this case, the log-linear model will be

$$\ln p_{ijk} = u^{(0)} + u_i^{(1)} + u_j^{(2)} + u_k^{(3)} + u_{i,j}^{(1,2)} + u_{i,k}^{(1,3)} + u_{j,k}^{(2,3)} + u_{i,j,k}^{(1,2,3)}, \tag{1}$$

where the probabilities must satisfy $\sum_{i,j,k} p_{ijk} = 1$. Obviously, the interaction terms, such as $u^{(1,2)}$, reflect the dependence among the factors and, for a log-linear model without any interaction effects, the factors are independent.

The identifiability constraints applicable to a log-linear model in the form of Equation (1) are somewhat inconvenient to write out, but they can be rewritten equivalently in the following form, which we illustrate for a 2 × 3 contingency table having the factor C_1 with two levels and the factor C_2 with three levels:

$$\begin{aligned}
u^{(0)} &= \beta_0,\\
u_1^{(1)} &= \beta_1,\\
u_2^{(1)} &= -\beta_1,\\
u_1^{(2)} &= \beta_2,\\
u_2^{(2)} &= \beta_3,\\
u_3^{(2)} &= -\beta_2 - \beta_3,\\
u_{1,1}^{(1,2)} &= \beta_4,\\
u_{1,2}^{(1,2)} &= \beta_5,\\
u_{1,3}^{(1,2)} &= -\beta_4 - \beta_5,\\
u_{2,1}^{(1,2)} &= -\beta_4,\\
u_{2,2}^{(1,2)} &= -\beta_5,\\
u_{2,3}^{(1,2)} &= \beta_4 + \beta_5.
\end{aligned}$$

Therefore, the logarithms of the probabilities p_{ij} ($i = 1, 2$; $j = 1, 2, 3$) can be expressed as a linear combination of the coefficients β_k ($k = 0, 1, \ldots, 5$). Clearly, β_1 measures the main effect $u^{(1)}$ of the first factor, $[\beta_2, \beta_3]^{\mathrm{T}}$ measures the main effect $u^{(2)}$ of the second factor, and $[\beta_4, \beta_5]^{\mathrm{T}}$ measures the interaction effect $u^{(1,2)}$ of the two factors. The preceding representation of factor effects in terms of coefficients can be extended to a general scenario.

The identifiability constraints dictate that the log-linear model for a p-way contingency table in which p factors are considered in the form of Equation (1) can be expressed in the following regression form (see Dahinden et al. (2007)):

$$\ln \mathbf{p} = \mathbf{1}\beta_0 + \sum_{i=1}^{2^p-1} \mathbf{X}_i \boldsymbol{\beta}_i, \tag{2}$$

where $\mathbf{p}$ is the $h \times 1$ probability vector corresponding to the h cells of the contingency table, $\mathbf{1}$ is a column vector of appropriate dimension consisting of 1 as all its entries, $\mathbf{X}_i$ is an $h \times q_i$ design submatrix corresponding to the ith main or interaction effect and containing 1, 0, or -1 as its elements, and $\boldsymbol{\beta}_i$ is the coefficient subvector of size $q_i \times 1$. For the preceding illustrative 2 × 3 contingency table example with $p = 2$ factors, $\mathbf{p} = [p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}]^{\mathrm{T}}$ and $\mathbf{1} = \mathbf{1}_6$ are both of size 6×1, and $\boldsymbol{\beta}_1 = \beta_1$,

$\boldsymbol{\beta}_2 = [\beta_2, \beta_3]^{\mathrm{T}}$, $\boldsymbol{\beta}_3 = [\beta_4, \beta_5]^{\mathrm{T}}$, and the design submatrixes are

$$\mathbf{X}_1 = \begin{bmatrix} \mathbf{1}_3 \\ -\mathbf{1}_3 \end{bmatrix}, \qquad \mathbf{X}_2 = \begin{bmatrix} \mathbf{J} \\ \mathbf{J} \end{bmatrix},$$

and

$$\mathbf{X}_3 = \begin{bmatrix} \mathbf{J} \\ -\mathbf{J} \end{bmatrix},$$

where

$$\mathbf{J} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}.$$

Note that the column sums in $\mathbf{J}$ are all zero, which assures identifiability.

Denoting the design matrix by $\widetilde{\mathbf{X}} = [\mathbf{1}, \mathbf{X}]$ with $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{2^p-1}]$ and the coefficient vector by $\widetilde{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}^{\mathrm{T}}]^{\mathrm{T}}$ with $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{2^p-1}^{\mathrm{T}}]^{\mathrm{T}}$, Equation (2) can be rewritten as $\ln \mathbf{p} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}$. The log-linear model (2) is at the effect level, and there are, in total, $2^p - 1$ effects from the main effects up to the p-factor-interaction effect. Here, $\mathbf{p}^{\mathrm{T}}\mathbf{1} = 1$ and β_0 is a scalar representing the intercept. Therefore, only $h - 1$ coefficients are free to vary independently and β_0 is included to ensure the constraint $\mathbf{p}^{\mathrm{T}}\mathbf{1} = 1$. Hereafter, attention will focus on the coefficient subvectors $\boldsymbol{\beta}_i$ $(i = 1, \ldots, 2^p - 1)$. In the case of factors all with two levels, the derivation of the design matrix $\widetilde{\mathbf{X}}$ is identical to that of the design matrix of 2^k full-factorial experiment with 1 and -1 representing the high and low levels, respectively. However, this becomes a little complex if at least one factor has more than two levels. Refer to the additional File 1 in Dahinden et al. (2007) for the general result of deriving $\widetilde{\mathbf{X}}$. For convenience, we provide the Fortran code for deriving $\widetilde{\mathbf{X}}$ once the number of factors and their levels are specified, which is available from the authors on request. In addition, an example of four factors is given in Appendix B in the supplemental file.

The design submatrixes, together with their corresponding coefficient subvectors, are arranged in the log-linear model from the overall mean β_0, the main effects, up to the effect of the highest order. For example, we consider three factors C_1, C_2, and C_3 with 2, 3, and 3 levels, respectively. The sequence is the overall mean; the main effects C_1, C_2, and C_3; the two-factor interaction effects C_1C_2, C_1C_3, and C_2C_3; and finally the three-factor interaction effect $C_1C_2C_3$. Hence, the coefficient vector is

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} = [\beta_0 \quad & \beta_{(1)} & \beta_{(2_1)} \\ \beta_{(2_2)} \quad & \beta_{(3_1)} & \beta_{(3_2)} \\ \beta_{(1,2_1)} \quad & \beta_{(1,2_2)} & \beta_{(1,3_1)} \\ \beta_{(1,3_2)} \quad & \beta_{(2_1,3_1)} & \beta_{(2_1,3_2)} \\ \beta_{(2_2,3_1)} \quad & \beta_{(2_2,3_2)} & \beta_{(1,2_1,3_1)} \\ \beta_{(1,2_1,3_2)} \quad & \beta_{(1,2_2,3_1)} & \beta_{(1,2_2,3_2)}]^{\mathrm{T}}. \end{aligned}$$

Note that, by the identifiability constraints, the scalar $\beta_{(1)}$ is sufficient to determine the main effect of the factor C_1 with only two levels. However, we need the two coefficients $\beta_{(3_1)}$ and $\beta_{(3_2)}$ consisting of the subvector $\boldsymbol{\beta}_3$ to jointly represent the main effect of the factor C_3 with three levels. In fact, for an effect that contains a factor with three levels or more, its corresponding coefficient subvector has two elements or more, instead of reducing into a scalar. Similarly, $\boldsymbol{\beta}_4 = [\beta_{(1,2_1)}, \beta_{(1,2_2)}]^{\mathrm{T}}$ measures the two-factor interaction effect C_1C_2, and $\boldsymbol{\beta}_7 = [\beta_{(1,2_1,3_1)}, \beta_{(1,2_1,3_2)}, \beta_{(1,2_2,3_1)}, \beta_{(1,2_2,3_2)}]^{\mathrm{T}}$ measures the three-factor-interaction effect $C_1C_2C_3$. We see that the ith main or interaction effect, the design submatrix $\mathbf{X}_i$, and the coefficient subvector $\boldsymbol{\beta}_i$ $(i = 1, \ldots, 2^p - 1)$ correspond to each other. Therefore, the probability vector is determined by the magnitudes of these coefficient subvectors.

The log-linear model (2) is at the effect level, but it can be rewritten equivalently at the coefficient level as

$$\ln \mathbf{p} = \mathbf{1}\beta_0 + \sum_{i=1}^{h-1} \mathbf{x}_i \beta_i, \tag{3}$$

where $\mathbf{x}_i$ is the ith column vector of the matrix $\mathbf{X}$ and the scalar β_i is its corresponding coefficient. For instance, in the above three-way contingency table of size $2 \times 3 \times 3$, $\beta_1 = \beta_{(1)}$, $\beta_9 = \beta_{(1,3_2)}$, and $\beta_{17} = \beta_{(1,2_2,3_2)}$. We see $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_{2^p-1}] = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{h-1}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_{2^p-1}^{\mathrm{T}}]^{\mathrm{T}} = [\beta_1, \beta_2, \ldots, \beta_{h-1}]^{\mathrm{T}}$. There is also a correspondence between the ith column $\mathbf{x}_i$ of $\mathbf{X}$ and the ith coefficient β_i $(i = 1, \ldots, h - 1)$. By analogy with a linear regression model, the log-linear model is also essentially a regression model with the probabilities as the responses, the coefficients as the regressors, and the column vectors composing the design matrix.

A Phase II control chart requires that the IC process parameters be known, which requires estimation of the coefficient vector $\widetilde{\boldsymbol{\beta}}$ in the log-linear model (2) or (3) from an IC dataset. But as will be seen next, some simple approximations allow using only the probability vector $\mathbf{p}$ in the IC state, skipping the estimation of the coefficient vector $\widetilde{\boldsymbol{\beta}}$. Once the IC dataset is known, the IC probability vector $\mathbf{p}^{(0)}$ can be obtained immediately by dividing the cell counts by the IC dataset size.

Log-Linear Directional Monitoring

In a log-linear model, the marginal distribution of one factor is mainly determined by its main effect, whereas the dependence among multiple factors is represented by their interaction effect. This provides a means of interpreting shifts in multivariate categorical processes. According to the one-to-one correspondence between factor effects and coefficient subvectors in a log-linear model, shifts in the marginal distribution of one factor lead to deviations of the coefficient subvector corresponding to its main effect, and shifts in the dependence among multiple factors result in deviations of the coefficient subvector reflecting their interaction effect.

The prespecified log-linear model can be summarized as

$$\ln \mathbf{p} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} \quad \text{and} \quad \mathbf{p}^{\mathrm{T}}\mathbf{1} = 1.$$

Denote this model as $F(\widetilde{\mathbf{X}}; \widetilde{\boldsymbol{\beta}})$. We assume that the jth on-line multivariate sampling observation vector $\mathbf{n}_j$, of size $h \times 1$, follows a multinomial distribution with a total size N and behaves over time according to the change-point model

$$\mathbf{n}_j \overset{\text{i.i.d.}}{\sim} \begin{cases} F(\widetilde{\mathbf{X}}; \widetilde{\boldsymbol{\beta}}^{(0)}), & \text{for } j = 1, \dots, \tau, \\ F(\widetilde{\mathbf{X}}; \widetilde{\boldsymbol{\beta}}^{(1)}), & \text{for } j = \tau + 1, \dots, \end{cases} \tag{4}$$

where τ is the unknown change point,and $\widetilde{\boldsymbol{\beta}}^{(0)} \neq \widetilde{\boldsymbol{\beta}}^{(1)}$ are the known IC and unknown OC process coefficient vectors, respectively.

Because β_0 can be determined from

$$\boldsymbol{\beta} = [\beta_1, \dots, \beta_{h-1}]^{\mathrm{T}},$$

the monitoring problem can be formulated as a hypothesis-testing problem,

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}^{(0)}. \tag{5}$$

A natural test for the hypothesis (5) can be constructed by using the idea of a generalized likelihood-ratio test (GLRT; Anderson (2003)), which incorporates all possible shifts in $\boldsymbol{\beta}^{(0)}$ and thus is quite general.

There is a correspondence between the ith main or interaction effect and the coefficient subvector $\boldsymbol{\beta}_i$. According to the effect-sparsity principle (see Wu and Hamada (2000)), which states that the number of relatively important effects is small in DOE, it is usually reasonable to assume that, in practical applications, any changes involve only a few coefficient subvectors or only a few coefficients in the appropriate model. Suppose, however, that we have some *a priori* knowledge that, in the OC state, only one coefficient β_i $(1 \leq i \leq h-1)$ is incremented by an unknown constant δ_i. The hypothesis then becomes

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} \quad \text{versus} \quad H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + \mathbf{d}_i\delta_i,$$

where $\mathbf{d}_i$ is the direction vector of size $(h-1) \times 1$ with a 1 as its ith component and 0s elsewhere.

Next consider the more practical case that, in the OC state, only one coefficient changes, but its location is unknown. The alternative in the hypothesis (5) reduces to

$$\begin{aligned} H_1 : \quad & \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + \mathbf{d}_1\delta_1 \quad \text{or} \\ & \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + \mathbf{d}_2\delta_2 \dots \quad \text{or} \\ & \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + \mathbf{d}_{h-1}\delta_{h-1}, \end{aligned} \tag{6}$$

where δ_i $(i = 1, 2, \dots, h-1)$ are the unknown shift magnitudes, and the possible shift directions $\mathbf{d}_1$, $\mathbf{d}_2$, $\dots$, $\mathbf{d}_{h-1}$ are defined in a similar way, applying to β_1, β_2, $\dots$, β_{h-1}, respectively. The GLRT derived from the hypothesis (6) should be better than that from the hypothesis (5) because it makes full use of more constructive information about potential shift directions. In fact, a GLRT based on the hypothesis (6) that incorporates directional knowledge about potential changes is very much like the GLRTs used in multistage process monitoring and diagnosis, which exploit the information of shift directions from the first stage to the last one (see Zou and Tsung (2008), Zou et al. (2008)).

The hypothesis (6) may be further generalized. There is also the hierarchical-ordering principle in DOE (see Wu and Hamada (2000)), which states that lower order effects are more likely to be important than higher order effects and that effects of the same order are equally likely to be important. So deviations involving fewer factors may appear more frequently, and it is reasonable to believe that, in applications, most shifts involve lower order effects rather than higher order ones. Unlike the hypothesis (6), which considers all one-coefficient shifts from the main-factor effects up to the highest p-factor-interaction effect, instead we may focus on effects involving the first few, say q, orders. Denote by g the number of coefficients corresponding to effects of the first q orders. Then the hypothesis (6) can be further extended as

$$H_1 : \quad \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} + \mathbf{d}_1\delta_1 \quad \text{or}$$