Ferrer, A. (2021). Multivariate six sigma: A key improvement strategy in industry 4.0. Quality Engineering. 33(4):758-763. https://doi.org/10.1080/08982112.2021.1957481

Multivariate Six Sigma: a key improvement strategy in Industry 4.0

Alberto Ferrer

Department of Applied Statistics, Operations Research and Quality,

Universitat Politècnica de València, Valencia, Spain

Camino de Vera s/n. Edificio 7A, 46022, Valencia (Spain). aferrer@eio.upv.es

Alberto Ferrer is Head of the Multivariate Statistical Engineering Research Group (https://giem.blogs.upv.es/) and Professor of Statistics at the Department of Applied Statistics, Operations Research and Quality at Universitat Politècnica de València (Spain). His main interest focuses on (big) data analytics and statistical techniques for quality and productivity improvement, especially those related to latent variables-based multivariate statistical methods for both continuous and batch processes (chemical, bio, pharma, …). He is active in industrial teaching and consultancy activities on (Big) Data Analytics, Six Sigma, Process Analytical Technology (PAT), Multivariate Image Analysis (MIA), Process Chemometrics and Statistical Methods for Knowledge Discovery.

Multivariate Six Sigma: a key improvement strategy in Industry 4.0

This paper aims to generate a reflection on the changes that need to be made in the Six Sigma strategy so that it continues to be a successful improvement strategy capable of facing the new challenges derived from Industry 4.0.

Keywords: Six Sigma; Industry 4.0; data science; Big Data; multivariate statistics; principal component analysis (PCA); partial least squares regression (PLS)

Motivation

Is this true that Data Science, with its powerful machine learning tools, is the future in which we must bet to guarantee the survival of companies in the new Big Data environment of Industry 4.0? Does it make sense to continue betting on the scientific method and statistical thinking with subject matter knowledge for understanding the root causes that cause problems, or all we need is exploiting the abundance of data generated in this new environment with powerful algorithms and computing facilities? What are the challenges for Six Sigma to remain a successful improvement strategy in Industry 4.0?

The "Big" Data tsunami

Industrial enterprises of the 21st century are facing a big new challenge: the so-called data tsunami, creating the Industry 4.0 paradigm. Digitalization is its driving force fostered by the Industrial Internet of Things (IIoT). In other sectors -such as sales, finance, marketing, and social networks- smartphones and Internet are contributing in a similar way to increase dramatically the amount of data registered creating the Big Data movement. This is characterized by the four V´s: volume, variety, velocity and veracity. For the first time in history, we have data everywhere and there is a belief that data contain useful information that has to be mined for helping the decision-making process

(Ferrer 2020).

In this new environment a new discipline has emerged: Data Science. Lots of companies are offering a new type of job: data scientist; and an apparent new discipline is emerging: data analytics, with the intended goal of effectively transform data to knowledge, creating added value by enabling cost reductions, productivity gains, or revenue increases (White 2016).

Data scientist are highly trained in processing large data sets, programming in Python, SQL, R, etc., applying machine learning methods such as neural networks, support vector machines and random forests, and mastering impressive visualization tools. Except in few exceptions, scarce training (if any) in statistical thinking is usually included in the syllabus (Hardin et al. 2015). Their interest is usually focused on forecasting the future or classifying new observations, i.e., extracting features or discovering patterns from databases that, used as predictors in the algorithms developed, yield good predictions or classifications. Given the powerful computing resources available and the ease of modeling nonlinearities, hundreds of complex predictive models can be fitted and updated at a very low cost. For these predictive goals only correlation (not causation) is needed, and good predictions may be obtained just by playing with the (abundance) available data (by trial and error).

In industry, this approach is used for building soft sensors that forecast the value of a low sampling frequency and hard-to-measure response variable, based on high sampling frequency and easy-to-measure process variables. In these situations, there is really no new scientific knowledge (i.e., process understanding) acquired; only features, patterns and good predictions (i.e., the so-called high-level knowledge extracted from low-level data), but not science.

As an industrial consultant, I am witnessing a lot of experts in information technology (IT) landing in the industry under the claim of the digitization required by industry 4.0, spreading two worrying statements. The first one comes from Anderson (2008). In this paper, entitled “The End of Theory: the Data Deluge that Makes the Scientific Method Obsolete”, it is stated that because of the quantity and speed of data production, new technologies could now solve major scientific and industrial problems solely through empirical data analysis, without the use of scientific models, theory, experience, or domain knowledge. This is generating a (false) belief that all we need are lots of data, computer facilities and machine learning algorithms, and by “pushing the bottom” we would expect successful results. This is based on what Crawford (2013) calls Data Fundamentalism, forgetting the fact that data quality matters, and big data do not necessarily mean value data because they may be biased data (i.e. large-sized samples not representing the whole picture of the population under study). A second worrying statement is that we no longer have to be fixated on causality and the world is shifting from causation to correlation (Mayer-Schönberger and Cukier 2013). The consequence is that given the “black-box” nature of the algorithms used and the usual correlation of predictors, model interpretation to gain process understanding (i.e., scientific knowledge) of the problem addressed is not possible (Ferrer 2020).

But, where do these beliefs come from? One possible explanation is that this forecast-based data analysis culture with "black box" models have generated significant economic revenues for some companies analyzing customer data or in applications where the only thing that matters is to obtain a good prediction or classification such as, for example, in soft sensors for feeding control loops, voice recognition or image identification systems. The problem is believing that what works in a certain area will work in whatever area is applied.

Six Sigma challenges

Certainly, the predictive applications commented in the previous section, typical of most data science projects, are not suited for troubleshooting, process improvement and optimization. These are critical goals in industry and technology, and scientific method (i.e. iterative inductive/deductive approach) and causal models are required. Causality implies that for any active changes in the adjustable variables in the process, the model will reliably predict the changes in the output of interest (MacGregor 2018). The question is how can we pursue these goals in Industry 4.0?

For the past forty years, Six Sigma has been promoting a never-ending improvement culture based on a strong and professionalized organization for improvement, a clear and well thought five-step DMAIC cycle (i.e., Define, Measure, Analyze, Improve and Control) empowered by the scientific method, and using statistical thinking as a catalyst of the learning process (Snee and Hoerl 2003). But Six Sigma, conceived in the 80s of the last century, must be revisited, to meet the challenges of data-rich environments in Industry 4.0.

What do come first: questions or data?

Six Sigma was created in a data-scarce context: low number of variables with small sample sizes (usually more observations than variables) that in most of the cases had to be measured (or generated) (i.e., there were no data available at the beginning of the Six Sigma project). Once defined the project, the Measure phase begins not by measuring but by asking questions, following the typical Question-Data-Analysis (QDA) paradigm in Statistics (Cao 2019). However, in the data-rich context typical of the Big Data era, in most of the situations, a lot of data are already available before the project is even defined. Some people think that the QDA has to be changed to a Data-Questions-Analysis (DQA) paradigm, claiming that data come first. Nevertheless, data have no

meaning in themselves, the focus has to be on the problem to be solved and data is just one of the resources to help accomplish the goals. Therefore, even in data-rich contexts I advocate for using the classical QDA paradigm following the fundamental principles of statistical methodology, often ignored in some data science applications: i) critical evaluation of data quality (i.e. the *data pedigree*); ii) integration of sound subject-matter knowledge; iii) development of an overall strategy for attacking the problem; and iv) sequential approaches versus "one-shot studies" (Hoerl, Snee and De Veaux 2014). In this QDA paradigm, the problems to be solved that define the critical-to-quality (CTQ) attributes of Six Sigma projects generate relevant questions that guide the searching for the relevant data to be selected and analyzed for getting the answers. Having lots of data is not any guarantee of success, quantity does not mean quality, data may be biased and not representative of the population under study.

Nature of Data in Industry 4.0

IT revolution in Industry 4.0 has not caused only a change in the *number* of the variables (that in some cases is even higher than the number of observations) but also a change in the *nature* of the registered data. Nowadays it is possible to register data from customers, quality, process and even from equipment. Sensors are providing different type of signals: chemical (i.e., spectra), physical (i.e., pressures, temperatures, flows, etc.), biochemical (i.e., pH, conductivity, dissolved oxygen, etc.), digital (i.e., electronic eyes), potentiometric (electronic noses and tongues), acoustic (i.e., electronic ears), and so on. These data are mostly collected from daily production and often exhibit high auto and cross correlation, rank deficiency, low signal-to-noise ratio, multi-stage and multi-way structure, and missing values. In most of the cases they are happenstance data (i.e., data from daily routine production and not generated from any experimental design) and, therefore, correlation does not necessarily mean causation. The complexity of this

type of data requires special skills to manage them. This is why some companies are defining a new job position as *data engineer*. These people should be involved in the Six Sigma project teams for helping in the data acquisition tasks.

Revisiting the Six Sigma statistical toolkit

Process data in industry, although shares many of the characteristics represented by the four V´s (i.e., volume, variety, veracity and velocity), may not really be Big Data in comparison to other sectors such as social networks, sales, marketing and finance. However, the complexity of the questions we are trying to answer with industrial process data is really high, and the information that we wish to extract from them is often subtle. This info needs to be analyzed and presented in a way that is easily interpreted and that is useful to process engineers. Not only do we want to find and interpret patterns in the data and use them for predictive purposes (as in classical data science projects), but we also want to extract meaningful relationships that can be used to improve and optimize a process (García-Muñoz and MacGregor 2016).

Traditional Six Sigma statistical toolkit, mainly focused on classical statistical techniques (such as scatterplots, correlation coefficients, hypothesis testing, and linear regression models from experimental designs), is seriously handicapped for problem solving using Industry 4.0 process data. New approaches have to be incorporated. I suggest augmenting the Six Sigma toolkit with machine learning tools and latent variable-based multivariate statistical techniques such as principal component analysis-PCA (Jackson 1991)) and partial least squares regression – PLS (Wold, Sjöström and Eriksson 2001). Machine learning tools may play a critical role in the Measure and Analyze phases to find and interpret patterns and for predictive purposes (as in classical data science projects). These tools may help uncover potential inputs variables related to CTQ attributes. Therefore, data scientist should be part of the Six Sigma projects teams.

PCA is a superb exploratory tool to be used also in the Measure phase and at the beginning of the Analyze phase. PCA and PLS can be used for (multivariate) process monitoring, fault detection and diagnosis, typical tasks in multivariate statistical process control (MSPC). Traditional univariate SPC (taught in most of the Six Sigma training courses) should be enriched with MSPC. PLS can also be used for building predictive models (as machine learning tools). But the most remarkable characteristic of PLS models is that they not only model the relationship between X and Y (as classical linear regression and machine learning models do), but also provide models for both the X and Y spaces. This fact gives them very nice properties: uniqueness and causality in the reduced latent space (this is the only space within which the process has varied) no matter if the data come either from a design of experiments (DOE) or daily production process (historical/happenstance data) (MacGregor et al. 2015). These properties make them suitable for process understanding, troubleshooting and optimization in Industry 4.0.

Causal models: the role of DOE

The goal of any Six Sigma project is to gain process understanding, that is, to discover which adjustable input variables X (i.e. raw material properties, processing conditions, etc.) are causally related to the process output Y (i.e., CTQ attribute) and predict what change in the CTQ attribute will cause any change in the X variables. This is expressed mathematically as $Y=f(X_1, X_2,\ldots, X_J)$. By inverting these causal models it is possible to obtain the values of the inputs that give rise to the desired value for the CTQ attributes, allowing troubleshooting and process optimization. Data-driven models are usually resorted to for fitting these causal models.

To guarantee causality when using data-driven approaches, however, independent variation in the input variables is required (Box, Hunter and Hunter 2005). The design

of experiments (DOE) is the "jewel in the crown" of the Six Sigma tools, specially designed to obtain causal empirical models. The problem is that in Industry 4.0 classical DOE using fractional factorial designs, as usually done in many Six Sigma projects, may be difficult to carry out, if not unfeasible (i.e. the number of potential factors to consider as inputs can be really high, and due to the complex correlation structure among them there are a lot of restrictions that prevent moving some factors independently from others). On the other hand, nowadays large amounts of historical data are available in most production processes. The problem is that these data are highly collinear and low rank, therefore, input-output correlation does not mean necessarily causation, and classical predictive models (such as linear regression and machine learning models) cannot be used for process optimization.[1]

So, is this possible to use the abundant historical data to guide the process optimization? The answer is yes, but using the DOE approach in a different way as it is usually done in classical Six Sigma projects. One approach is to sample (filter) the large data base of process data for observations that were closed to the design points of a fractional factorial design in X-space, and fit a classical linear regression model with simple effects and interactions. This retrospective DOE approach is simple but requires selecting a small number of the potential X variables that can be manipulated independently (Wold et al. 2004).

A second approach is by fitting PLS models. As already commented, PLS models provide uniqueness and causality in the reduced latent space no matter if the data come either from a DOE or daily production process, therefore, optimization can be done in the latent space (Jaeckle and MaGregor 2000). This implies that it is possible to

[1] Note that this is a mistake that is beginning to occur in industry 4.0 with the indiscriminate use of machine learning tools.

estimate the values of the latent variables that guarantee the desired values in the CTQ attributes. From these latent variables values the settings of the original X variables are obtained. Nevertheless, there are limitations in the optimization in the latent space. The latent variables cannot be explicitly manipulated by the user, but the original X variables can be manipulated in a way that changes on the raw materials and process conditions are done along the directions of the latent variables, which is equivalent to implicitly "manipulating" the latent variables themselves. This, of course, implies that the optimization in the latent space is highly restricted, since it will only provide solutions that respect the correlation structure of the PLS model (i.e., it will only allow us to modify the process in specific ways, so that the original X variables are not varied independently from each other, and any solution will abide by the correlation structure defined by the subspace of the PLS regression model). Furthermore, since the initial number of X variables involved is reduced to a smaller number of uncorrelated latent variables, the computational cost of any optimization problem in the latent space will decrease in comparison to the equivalent problem in the original space (Palací-López et al. 2019, Tomba et al. 2012).

A third approach is to run a DOE in the latent space. This implies that the design variables are not selected from the high number of original correlated X variables but from the few orthogonal scores from the PCA or PLS models fitted with the abundant historical process data. Some people call this as multivariate design-MVD (Wold et al. 1986). Because the latent variables are independent, as opposed to the original X variables, which may be correlated, the latent variable space is the only space in which a truly orthogonal design can be accomplished. Each latent variable is a linear combination of these original variables, therefore choosing levels of latent variables for a factorial experiment correspond to moving several of the original X variables up

and/or down together (Nichols 2011). Some examples of these MVD can be found in Wold et al. (2004).

Multivariate Six Sigma

In order to address the challenges that Big Data is creating in Industry 4.0, Six Sigma statistical toolkit, based mainly in least squares techniques, has to incorporate new tools such as machine learning and latent variables-based multivariate statistical techniques, giving rise to the so-called Multivariate Six Sigma (Palací-López et al. 2020).

Some examples of the integration of multivariate statistical tools into the Six Sigma toolkit are available in the literature. For example, Peruchi et al. (2020) integrated principal component analysis (PCA) into a Six Sigma DMAIC project for assessing the measurement system, analyzing process stability and capability, as well as modeling and optimizing multivariate manufacturing processes in a hardened steel turning case involving two CTQ attributes. In Ismael et al. (2018), discriminant analysis and PCA were integrated into the DMAIC Six Sigma framework in order to improve the quality of oil type classification from oil spills chromatographic data.

Palací-López et al. (2020) show a case study of a successful integration of latent variables models (PCA and PLS) into the DMAIC problem solving strategy using historical data from past production of a batch process at a chemical plant. In the Measure phase, validation of the data to detect and diagnose potential outliers was done using PCA. In the Analyze phase PCA was used to explore the correlation structure among the process variables and CTQ attributes and to detect cluster of batches that were operated in a similar way. Afterwards, a PLS regression model permitted predicting the CTQ attributes from the process variables and determining which of these process variables have a significant relationship on the CTQ attributes. The discriminant version of PLS (i.e., PLS-DA) was used to identify which process variables behaved in

a different way between batches with good and bad performance. As a result of the analyses performed, in the Improve phase the team responsible for this process was able to pinpoint a specific behavior in the process potentially related to the bad performance. Once the causes of the bad performance were detected and addressed, in the Control phase a multivariate monitoring scheme was implemented in the plant to detect and diagnose possible deviations in the process variables. The solution implemented remains a success to this day, with estimated benefits/savings above 140,000 €/year (40,000€ higher that the initial estimation in the Define phase).

Lessons learned

Data Science projects without a scientific method approach and a proved problem-solving strategy are doomed to failure in Industry 4.0 when process understanding is needed for troubleshooting and process optimization. The credibility of the Six Sigma strategy as a useful methodology for process improvement in Industry 4.0 depends on the speed with which it integrates new tools from the field of artificial intelligence (such as machine learning techniques) and multivariate statistics (such as latent variable techniques), as well as the skill in using them for what they serve in the DMAIC cycle. This evolution of Six Sigma into what has come to be called Multivariate Six Sigma requires more successful case studies to convince managers of its usefulness.

References:

Anderson, C. 2008. The end of theory: the data deluge makes the scientific method obsolete. *WIRED Magazine*. 16(7).

Box, G.E.P., Hunter, W.G., and Hunter, J.S. 2005. *Statistics for Experimenters: Design, Discovery and Innovation*. 2nd ed. Hoboken, NJ: John Wiley and Sons.

Cao, R. 2019. Comments on: Data Science, big data and statistics. *Test.* 28(3):664-670.

Crawford, K. 2013. *The Hidden Biases in Big Data*. Cambridge, UK: Harvard Business Review.

Ferrer, A. 2020. Discussion of "A review of data science in business and industry and a future view" by Grazia Vicario and Shirley Coleman. *Appl. Stochastic Models Bus. Ind.,* 1–7.

García Muñoz, S., and MacGregor, J.F. 2016. Big Data. Success Stories in the Process Industries. *Chemical Engineering Progress* 112(3):36-40.

Hardin, J., Hoerl, R., Horton, N.J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D.T., and Ward, M.D. 2015. Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician*. 69(4):343-353.

Hoerl, R.W., Snee R.D., and De Veaux R.D. 2014. Applying Statistical Thinking to 'Big Data' Problems. *Wiley Interdisciplinary Reviews: Computational Statistics*. 6:222-232.

Ismail, A., Mohamed, S.B., Juahir, H., Toriman, M.E., and Kassim, A. 2018. DMAIC Six Sigma Methodology in Petroleum Hydrocarbon Oil Classification. *Int. J. Eng. Technol.* 7:98–106.

Jackson, J.E. 1991. *A User's Guide to Principal Components*. New York:Wiley.

Jaeckle, C.M., and MacGregor, J.F. 2000. Industrial applications of product design through the inversion of latent variable models. *Chemometrics and Intelligent Laboratory Systems*. 50:199–210.

MacGregor, J.F. 2018. Empirical Models for Analyzing "big" data - what´s the difference. In: Spring AIChE Conf., Orlando, Florida, USA.

MacGregor, J.F., Bruwer, M.J., Miletic, I., Cardin, M., and Liu, Z. 2015. Latent variable models and big data in the process industries, *IFAC-PapersOnLine*. 28:520–524.

Mayer-Schönberger, V., and Cukier K. 2013. *Big Data: A Revolution that Will Transform How We Live, Work and Think*. Boston, MA: Eamon Dolan/Houghton Mifflin Harcourt.

Nichols, E. 2011. *Latent variable methods: case studies in the food industry*. Open Access Dissertations and Theses. Paper 6294.

Palací-López, D., Facco, P., Barolo, M., and Ferrer, A. 2019. New tools for the design and manufacturing of new products based on Latent Variable Model Inversion. *Chemometrics and Intelligent Laboratory Systems*. 194:103848.

Palací-López, D., Borràs-Ferrís, J., da Silva de Oliveira, L.T., and Ferrer, A. 2020. Multivariate Six Sigma: A Case Study in Industry 4.0. *Processes*. 8:1119.

Peruchi, R.S., Rotela Junior, P., Brito, T.G., Paiva, A.P., Balestrassi, P.P., and Mendes Araujo, L.M. 2020. Integrating Multivariate Statistical Analysis Into Six Sigma DMAIC Projects: A Case Study on AISI 52100 Hardened Steel Turning. *IEEE Access*. 8:34246–34255.

Snee, R.D., and Hoerl, R.W. 2003. *Leading Six Sigma: A Step by Step Guide Based on experience with GE and Other Six Sigma Companies*. Upper Saddle River, NJ: Pearson Education.

Tomba, E., Barolo, M., and García-Muñoz, S. 2012. General framework for latent variable model inversion for the design and manufacturing of new products, *Industrial Engineering Chemistry Research*. 51:12886–12900.

White, D. 2016. Big Data. What is it? *Chemical Engineering Progress*, 112(3):32-35.

Wold, S., Sjöström, M., Carlson, R., Lundstedt, T., Hellberg, S., Skagerberg, B., Wikström, C., and Öhman, J. 1986. Multivariate Design. *Analytica Chimica Acta*. 191:17-32.

Wold, S., Sjöström, M., and Eriksson, L. 2001. PLS-Regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 58:109–130.

Wold, S., Josefson, M., Gottfries, J., and Linusson, A. 2004. The utility of multivariate design in PLS modelling. *Journal of Chemometrics*, 18:156-165.