Quality Engineering, 26:96–98, 2014

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2014.846102

Taylor & Francis
Taylor & Francis Group

Comments: Some Challenges for Multivariate Statistical Process Control

Dennis K. J. Lin

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania

I would like to congratulate Alberto Ferrer on his inspiring work on multivariate statistical process control. The author laid out the need for a paradigm shift, proposed a latent structure–based multivariate statistical process control (LSb-MSPC) with a case study, and finally discussed potential challenges of the proposed LSb-MSPC for data-rich environments. The new era of big data impacts many fields, including multivariate statistical process control. It is thus valuable to investigate the appropriateness of conventional SPC and then propose future SPC methodologies. This work is insightful and inspiring. I found myself very much agreeable with most his views and thus am only able to provide some more connections and challenges below, especially under the big data environment.

BIG DATA AND SPC

As mentioned by the author, the conventional SPC typically assumed data sets with low-frequency sampling (small number of observations). With today technologies, such as sensors and radio frequency identification (see, for example, Bi and Lin 2009; Wadhwa and Lin 2008), data can be collected efficiently, rapidly, and automatically. This is known as *big data*. The so-called big data typically refers to its three Vs—volume, velocity, and variety:

- Volume of the data, measured by the computer storage space, has been increased from bytes to megabytes (MB $= 10^3$ bytes), to gegabytes (GB $= 10^6$ byte), to exabytes (EB $= 10^{18}$ bytes). With today's computer facilities (both hardware and software), data can be easily collected, and even analyzed. The immediate impact is the computing issue—those statistical methodologies that work nicely for small data sets may not be feasible for large data sets. For example, any methodology with complexity $O(n^2)$ is not recommended for large data sets. Furthermore, automatically collected data typically are unstructured and contain very little useful information. Most statistical assumptions, such as independent and identical distribution, are no longer valid for large data. Consequently, many statistical theorems and methods, including SPC, need to be updated or even reinvestigated.

Address correspondence to Dennis K. J. Lin, Department of Statistics, The Pennsylvania State University, 317 Thomas Building, University Park, PA 16802-3603. E-mail: DKL5@psu.edu

- Velocity of the data has been speeded up from streaming data to second to respond, or so-called real-time, data. Automatically, (statistical) analysis is applied to gain a fast feedback to the field. For example, supply chain event management focuses on real-time feedback from the inventory to stores. SPC could and should play a critical role here.
- Variety of the data has been varied from structured (such as number, text, or even photo/image data) to unstructured (such as multimedia, YouTube video data, social Facebook data, and mobile data). SPC methodologies are rather mature for data in number format, somehow doable for text data, but are far behind for data of other types. If we attempt to monitor these data, what types of SPC techniques do we need?

It is so temping to mention the fourth V for big data. For other disciplines, the fourth V for big data should be value. However, this is rarely the case. Big data is so big, mainly because it is automatically collected. Thus, in many cases, it does not contain much information. This is sometimes called a data-rich, information-poor environment. Therefore, the fourth V cannot be value. Most of us tend to believe that the fourth V ought to be veracity—because most data are in doubt and do not contain much information. As such, statistical methodologies with stochastic features must play a critical role here. SPC is no exception.

WHAT TYPES OF DATA TO BE MONITORING: SOME EXAMPLES

As an illustrative example, consider the following five scenarios for a sequence of data to be monitored using SPC approaches:

- Time series 101—each observation is a number (scalar).
- Time series 201—each observation is a vector.
- Time series 301—each observation is a function (between response and covariates).
- Time series 401—each observation is a network.
- Time series 501—each observation is a graphic.

For time series 101, the data has the form $\{y_1, y_2, \ldots, y_T\}$, where each y_t scalar. This is a univariate time series where the conventional SPC works well and much is known. For time series 201, the data has the form $\{\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_T\}$, where each $\vec{y}_i$ is a vector. This is a multivariate time series where the LSb-MSPC is needed—although many are known, there is room for improvement. This is especially true for high-dimensional data. Time series 301 has the form $\{y_1 = f_1(X),\ y_2 = f_2(X), \ldots,\ y_T = f_T(X)\}$ where the functions are to be monitored. In reality, these f_is are unknown and need to be estimated from the data (y_t, X_t). This is the so-called functional data (profile). Some nice work has been proposed, but much more needs to be done (see also Woodall and Montgomery 2013). Time series 401 has the form $\{N_1, N_2, \ldots, N_T\}$ where N_t is a network—typically represented by a graphic or a matrix—to capture the relationships among people or company (nodes); for example, at the specific time t. Time series 501 has the form $\{G_1, G_2, \ldots, G_T\}$ where G_is are graphics in general (directional/nondirectional, unipartite/bipartite, etc.). SPC techniques to monitor data like time series 401 or 501 are lacking and seem to have strong demanding, especially for the upcoming big data era.

DIMENSIONAL ANALYSIS AND SPC

The author also discussed the use of a projection method, such as principal component analysis or partial least squares, for potential dimension reduction. This is indeed the basic idea behind LSb-MSPC. It is questionable, however, how much information can be kept when projected into low (such as one or two) dimensions. This is especially troublesome for high-dimensional data. Dimensional analysis (DA), on the other hand, has been proposed and used to reduce the number of variables.

DA is a well-developed, widely employed methodology in the physical and engineering sciences. The application of DA in statistics leads to three advantages: (1) a reduction of the number of potential causal factors that we need to consider, (2) analytical insights into the relations among variables that it generates, and (3) scalability of results. The formalization of the DA method in statistical design and analysis gives a clear view of its generality and overlooked significance. For a brief understanding of DA, see Shen et al. (2013) as well as the recent paper by Albrecht et al. (2013) with discussions is recommended. Could we somehow

apply DA to the SPC world? This seems to be a promising research area to be explored.

CRITERION FOR THE PERFORMANCE OF A CONTROL CHART: CONTINUOUS RANKED PROBABILITY SCORE

Another challenge for SPC under big data is the comparison criterion. An ideal criterion is not only informative but must also be computational easy. Run length distribution is well recognized as a key measurement for the performance of a control chart. Comparison between the run length distributions is difficult in general. Thus, average run length is typically used as a simple criterion for comparison. However, using the average to present the entire probability distribution is not informative and sometimes may be misleading. Wang and Lin (2013) proposed the continuous ranked probability score (CRPS) as a new criterion for the performance of control charts. It is shown that if a single index criterion is to be used, CRPS outperforms other existing criteria, including average run lenth.

The CRPS is defined as

$$CRPS(F,x) = \int_{-\infty}^{\infty} (F(y) - H(y-x))^2 \, dy,$$

where $H(y-x)$ denotes the Heaviside function, which takes the value 0 when $y < x$ and the value 1 otherwise. The CRPS is basically a distance measure between the target value x and the distribution function F. To apply CRPS for comparison of run length distribution, we set F to be the run length distribution and take $x = 1$. Wang and Lin (2013) provided theoretical and empirical support for using such a new criterion.

Though other powerful criteria may be raised in the near future, the issue of a suitable criterion for comparing the goodness of new control charts remains an important one.

CONCLUSION

Ferrer's article presents a fresh perspective for MSPC for the big data environment. The author proposed an LSb-MSPC and discussed potential challenges of the proposed LSb-MSPC. A broader area in general is quality assurance (for example, total quality management, Six Sigma, etc.). Facing the incoming big data era, many important concepts require updated (or even brand new) techniques to be implemented. This article is timely and could serve as an excellent example for such an evolutionary event. Congratulations again for his outstanding work. I am grateful for this opportunity to be part of the discussion. Perhaps I should add that Stu Hunter has been a true leader in our society. His original work, especially on design of experiments, had a significant impact on my research career. It is my great privilege to participate in the first Hunter Conference.

ABOUT THE AUTHOR

Dr. Dennis K. J. Lin is a university distinguished professor of supply chain and statistics at Penn State University. His research interests are quality assurance, industrial statistics, data mining, and response surface. He has published near 200 papers in a wide variety of journals. He currently serves or has served as associate editor for more than 10 professional journals and was coeditor for Applied Stochastic Models for Business and Industry. Dr. Lin is an elected fellow of ASA, IMS and ASQ, an elected member of ISI, a lifetime member of ICSA, and a fellow of RSS. He is an honorary chair professor for various universities, including Renmin University of China (as a Chang-Jiang Scholar), Fudan University, and National Chengchi University (Taiwan). His recent awards including, the recipient of the 2004 Faculty Scholar Medal Award (Penn State), the Youden Address (ASQ, 2010), the Shewell Award (ASQ, 2010), the Don Owen Award (ASA, 2011), and the Loutit Address (SSC, 2011).

REFERENCES

Albrecht, M. C., Nachtsheim, C. J., Albrecht, T. A., Cook, R. D. (2013). Experimental design for engineering dimensional analysis. *Technometrics*, 55:257–270.

Bi, H. H., Lin, D. K. J. (2009). RFID-enabled discovery of supply networks. *IEEE Transactions on Engineering Management*, 56:129–141.

Ferrer, A. (2014). Latent structures based-multivariate statistical process control: A paradigm shift. *Quality Engineering*, 26:72–91.

Shen, W. J., Davis, T., Lin, D. K. J., Nachtsheim, C. J. (2013). Dimensional analysis and its applications in statistics. *Journal of Quality Technology*.

Wang, W., Lin, D. K. J. (2013). Another look at run length distribution.

Wadhwa, V., Lin, D. K. J. (2008). Radio frequency identification: A new opportunity for data science. *Journal of Data Science*, 6:369–388.

Woodall, W. H., Montgomery, D. C. (2013). Some Current Directions in the Theory and Applications of Statistical Process Monitoring, submitted for publication.