Statistical Monitoring of Longitudinal Categorical Survey Data

Chen Zhang, Nan Chen
Department of Industrial and Systems Engineering, National University of Singapore, Singapore
(zhangchen@u.nus.edu & isecn@nus.edu.sg)

Abstract - **The longitudinal survey is conducted to collect responses from the target population repeatedly over long periods of time, aiming to analyze and monitor the response development as time goes on. However so far systematic detection methodology for survey response changes is still in its infancy. In this regard, this paper sheds light on this field by applying statistical process control (SPC) methodology into monitoring of longitudinal survey responses. Specifically, since generally the longitudinal survey responses are categorical variables which have temporal dependence on their previous values, i.e., autocorrelation, this research firstly proposes a hierarchical model to describe these categorical time series. The model can provide flexible autocorrelation structures for the categorical time series and therefore can be widely applied. Based on this model, this research secondly designs a SPC scheme using likelihood ratio test to monitor the survey responses. Estimation of the in-control (IC) response distribution is also discussed. Numerical studies demonstrate the satisfactory monitoring performance of the proposed scheme. Finally as an empirical evaluation, the scheme is applied to a real survey dataset to detect changes of consumer attitudes towards the economic conditions during a economic crisis.**

Keywords - **Statistical process control (SPC), longitudinal survey data analysis, categorical time series, hierarchical state space model, likelihood ratio test, particle filtering, Monte Carlo expectation maximization (MCEM)**

I. Introduction

Survey research has been widely used in social sciences, marketing and politics as a standard tool to gather information from the target population and to learn the opinions of the population on certain interesting topics [1]. For examples, customer satisfaction surveys help the company to evaluate its performance and to orientate management and strategies. Political polls help the government to know the residence opinions on a forthcoming policy or law. One crucial aspect of the success of a survey is statistical analysis of the collected data. So far plenty of literature has been focusing on how to design and analyze surveys to guarantee accurate statistical properties. See [1] and the references therein for more background knowledge. One trend is that, as data collecting technologies become easier and cheaper, now we can conduct a series of surveys from the same population over long periods of time, sometimes many years, which are called longitudinal surveys (studies) [2]. They aim to measure the development of the responses (characteristics) of the target population. Current analyses of longitudinal surveys focus on building models to describe the responses and forecasting their future values [4]. However, another equally important yet unnoticed point is statistical monitoring of these responses and efficient detection of their changes. A fast detection will provide timely feedback to the survey sponsor for policy adjustment. However, the only pioneering work for monitoring longitudinal surveys is [5], which presents a regression model to analyze the impact of service changes on customer attitudes by comparing the regression coefficients before and after the changes. To our best knowledge, so far systematic statistical monitoring schemes are still under-developed. One possible direction is to apply statistical process control (SPC) methodology to monitor longitudinal survey data. This is motivated by that the survey data consist of samples sequentially collected from the population, such as those taken as part of a quality control process.

As we know, one most typical question type of a survey is the tick-box question where respondents are asked to select one (or potentially more) from a fixed number of possible options. In statistics we call this kind of responses as categorical data and assume them following the multinomial distribution. So far many SPC schemes have been proposed for categorical data with applications in manufacturing. [6] presented a generalized p chart to monitor multinomial processes based on Pearson χ^2 statistics. [7] designed a multinomial cumulative sum (CUSUM) chart based on likelihood ratio test. [8] represented a k-category process by a probability tree with $k-1$ binary stages and monitored them with $k-1$ independent p charts.

However one common limitation of the aforementioned charts is to treat the multinomial samples as independent ones. While for longitudinal survey data, the responses usually have temporal correlations with the previous values, i.e., autocorrelation. As a result, efficient SPC schemes should take this point into consideration as well. Unfortunately, to our best knowledge, current SPC schemes for categorical time series can only handle binomial or binary data series. [9] used the correlated binomial model to monitor a product process with almost-zero nonconformity probability. [10] proposed a stationary Markov chain model to monitor binary series. [11] developed a binomial integer-valued autoregressive (INAR) model and proposed several control schemes based on the model. However their extensions to multinomial or categorical cases are yet to be addressed.

To fill this research gap, here we propose an easy-to-interpret SPC scheme for longitudinal categorical survey data. It is also implementable for general categorical time series with modest modifications. Our contributions are threefold. Firstly we propose a hierarchical model to describe the multinomial time series whose distribution parameters are assumed to be driven by some latent vari-

978-1-4673-8066-9/15/$31.00 ©2015 IEEE

ables evolving according to a state space model. It is this latent process that introduces temporal correlation of the category data. Based on this model, we secondly propose a control scheme for categorical time series using likelihood ratio test. EWMA technique is also integrated into the scheme to improve its detection power for small distributional changes. Thirdly, with regard to implementable issues, we put stress on model parameter estimation of the in-control (IC) process. Monte Carlo expectation maximization (MCEM) algorithm together with particle filtering & smoothing technique is used here, which provides satisfactory estimation results. Numerical studies show the chart can detect general distributional changes of the categorical time series efficiently. An empirical evaluation from a real longitudinal survey dataset, which records the consumer attitudes towards the economic conditions in the United States for consecutive years, demonstrates this point as well.

II. Methodology

A. A Motivating Example

Our motivating application concerns a real survey dataset which measures changes in American consumer attitudes towards the economic conditions. This survey was carried out since 1974 and monthly thereafter by the Survey Research Centre of University of Michigan. The respondent samples are selected from dwelling units randomly by area and therefore representative of the adult population of the United States. The responses come from three alternatives: "better situation now", "about the same situation now", or "worse situation now" compared with that of the previous year. Therefore it is a 3-dimensional multinomial process. Raw plots of the empirical proportions of these 3 categories (e.g., Fig 3) reveal their similar or opposite change patterns. These patterns can be justified by the existence of factors, not precisely identified, that relate to the economic conditions and affect different series in similar or opposite ways. For example, the decrease of deposit interest rates seems positive for some people but negative for some others. Hence it increases the number of responses in the categories of "better now" and "worse now" in similar ways with different magnitudes, and reduces the number of responses in the category of "about the same". This kind of considerations motivates our choice of jointly modelling the categorical series in a way that takes into account the existence of some latent variables that evolve through time, as introduced in detail in the following.

B. A State Space Model For Categorical Time Series

For k-categorical survey data collected sequentially at time $t = 1, \cdots, T$, suppose at t we collect n_t responses, every one of which comes from one of these k options. We denote $\mathbf{Y}_t = [Y_{t1}, \ldots, Y_{tk}]$ as a k-dimensional vector with $Y_{ti}, i = 1, \cdots, k$ as the number of responses in the category i. Let p_{ti} define the probability that one response falls into the category i at t, with the constraint $\sum_{i=1}^{k} p_{ti} = 1$. Then conditioning on $\mathbf{p}_t = [p_{t1}, \cdots, p_{tk}]$, $\mathbf{Y}_t$ follows a multinomial distribution, i.e.,

$$\mathbf{Y}_t | \mathbf{p}_t \sim \mathrm{Multinom}(n_t, \mathbf{p}_t), \text{for } t = 1, \cdots, T.$$

Now we introduce an elementwise transformation of $\mathbf{p}_t$ to a vector $\mathbf{X}_t$ whose elements are real values. This transformation aims to define the distribution of $\mathbf{p}_t$ indirectly and to specify multivariate normal models for $\mathbf{X}_t$, since a multivariate normal distribution allows for more flexible model structures. Here we consider the logit transformation $\mathbf{X}_t = \mathrm{logit}(\mathbf{p}_t)$, where $X_{ti} = \log(\frac{p_{ti}}{p_{tk}})$ for $i = 1, \cdots, k-1$. Then we have

$$p_{ti} = \begin{cases} \frac{\exp(X_{ti})}{1+\sum_{j=1}^{k-1}\exp(X_{tj})} & \text{for } i = 1, \cdots, k-1 \\ \frac{1}{1+\sum_{j=1}^{k-1}\exp(X_{tj})} & \text{for } i = k \end{cases}. \quad (1)$$

Other similar transformations such like the arcsine transformation are also possible. Then we treat $\mathbf{X}_t = \left[X_{t1}, \cdots, X_{t(k-1)}\right]$ as a latent variable and introduce process autocorrelation by assuming that $\mathbf{X}_t$ evolves according to a state space model, i.e.,

$$\mathbf{X}_t - \boldsymbol{\mu} = \boldsymbol{\Phi} \cdot (\mathbf{X}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\epsilon}_t, \quad (2)$$

where $\boldsymbol{\epsilon}_t$ is the white noise following a $(k-1)$-dimensional multivariate normal distribution $\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$. So far we have introduced all the model parameters $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}\}$.

As long as $\boldsymbol{\Phi}$ satisfies

$$\det(\mathbf{I} - z\boldsymbol{\Phi}) \neq 0, \quad \text{for all } z \in \mathcal{C} \text{ such that } |z| \leq 1,$$

we can guarantee that $\mathbf{X}_t$ together with $\mathbf{p}_t$ is stationary. Then $p_\Theta(\mathbf{X}_t)$ is a $k-1$-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is the solution of $\boldsymbol{\Gamma} = \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}' + \boldsymbol{\Sigma}$ according to the Yule-Walker relationship. Marginalizing $\mathbf{X}_t$, the unconditional distribution of $\mathbf{Y}_t$ can be expressed as

$$p_\Theta(\mathbf{Y}_t) = \int_{\mathbf{R}_+^{k-1}} \frac{n_t!}{Y_{tk}!\left(1 + \sum_{j=1}^{k-1}\exp(X_{tj})\right)^{n_t}} \quad (3)$$
$$\prod_{i=1}^{k-1} \frac{\exp(X_{ti}Y_{ti})}{Y_{ti}!} p_\Theta(\mathbf{X}_t) \mathrm{d}\mathbf{X}_t.$$

This hierarchical model allows for flexible patterns of the autocorrelation between different categories. The model ties the categorical series together but allows for individual stochastic components through the term $\boldsymbol{\epsilon}_t$. Higher order autocorrelation of $\mathbf{X}_t$ can also be accommodated in the model, which is not the focus of this paper.

III. A Monitoring Scheme

Based on the proposed state space model, we develop a SPC scheme for categorical time series. We focus on Phase II monitoring assuming that the in-control (IC) parameters $\Theta_0 = \{\boldsymbol{\mu}_0, \boldsymbol{\Phi}_0, \boldsymbol{\Sigma}_0\}$ are known exactly or estimated accurately from the historical data. The estimation procedure will be introduced in detail in Section IV. We further assume that the target or the out-of-control (OC) parameters Θ_1 are unknown, and define a change-point model as

$$\mathbf{Y}_t \overset{\text{i.i.d.}}{\sim} \begin{cases} p_{\Theta_0}(\mathbf{Y}_t) & \text{for} \quad t = 1, \ldots, \tau \\ p_{\Theta_1}(\mathbf{Y}_t), & \text{for} \quad t = \tau + 1, \ldots \end{cases},$$

where τ is the unknown change point we want to detect.

Suppose that the chart has not triggered an OC alarm before t, then we can estimate $\tilde{\mathbf{X}}_{t-1}$ from $p_{\Theta_0}(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})$ as

$$\tilde{\mathbf{X}}_{t-1} = \mathrm{E}_{\Theta_0}\left[\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1}\right], \tag{4}$$

and forecast $\hat{\mathbf{X}}_t$ via

$$\hat{\mathbf{X}}_t - \boldsymbol{\mu}_0 = \boldsymbol{\Phi}_0 \cdot (\tilde{\mathbf{X}}_{t-1} - \boldsymbol{\mu}_0).$$

In this way we can also forecast $\hat{\mathbf{p}}_t$ via the link function in Equation (1). According to $\hat{\mathbf{p}}_t$, we construct the likelihood ratio-based goodness-of-fit test at time t as

$$G_t = \sum_{i=1}^{k} Y_{ti} \ln \frac{Y_{ti}}{n_t \hat{p}_{ti}}. \tag{5}$$

To improve the detection power for small shifts, we integrate the EWMA technique into Equation (5) and get the final charting statistic as

$$Z_t = (1 - \lambda) Z_{t-1} + \lambda G_t, \tag{6}$$

where λ is the EWMA tuning parameter with usual settings as $0.05 \leq \lambda \leq 0.2$. We set the initial value $Z_0 = 0$. We define the chart triggers an OC alarm if $Z_t > h$. h is the critical value determined based on the pre-specific IC average run length (ARL$_0$).

One concern about the chart is, since the posterior distribution $p_{\Theta_0}(\mathbf{X}_t|\mathbf{Y}_{1:t})$ has no close form for arbitrary Θ_0, at first glance the inference of Equation (4) is intractable. However, we can tackle it perfectly by numerical integration method. Considering the sequential nature of the state space model, here we use particle filter algorithm (PF, also called Sequential Monte Carlo) [12]. PF approximates $p_{\Theta_0}(\mathbf{X}_t|\mathbf{Y}_{1:t})$ in a sequential way with the concept of *importance sampling*. It assumes the density $p_{\Theta_0}(\mathbf{X}_t|\mathbf{Y}_{1:t})$ can be approximated by N_p samples $\{\mathbf{x}_t^i, i = 1, \cdots, N_p\}$ and their associated weights $\{W_t^i, i = 1, \cdots, N_p\}$. Then the weighted approximation to the posterior density is given by

$$p_{\Theta_0}(\mathbf{X}_t|\mathbf{Y}_{1:t}) \approx \sum_{i=1}^{N_p} W_t^i \delta(\mathbf{X}_t - \mathbf{x}_t^i), \tag{7}$$

where δ is the Dirac delta function. The normalized weights W_t^i's satisfy $\sum_{i=1}^{N_p} W_t^i = 1$. Then the distribution of $\mathbf{X}_{t+1}$ can be predicted by Equation (7). Specifically, each particle $\mathbf{x}_{t+1}^i$ for $i = 1, \cdots, N_p$ can be generated by propagating the state function in Equation (2) with a random noise $\boldsymbol{\epsilon}_{t+1}^i$ drawn from the white noise distribution, and then used for prediction as

$$p_{\Theta_0}(\mathbf{X}_{t+1}|\mathbf{Y}_{1:t}) \approx \sum_{i=1}^{N_p} W_t^i \delta(\mathbf{X}_{t+1} - \mathbf{x}_{t+1}^i).$$

Then the next observation $\mathbf{Y}_{t+1}$ is used to update the weight for every particle by its corresponding likelihood according to the Bayes rule, i.e.,

$$W_{t+1}^i \propto W_t^i \cdot p(\mathbf{Y}_{t+1}|\mathbf{x}_{t+1}^i).$$

Therefore we have

$$p_{\Theta_0}(\mathbf{X}_{t+1}|\mathbf{Y}_{1:t+1}) \approx \sum_{i=1}^{N_p} W_{t+1}^i \delta(\mathbf{X}_{t+1} - \mathbf{x}_{t+1}^i).$$

Fig. 1 presents the estimation performance of PF with $N_p = 500$ for a 3-dimensional categorical time series as an example. The close overlap of the true and estimated values demonstrates the efficiency of the algorithm.

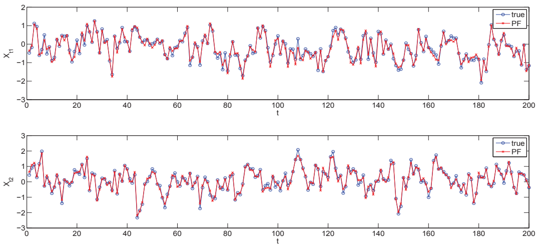


Fig. 1: The estimate $\tilde{\mathbf{X}}_t = \mathrm{E}_{\Theta_0}\left[\mathbf{X}_t|\mathbf{Y}_{1:t}\right]$ based on PF for a 3-dimensional categorical time series with parameter $n_t = 400, \boldsymbol{\Phi} = [0.4, 0.1; -0.2, 0.5]$, $\boldsymbol{\mu} = [-0.3, 0.3]$, $\boldsymbol{\Sigma} = 0.36\mathbf{I}$, and $N_p = 500$.

IV. Parameter Estimation

Usually the IC parameters Θ_0 need to be estimated from the historical IC data, noted as $\{\mathbf{Y}_{1:T}\}$ where T is the IC sample size. As such, now we discuss the general estimation procedure of Θ for the proposed state space model. The challenge is since its latent variable structure makes the marginal unconditional distribution $p_\Theta(\mathbf{Y}_t)$ in Equation (3) have no close form, direct model estimation based on maximum likelihood is impossible here. One natural and efficient solution to deal with these latent variables is expectation maximization (EM) algorithm. EM algorithm is an iterative procedure to seek for Θ^{k+1} at the $(k+1)^{th}$ step such that the likelihood of $\{\mathbf{Y}_{1:T}\}$ is increased from that at the k^{th} step. Its key idea is to postulate the "missing" data set $\{\mathbf{X}_{1:T}\}$ and consider maximizing the joint log-likelihood of the complete data $\{\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}\}$. Thanks to the Markovian structure of the state space model, the complete data log-likelihood has the separable form as

$$\log p_\Theta(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) = \log p_0(\mathbf{X}_1) + \sum_{t=1}^{T-1} \log p_\Theta(\mathbf{X}_{t+1}|\mathbf{X}_t) + \sum_{t=1}^{T} \log p(\mathbf{Y}_t|\mathbf{X}_t).$$

where $p_0(\mathbf{X}_1)$ is the prior distribution of $\mathbf{X}_1$. Since $\{\mathbf{X}_{1:T}\}$ are unavailable, EM approximates the complete log-likelihood by $\mathcal{Q}(\Theta, \Theta^k)$, which is the conditional expectation of $\log p_\Theta(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T})$ given the observations

$\{\mathbf{Y}_{1:T}\}$ with the current parameters Θ^k, i.e.,

$$\text{E step:} \quad \mathcal{Q}(\Theta, \Theta^k) = \\ \int p_{\Theta^k}(\mathbf{X}_{1:T}|\mathbf{Y}_{1:T}) \cdot \log p_{\Theta}(\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}) \mathrm{d}\mathbf{X}_{1:T}. \tag{8}$$

Then we want to find the revised parameter estimate Θ^{k+1} that maximizes the function

$$\text{M step:} \quad \Theta^{k+1} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^k).$$

Unfortunately, since in our model $p_{\Theta^k}(\mathbf{X}_{1:T}|\mathbf{Y}_{1:T})$ also has no close form, direct EM estimation is intractable. Here we propose to use Monte Carlo EM (MCEM) with particle filtering & smoothing algorithm to approximate Equation (8). Specifically, recalling the filtering distribution in Equation (7), we can also get a direct approximation to the smoothed distribution $p_{\Theta^k}(\mathbf{X}_t|\mathbf{Y}_{1:T})$ for $t = 1, \cdots, T-1$ with the same particles yet different weights as filtering by considering the information of the future observations $\{\mathbf{Y}_{t+1:T}\}$ into the current state estimation [12]. Based on the smoothed distribution, we can further get an approximation of Equation (8) as

$$\hat{\mathcal{Q}}(\Theta, \Theta^k) \approx \sum_{t=1}^{T-1} \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} W_{t,t+1|T}^{ij} \log p_{\Theta}(\mathbf{x}_{t+1}^j | \mathbf{x}_t^i), \tag{9}$$

where the pairwise particle weight is given by

$$W_{t,t+1|T}^{ij} = W_t^i \frac{W_{t+1|T}^j p_{\Theta^k}(\mathbf{x}_{t+1}^j|\mathbf{x}_t^i)}{\sum_{l=1}^{N_p} W_t^l p_{\Theta^k}(\mathbf{x}_{t+1}^j|\mathbf{x}_t^l)}, \tag{10}$$

$$\text{and} \quad W_{t|T}^i = W_t^i \sum_{j=1}^{N_p} \frac{W_{t+1|T}^j p_{\Theta^k}(\mathbf{x}_{t+1}^j|\mathbf{x}_t^i)}{\sum_{l=1}^{N_p} W_t^l p_{\Theta^k}(\mathbf{x}_{t+1}^j|\mathbf{x}_t^l)}.$$

Then in the M step, with the gradient available for Equation (9), we get $\Theta^{k+1} = \{\boldsymbol{\mu}^{k+1}, \boldsymbol{\Phi}^{k+1}, \boldsymbol{\Sigma}^{k+1}\}$ as

$$\boldsymbol{\Pi}^{k+1} = [(\mathbf{I} - \boldsymbol{\Phi}^{k+1})\boldsymbol{\mu}^{k+1}, \boldsymbol{\Phi}^{k+1}]' \\ = (\sum_{t=1}^{T-1}\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} W_{t,t+1|T}^{ij} \mathbf{x}_{t+1}^j \mathbf{z}_t^{i'})(\sum_{t=1}^{T-1}\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} W_{t,t+1|T}^{ij} \mathbf{z}_t^i \mathbf{z}_t^{i'})^{-1}$$

and

$$\boldsymbol{\Sigma}^{k+1} = \frac{1}{T-1} \sum_{t=1}^{T-1}\sum_{i=1}^{N_p}\sum_{j=1}^{N_p} W_{t,t+1|T}^{ij} (\mathbf{x}_{t+1}^j - \boldsymbol{\Pi}^{k+1}\mathbf{z}_t^i)(\mathbf{x}_{t+1}^j - \boldsymbol{\Pi}^{k+1}\mathbf{z}_t^i)'$$

with $\mathbf{z}_t^i = \left[1, \mathbf{x}_t^i\right]$.

The particle filter & smoothing-based EM algorithm inherits the good convergence property of the conventional EM algorithm when N_p is large enough. The following simulation results illustrate this point by setting $N_p = 200$ to estimate a 3-dimensional categorical time series with $n_t = 400$ and $T = 500$. The process parameters are set to be $\boldsymbol{\Phi} = [0.4, 0.1; -0.2, 0.5]$, $\boldsymbol{\mu} = [-0.3, 0.3]$ and $\boldsymbol{\Sigma} = 0.36\mathbf{I}$, from which 200 different replications of $\{\mathbf{X}_{1:T}, \mathbf{Y}_{1:T}\}$ are simulated. For every replication, we assume its $\{\mathbf{X}_{1:T}\}$ and Θ are yet unknown and to be estimated by the MCEM algorithm. The MCEM algorithm randomly picks an initial value Θ^0 and iterates the E step and M step until the terminating condition $\hat{\mathcal{Q}}(\Theta^{k+1}, \Theta^k) - \hat{\mathcal{Q}}(\Theta^k, \Theta^k) \leq \epsilon$ is satisfied for an extremely small $\epsilon \geq 0$. Then we totally have 200 estimates of Θ and use them to analyze the bias and the rooted mean square error (RMSE) of the estimation method. As Table I lists, the bias and RMSE are acceptably small, illustrating the satisfactory estimation accuracy and stability of MCEM.

TABLE I: Parameter estimation by MCEM Algorithm based on 200 replications.

Para.	True	Bias(RMSE)	Para.	True	Bias(RMSE)
ϕ_1	0.4	-0.012 (0.044)	μ_2	0.3	-0.001 (0.056)
ϕ_2	0.1	-0.002 (0.044)	σ_{11}	0.36	0.006 (0.025)
ϕ_3	-0.2	-0.006 (0.040)	σ_{12}	0	0.002 (0.017)
ϕ_4	0.5	-0.010 (0.036)	σ_{22}	0.36	0.006 (0.025)
μ_1	-0.3	-0.003 (0.044)			

V. Numerical Studies

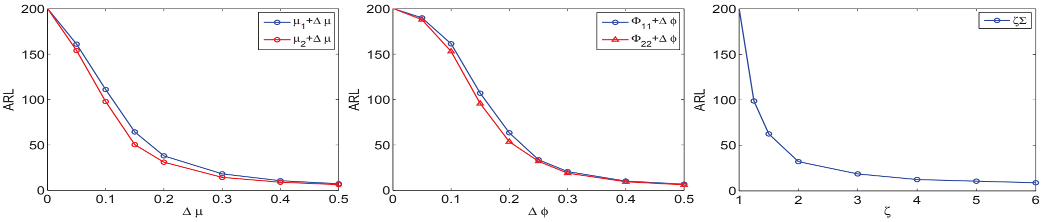
Here we use some numerical studies to present the charting performance of Equation (6). As we know, any shift in one of the three parameters, $\boldsymbol{\mu}, \boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$, may change the distribution of $\mathbf{Y}_t$ and trigger OC signals. Hence we discuss these three types of parameter changes respectively and study the chart performance in terms of average run length (ARL). Due to the limited space of this paper, we illustrates only one IC scenario here for a 3-dimensional process with $\boldsymbol{\mu}_0 = [-0.3, 0.3], \boldsymbol{\Phi}_0 = [0.4, 0.1; -0.2, 0.4], \boldsymbol{\Sigma}_0 = 0.01\mathbf{I}$ as an example. Some other simulation studies not shown here have been done as well and illustrated that the chart is quite robust for processes with different dimensions or IC parameters.

The OC scenarios considered here are as follows: i) shift in $\boldsymbol{\mu}$ with size $\Delta\mu$ for either dimension, i.e., $\mu_{1,i} = \mu_{0,i} + \Delta\mu$ for $i = 1, 2$; ii) shift in $\boldsymbol{\Phi}$ with size $\Delta\phi$, i.e., $\Phi_{1,ii} = \Phi_{0,ii} + \Delta\phi$ for $i = 1, 2$; iii) shift in $\boldsymbol{\Sigma}$ with magnitude of ζ, i.e., $\boldsymbol{\Sigma}_1 = \zeta\boldsymbol{\Sigma}_0$. We consider the target $\text{ARL}_0 = 200$ and calculate the responding control limit $h = 3.0906$ via simulation. We set the EWMA parameter $\lambda = 0.1$.

Fig. 2 shows that the chart has satisfactory OC performance against the change in magnitudes for different structure parameters. Particularly, it has a slight better detection power for changes in the second dimension. This might be caused by two factors. On one hand, due to the asymmetry of $\boldsymbol{\mu}_0$, a bigger $\mu_{0,2}$ in the second dimension means a higher proportion and more observations than the first dimension, and thereby magnifies the influence of the changes. On the other, due to the asymmetry of $\boldsymbol{\Phi}_0$, the same $\Delta\phi$ in $\Phi_{0,22}$ causes a bigger change of $\boldsymbol{\Gamma}$ for $p_{\Theta}(\mathbf{X}_t)$, and correspondingly a bigger distributional change for $\mathbf{p}_t$.

VI. A Real Application in Survey of Consumer Attitudes Towards Economic Conditions

Now we continue the real survey data analysis as introduced in Section II-A. Actually, we find the survey responses change periodically due to the influence of economic crises. In this sense, we can use the response data between two consecutive economic crises as IC



(a) $\boldsymbol{\mu}$ shift

(b) $\boldsymbol{\Phi}$ shift

(c) $\boldsymbol{\Sigma}$ shift

Fig. 2: OC performance for different change patterns: (a) $\boldsymbol{\mu}$ shift; (b) $\boldsymbol{\Phi}$ shfit; (c) $\boldsymbol{\Sigma}$ shift.

samples for model estimation, and then implement the control chart to detect the OC response changes during the second economic crisis. Here we use the data from 2003 July, when the economic conditions were just recovered from the crisis in 2000, to 2007 August with series length $T = 50$ as IC samples to construct the chart, and then use the chart to detect the response changes in the coming crisis in 2008.

The model parameter estimates based on the IC samples are given by

$$\boldsymbol{\Phi}_0 = \begin{pmatrix} 0.85 & -0.16 \\ 0.23 & 0.27 \end{pmatrix}, \boldsymbol{\mu}_0 = \begin{pmatrix} -0.14 \\ -1.38 \end{pmatrix}, \boldsymbol{\Sigma}_0 = \begin{pmatrix} 0.051 & 0.014 \\ 0.014 & 0.051 \end{pmatrix}.$$

Based on them, the correspondingly tracked $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{p}}_t$ by particle filtering are shown in Fig. 3. We can see that for the IC samples, $\hat{\mathbf{p}}_t$ match the empirical proportions well, illustrating that the fitted model can provide a good description of the survey data. Further model diagnoses which are not shown here also demonstrate this point. However after 2007 August, the fitted model loses its tracking accuracy, meaning that the process undergoes certain parameter changes. We apply the proposed SPC scheme to detect the changes. The control limit $h = 7.7596$ is calculated via simulation with the pre-specific IC $\text{ARL}_0 = 200$ and $\lambda = 0.1$. Fig. 4 plots the monitoring statistic Z_t through time. Noted that though we focus on Phase II monitoring, i.e., only monitoring samples after $t = T$, we still draw Z_t for $t = 2, \cdots, T$ for mere illustration. We can see that $\{Z_{2:T}\}$ are relatively small and stably below the control limit. After T, Z_t increases monotonously and reaches h at $t = 55$ (2008 January), triggering the OC alarm.

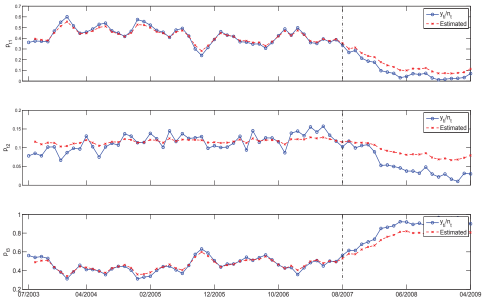


Fig. 3: The estimated $\mathbf{p}_t$ from samples (blue circle) and the state space model (red cross).

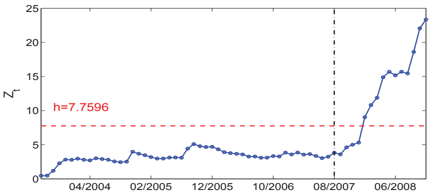


Fig. 4: The monitoring statistic Z_t for longitudinal categorical survey data, with the control limit $h = 7.7596$.

VII. Conclusion

This paper first proposes a state space model to describe the categorical time series with flexibility of the autocorrelation structure between different categories. Then based on the proposed model, it designs a SPC scheme for Phase II monitoring by likelihood ratio test. Numerical studies report the satisfactory detection power of the proposed chart. An empirical evaluation from a real survey dataset also demonstrates this point.

References

[1] R. L. Chambers and C. J. Skinner, *Analysis of survey data*. John Wiley & Sons, 2003.

[2] S. L. Zeger and K.-Y. Liang, “Longitudinal data analysis for discrete and continuous outcomes,” *Biometrics*, pp. 121–130, 1986.

[3] A. Pettitt, T. T. Tran, M. Haynes, and J. Hay, “A bayesian hierarchical model for categorical longitudinal data from a social survey of immigrants,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 169, no. 1, pp. 97–114, 2006.

[4] R. N. Bolton and J. H. Drew, “A longitudinal analysis of the impact of service changes on customer attitudes,” *The Journal of Marketing*, pp. 1–9, 1991.

[5] M. Marcucci, “Monitoring multinomial processes,” *Journal of Quality Technology*, vol. 17, no. 2, pp. 86–91, 1985.

[6] A. G. Ryan, L. J. Wells, and W. H. Woodall, “Methods for monitoring multiple proportions when inspecting continuously,” *Journal of quality technology*, vol. 43, no. 3, pp. 237–248, 2011.

[7] R. I. Duran and S. L. Albin, “Monitoring and accurately interpreting service processes with transactions that are classified in multiple categories,” *IIE Transactions*, vol. 42, no. 2, pp. 136–145, 2009.

[8] C. Lai, K. Govindaraju, and M. Xie, “Effects of correlation on fraction non-conforming statistical process control procedures,” *Journal of Applied Statistics*, vol. 25, no. 4, pp. 535–543, 1998.

[9] D. K. Shepherd, C. W. Champ, S. E. Rigdon, and H. T. Fuller, “Attribute charts for monitoring a dependent process,” *Quality and Reliability Engineering International*, vol. 23, no. 3, pp. 341–365, 2007.

[10] C. H. Weiß, “Monitoring correlated processes with binomial marginals,” *Journal of Applied Statistics*, vol. 36, no. 4, pp. 399–414, 2009.

[11] A. Doucet, S. Godsill, and C. Andrieu, “On sequential monte carlo sampling methods for bayesian filtering,” *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.