Rejoinder

I would like to thank Dennis Lin and Bill Woodall for their stimulating discussion of the role of SPC in data-rich (*i.e.* BIGdata) environments. I appreciate the opportunity to comment on their valuable contributions.

Dimension reduction

Although dimension reduction is achieved with LSb-MSPC, I would like to stress that the most important benefit of these multivariate projection techniques is their ability to cope with Big Data sets by projecting them into low-dimensional subspaces. This way each original multivariate observation is decomposed in two orthogonal parts: the projection onto the subspace and the residual distance from the observation to the projection. This residual distance plays a critical role in process monitoring. In fact, in many processes failures generate a breakage of the in-control correlation structure yielding an out-of-control signal in the residual distance control chart. Caution should be taken to avoid simplistic approaches discarding residual information and using only control charts based on the projected information (*i.e.* scores).

Understanding of LSb-MSPC methods

I do not think that understanding LSb-methods is more difficult than getting familiar with traditional control chart statistical theory. In fact, low-statistically trained people from industry regularly attend LSb-teaching courses taught by companies who develop commercial LSb-software (as, e.g. SIMCA or ProSensus).

Regarding the application in practice, I guess that not only LSb-methods but also any statistical method requires a team-based approach in collaboration with subject matter experts. I guess this is the key for scientific relevance.

Work done by members of Group 3 over the past

Bill refers to a review paper (Woodall and Montgomery 2013) on some work done on process monitoring during the last ten years in the area of dimension reduction with large data sets. These newer methods include tools for monitoring functions, monitoring multistage processes and monitoring with spatiotemporal data. I will not go into a detailed study but just comment on some issues regarding the need of a paradigm shift especially for members of Group 3.

After browsing some of the references cited by Bill I have found one of the key characteristics regarding conventional MSPC: the focus is on monitoring quality attributes. As an example, in Woodall and Montgomery (2013) (Section 4. Multivariate Methods) authors state "*multivariate methods are needed whenever one wants to monitor several <u>quality</u> variables and take advantage of any relationships among them*". As commented in my paper a paradigm shift is needed to also look at all the process and input variables involved.

Monitoring of functions (i.e. profile monitoring) is also addressed in LSb-MSPC as one of the potential approaches for dealing with batch processes. In Wold *et al.* (2009) this is called "landmark feature extraction" approach. The paradigm shift in this case calls for using not only the quality profile but also the process variables that have been measured during the manufacturing of the part.

A *multistage* (multi-step) system refers to a system consisting of multiple components, stations, steps or stages required to finish the final product or service. In one of the key references provided the focus in again on quality attributes: "the quality characteristics at one stage are not only influenced by local variations at that stage, but also by variations propagated from upstream stages" (Shi and Zhou 2009). Multistage systems are particular cases of more complicated scenarios where the number of variables is large and additional information is available for blocking the variables into conceptually meaningful blocks. In these cases, there may be a strong temptation to drastically reduce the number of variables to a smaller, more manageable number. This temptation is further strengthened by the 'regression tradition' to reduce the variables as far as possible to get the data matrix well conditioned. Such a reduction of variables, however, often removes information and makes the fault detection and diagnosis misleading. A better alternative is to divide the variables into conceptually meaningful blocks and then apply LSb-multiblock methods (Wold *et al.* 1996, Westerhuis *et al.* 1998). In batch processes these methods allow one to use not only the measured trajectory data on all the process variables and information on measured final quality variables but also information on initial conditions for the batch such as raw material properties, initial ingredient charges and discrete operating conditions.

The approach proposed by Megahed *et al.* (2011) for adapting multivariate charts for *spatiotemporal surveillance* follows the parametric statistical distribution-based (*i.e.* theoretical) approach of control charting advocated by Group 3. The approach is illustrated with grayscale images under several assumptions. I wonder to what extent this method could also be applied to more informative images as color or hyperspectral images. A comparative study with other multivariate image analysis methods such as those reviewed by Duchesne *et al.* (2012) and Prats-Montalbán *et al.* (2011) would be welcomed.

Another reference provided by Bill (Schall and Chandra 1987) also refers to multivariate quality control. In my opinion their approach lacks applicability in data-rich environments. They assume there are fewer input variables than output variables and that the sample size is large compared to the number of variables (usually the opposite occurs in practice). A PCA is fitted to the output variables assuming the data matrix is full rank. The method is illustrated with a numerical (not real) study with 5 inputs and 10 outputs. Although multivariate, this is not an example typical of a data-rich environment.

Automobile manufacturing case study (Ferrer 2007)

The automobile manufacturing case study used by Ferrer (2007) is a simpler scenario than the petrochemical case study discussed in this paper. The motivation of that paper was not to make a comparison study with different techniques based on principal component methods but to illustrate the drawbacks of conventional (M)SPC techniques. Note that at the time the study was done the process was being monitored by univariate control charts of selected "critical" dimensions.

Continuous Ranked Probability Score (CRPS)

Dennis proposes the CRPS as a new criterion for comparing the performance of control charts. This index relies on the distribution function of the monitored characteristic. I wonder to what extent this new index is really useful and applicable to BIGdata scenarios.

Acknowledgements

It has been my great privilege to participate in this conference honouring Stu Hunter in his 90th birthday. I wish this particular view of the challenges for multivariate statistical process control provided will foster interactions between people interested in this field.

Quoting George E.P. Box, who passed away few days after this conference (March 28th 2013), "*in any feedback loop it is, of course, the error signal – for example, the discrepancy between what tentative theory suggests should be so and what practice says is so – that can produce learning. The good scientist must have the flexibility and courage to seek out, recognize, and exploit such errors – especially his own*" (Box 1976),

References

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association,* 71(356): 791-799.

Duchesne, C., Liu, J.J., MacGregor, J.F. (2012). Multivariate image analysis in the process industries: A review. *Chemometrics and Intelligent Laboratory Systems*, 117:116-128.

Ferrer, A. (2007). Multivariate Statistical Process Control based on Principal Component Analysis (MSPC-PCA): Some Reflections and a Case Study in an Autobody Assembly Process. *Quality Engineering*, 19:311-325.

Megahed, F. M., Wells, L. J., Camelio, J. A., Woodall, W. H. (2012). A spatiotemporal method for the monitoring of image data. *Quality and Reliability Engineering International*, 28: 967- 980.

Prats-Montalbán, J. M., de Juan, A., Ferrer, A. (2011). Multivariate image analysis: a review with applications. *Chemometrics and Intelligent Laboratory Systems*, 107:1–23.

Schall, S., Chandra, J. (1987). Multivariate quality control using principal components. *International Journal of Production Research*, 25(4): 571-588.

Shi, J., Zhou, S. (2009). Quality control and improvement for multistage systems: a survey. *IIE Transactions*, 41(9): 744-753.

Westerhuis, J. A., Kourti, T., Macgregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* 12:301–321.

Wold, S., Kettaneh, N., Tjessem K., (1996). Hierarchical multiblock PLS and PC models for easier interpretation and as an alternative to variable selection. *J. Chemometrics*, 10:463–482.

Wold, S., Kettaneh-Wold, N., MacGregor, J.F., Dunn, K.G. (2009). Batch process modelling and MSPC. In: Brown, S.D., Tauler, R., Walczak, B. (Eds.), Comprehensive Chemometrics (Vol. 2), Elsevier B.V., Oxford, 163-195.

Woodall, W. H., Montgomery, D. C. (2013). Some current directions in the theory and application of statistical process monitoring. To appear in the *Journal of Quality Technology*.