Statistical Process Control Can Be a Powerful Tool for Certain Big Data Applications

Peihua Qiu

Department of Biostatistics
University of Florida
E-mail: pqiu@ufl.edu

October 14, 2020
ICW-HDDA-X
Monitoring **production lines** in manufacturing industries

Distinguish **special cause variation** from **common cause variation**

*In-control (IC) versus out-of-control (OC)*

**Basic SPC charts**
- Shewhart chart (Shewhart 1931)
- CUSUM chart (Page 1954)
- EWMA chart (Roberts 1959)
- CPD chart (Hawkins, Qiu and Kang 2003)
Conventional CUSUM chart

- Observed data: $X_1, X_2, \ldots$
- IC process distribution: $N(\mu_0, \sigma^2)$
- CUSUM Chart (for detecting upward shifts):

$$C_n^+ = \max(0, C_{n-1}^+ + (X_n - \mu_0) / \sigma - k)$$

It gives a signal when $C_n^+ > h$. 

![CUSUM Chart with data points and a trend line showing upward shift detection.](chart.png)
- Performance measures: IC average run length \( ARL_0 \) versus OC average run length \( ARL_1 \)
- Assumptions: process observations are independent and identically distributed with a parametric (e.g., normal) distribution.
- Conventional charts are unreliable or even misleading when their model assumptions are violated.

Figure 1: Observed influenza-like illness incidence rates in Florida on 06/01/2012 (plot (a)) and 12/01/2012 (plot (b)).

- There are many global, national and regional reporting systems for collecting disease incidence data.
- **Disease surveillance:** detecting disease outbreaks
Monitoring Earth’s surface and Earth’s resource: land use, forest science, climate science, agriculture forecasting, ecological and ecosystem monitoring, water resources, and other natural science disciplines.
Recent SPC Research: Nonparametric Charts

Major Goal: Remove the **Normality** assumption

- **Multivariate SPC**
  - Nonparametric charts by data categorization and log-linear modeling (Qiu 2008)
  - Spatial ranks (Zou et al. 2012)
- **Univariate SPC** (Chakrabort et al. 2001, Qiu and Li 2011)
- An overview (Qiu 2018, *JQT*)
Dynamic Screening System (DySS)

Major Goal: Remove the **Identical Distribution** assumption

Motivating Example

- **SHARE** Framingham Heart Study.
- Many residents at Framingham MA were involved.
- Major risk factors of cardiovascular diseases: blood pressure, total cholesterol level (TCL), smoking, obesity, ...
- Identify patients with irregular longitudinal patterns of the disease risk factors as early as possible.
- Disease early detection and prevention
- Dynamic screening (DS) problem
DS problem is popular

Most products (e.g., airplanes, cars) are checked regularly or occasionally about certain variables related to their quality and/or performance.

If the observed values of a product are significantly worse than the values of a typical well-functioning product of the same age, then some adjustments or interventions should be made to avoid unpleasant consequences.
DySS (con’d)

Three-step approach:

- Estimate regular longitudinal pattern from an IC dataset
- Standardize observations of a new subject to monitor
- Monitor the standardized observations by a control chart
References:

- Qiu, Zi and Zou (2018, *Technometrics*): improved version
- You and Qiu (2019, *JSCS*): a faster computing algorithm
- You and Qiu (2019, *Technometrics*): risk quantification
- Qiu, Xia and You (2020, *Technometrics*): new performance metrics
Major Goal: Remove the **Independent Data** assumption

- Short-memory, stationary data correlation (Li and Qiu 2020, Qiu et al. 2020).
- Namely, $\gamma(q) = Cov(X_i, X_{i+q})$ only depends on $q$ when $i$ changes, and $\gamma(q)$ tends to 0 when $q$ increases.
This is the Big Data Era!

Big Data Landscape

- **Vertical Apps**
  - Predictive Policing
  - Bloomreach
  - MyRx

- **Log Data Apps**
  - Splunk
  - Loggly
  - Sumo Logic

- **Ad/Media Apps**
  - RocketFuel
  - Collective

- **Media Science**
  - Turn
  - DataXu

- **Business Intelligence**
  - Oracle
  - Datalogix
  - Rearden Future
  - Datanyze

- **Microsoft**
  - Business Intelligence

- **IBM**
  - COGNOS
  - birst

- **SAS**
  - Teradata
  - Teradata Aster

- **TIBCO**
  - Panopticon

- **Autonomy**
  - QlikView
  - Data Tree

- **Eloqua**
  - Marketo
  - Hootsuite

- **Adobe**
  - Tableau

- **Analytics and Visualization**
  - Tableau
  - MicroStrategy
  - Birst

- **Data As A Service**
  - GoodData

- **Analytics Infrastructure**
  - Factal
  - GNIP
  - DataStax

- **Operational Infrastructure**
  - Couchbase
  - 10gen

- **Infrastructure As A Service**
  - Amazon
  - Windows Azure

- **Structured Databases**
  - MySQL
  - Microsoft SQL Server

- **Technologies**
  - Hadoop
  - Hadoop MapReduce

Peihua Qiu

SPC

13 / 24
Big Data Applications in Manufacturing Industries

An imaging system for detecting defects on steel surface
Big Data Applications in Medical Studies

Electronic Health Record – Concept Overview

The EHR represents the integration of healthcare data from a participating collection of systems for a single patient.

Each patient encounter with a department results in the capture of data.

EHR Network Services:
- Data Discovery
- Data Management
- EHR Security
- System Data Registry
- EHR Business Rules
- EHR Patient Index

EHR Data

Electronic Health Record
Patient (x)
- Admin Data (x)
- Admin Meta Data (x)
- Lab Data (x)
- Lab Meta Data (x)
- Clinical Data (x)
- Clinical Meta Data (x)
- Radiology Data (x)
- Radiology Meta Data (x)
- Pharmacy Data (x)
- Pharmacy Meta Data (x)
- Coord of Care Data (x)
- EHR Patient ID (x)
- EHR Context Data (x)

Coordination of Care
Patient (x)

The EHR Network integrates data from the systems of participating organizations to create the EHR for a specific patient/subject.

* Using Terminology from Standard Nomenclature or Structured Vocabulary
Some Common Features of Big Data Applications

- Data volume is big
- Data structure is complicated
- **Data streams** (i.e., new data keep coming over time)
- **Fundamental research questions:** whether the underlying longitudinal process of the observed data changes over time and how it changes over time.

A major statistical tool for monitoring a sequential process is **Statistical Process Control (SPC).**

- Production lines in manufacturing industries
- Internet traffic
- Surveillance of infectious diseases
- Monitoring of PM 2.5
- ...
Sequential monitoring of **high-dimensional multivariate processes**

- Multivariate SPC charts based on variable selection and LASSO (Wang and Jiang 2009, Yan et al. 2018, Zou and Qiu 2009)
- Post-signal diagnostics (Li et al. 2019, Zou et al. 2015)
- Multivariate SPC charts based on machine learning
  - Supervised learning (labelled data) - *artificial contrasts* (Deng et al. 2012, Tuv and Runger 2003)
  - Unsupervised learning (unlabelled data) - *one-class classification* (He and Zhang 2011, Ning and Tsung 2013)
Monitoring of univariate and multivariate profiles

- Monitoring univariate profiles
  - Parametric methods (Kang and Albin 2000, Kim et al. 2003, Zou et al. 2006)
  - Nonparametric methods (Qiu et al. 2010, Zou et al. 2008)
- Monitoring multivariate profiles (Paynabar et al. 2016, Ren 2019)
Monitoring of Spatial Data

- **Complicated data structure**
  - Spatial and temporal data correlation
  - Spatial data variation may NOT be described by a parametric form
  - Dynamic longitudinal spatial process (e.g., seasonality)
  - Data distribution may NOT be normal, Poisson, Negative Bionomial, ...

- **Methods based on Knox or scan statistics** (Knox and Bartlett 1964, Kulldorff 1997): retrospective, ignore data correlation, non-dynamic, assumed parametric distributions.


- **Nonparametric local smoothing** (Zang and Qiu 2018a,b): non-dynamic, ignores data correlation

- **Dynamic spatial processes** (Yang and Qiu 2020)
**Complicated data structure**: edges, all other complicated structures of spatial data

**Geometric misalignment**: due to relative move between the camera and an object at different times

**Image registration** (e.g., Qiu and Xing 2013)

Two satellite images of the San Francisco bay area taken in 1990 and 1999, their difference before image registration, and their difference after image registration.
Monitoring of Images (con’d)

- Extract some features (PCA) and monitor the extracted features (e.g., Duchesne et al. 2012, Lin et al. 2008, Yan et al. 2017, 2018).
- Focus on certain pre-specified regions (ROIs) and monitor a summary statistic of the ROIs (e.g., Jiang et al. 2011, Megahed et al. 2012).
- Edges play and important role in image monitoring

- Image comparisons (Feng and Qiu 2018)
Monitoring of network data

- Network of a subset of the Enron email corpus.

- Aggregated measures of the topological characteristics (e.g., degree) of the entire network or relevant sub-networks (e.g., McCulloh and Carley 2011, Neil 2013).

- Network state-space modelling (Zou and Li 2017)

- Recent reviews (Jeske et al. 2018, Savage et al. 2014, Woodall 2017)
Feature-based process monitoring

- What kinds of features? How many features?
- Original goals of process monitoring might have been compromised! (e.g., Use the first principal component of a sequence of images: $Y_{1n} = a_1'X_n$, for $n \geq 1$. It can NOT detect all mean shifts $\delta$ satisfying $a_1'\delta = 0$.)

- Accommodation of complicated data structure (e.g., data fusion)
- Accommodation of covariates
- Dynamic process monitoring (need to specify an IC data to estimate regular longitudinal pattern, multiple shifts, ...)

Challenges in Monitoring Big Data
Concluding Remarks

- Data streams are common in big data applications
- SPC provides a major statistical tool for such applications
- Big data streams often have complicated data structure
- Big data analysis is not just for data simplification
- Misleading statistical methods/tools are worse than nothing
- Traditional SPC methods should be modified properly and new SPC methods should be developed
- We should learn knowledge in different areas/disciplines (e.g., image processing)