HIGH DIMENSIONAL DATA: A GROWING BUSINESS

Bart De Ketelaere and Paul Darius
MeBioS, Katholieke Universiteit Leuven, Belgium
bart.deketelaere@biw.kuleuven.be

When it comes down to understanding, predicting or optimizing business, the tools provided by statistics form an important corner stone. Training of statistics at university level is a crucial factor here and should be well aligned with the actual needs after graduation. Those needs depend on the actual business type involved, and are approached here from an engineer's point of view. We will briefly touch some general trends and issues that might be considered when forming the new generation of engineers and statisticians.

INTRODUCTION

When looking at history, statistical methods were often introduced in science and engineering disciplines by statisticians that were primarily great scientists–chemists, engineers, agronomists such as Fisher, Box, Tukey, … It was advantageous that those researchers clearly understood the way engineers and scientists thought, and they had a clear understanding of the real-life problems. In a sense, they developed new methods starting from these real life problems and not so much from a pure theoretical, fundamental/statistical point of view. As McGregor (1997) states, the period of early developments in statistics up to the 1970's might be considered to be the "golden years of applied statistics". From the 70's on, a shift towards statistics as an increasingly mathematical discipline can be observed, and advances in applied statistics were only limited. This shift can be attributed to the fact that the leadership in the statistical disciplines passed on to a new generation of mathematical statisticians.

During the most recent decade, there seems to be evidence that we are once again seeing a major shift in the leadership and direction of the statistical community. This new era is being catalyzed by new developments in the broad field of engineering and quality control: novel sensor technologies entered the broad market for laboratory quality assessment, a focus on online process control, the availability of powerful data acquisition and processing systems, ... The availability of such systems has totally changed the nature of the data we are dealing with: we now live in a highly multivariate, data-rich society and are being inundated with data from all directions.

This change in data characteristics that we are exposed to in daily routine has opened the door to the need for new statistical methods that are capable of handling such large volumes of data that are collected within narrow time spans–often thousands of variables are measured in a fraction of a second. Not only the processing of such datasets but also the interpretation forms an important aspect of the daily activity of our new generation of statisticians–from communications, image analysis, biotechnology, management to chemistry and process industries–so that *quantity is translated into quality*.

We believe that it is crucial to consider these trends when forming the new generation of statisticians–a generation that might have a bright future ahead when scientists that are the owners of the problems and statisticians mix ideas, paralleling the earlier era of applied statistics.

This paper focuses on the shift in data characteristics of the modern world and treats some general and specific issues that might be considered when forming the new generation of statisticians. It encompasses issues such as multivariate data analysis, statistical process monitoring and sequential design of experiments.

DATA CHARACTERISTICS

Engineering is a rapid changing discipline that is characterized by an ever increasing amount of automation. The higher degree of automation and related increasing production speeds required together with the strong expectations of the customers have lead to the development of a broad range of novel sensor technologies that all produce mass data in a fraction of a second, at a cost efficient price. One can think of cheap camera technologies, vibration sensors and optical sensor technologies to name a few. Related, also the computer power has increased significantly, so that engineers have available a broad range of tools for attaining their goals. Those developments have as a consequence that the type of data and the way they are collected is subject

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

to change. Whereas the engineer would have had information about one single quality characteristic of a limited sample of the total production batch some decades ago, nowadays he may have collected hundreds or even thousands of different characteristics of each and every sample, possibly even with repeated measures over time.

Consider for instance the case of apple quality, for which color and internal quality are of utmost importance. The classical ways of assessing those attributes are often time consuming, subjective and destructive. A potential buyer of a stock of apples would quantify apple color using a simple color card score, whereas the internal quality is measured by a destructive test – pinching out a piece of apple flesh that is further used for sugar content measurements. This process is typically performed for a small number of apples, and the result is generalized over the whole batch. These classical methods for quality assessment are nowadays replaced using rapid, nondestructive sensors: a Near Infrared (NIR) sensor that produces light absorption information for a large number of wavelengths (spectra)–typically more than 100–gives accurate information about the internal quality and skin color. Such methods are able to inspect 10 fruits per second and are not only applied for offline use, but are recently also used for online purposes, inspecting each and every apple on a conveyor belt. An example of such spectral data is given in Figure 1 (left). Very typical in such datasets is that *the number of variables exceeds the number of samples* analyzed, and that the *correlation among the different variables is very high* (say, $r > 0.9$).

When talking about quality of a given process or product, it is also important to stress that *quality itself is more often than not a multivariate property and must be treated as such*. By this it is meant that a high quality product must simultaneously have the right combination of all the individual aspects. Each individual aspect by itself has little meaning.

HIGH DIMENSIONAL DATA EXPLORATION

Good statistical practice includes a data exploration step as the first critical step to go through once the data are collected. In case of low dimensional data, simple histograms, scatter plots or scatter plot matrices that visualize two-by-two relations amongst variables are easy to set up and to use. They offer the investigator a quick overview of possible outliers, the spread in the data and, in case of more than one variable, of the correlation structure. Besides graphical data exploration also simple statistics aiming at the detection of outliers are well described (Kutner et al., 2004). Due to the ease of use, the data exploration step is seldom a stumbling block for practitioners.

When the dimensionality of the data increases, the crucial step of data exploration becomes less straightforward and it is generally known that finding outliers becomes more difficult in such cases (Rocke & Woodruff, 1996). For the spectral data depicted in Figure 1 it is far more difficult to detect outliers using higher-mentioned graphical displays such as histograms or scatter plot matrices. Indeed, when inspecting one variable at a time by means of histograms, every data point might be considered “in control”, but when including the correlation structure this data point may be a multivariate outlier. Techniques that provide quick evidence of outliers in high dimensional datasets are very useful. The Hotelling’s T^2 statistic for instance provides a quick measure for outlyingness. However, in cases where the correlation among variables is very high ($r > 0.9$) and the number of variables becomes larger than the number of samples, the Hotelling’s T^2 is inappropriate and alternatives are needed.

Dimension reduction techniques often provide a useful means both for graphical inspection as well as for more quantitative analysis. One of the most widely used dimension reduction techniques is Principal Component Analysis (PCA). PCA involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible (Jolliffe, 2002). In highly correlated datasets such as spectral data, the first two principal components often explain a substantial amount of the total variability (say, more than 50 %), and a projection of the dataset onto these first two principal components provides already a first quick view upon the dataset. Gabriel (1971) added to this plot also the loadings–the relative importance of a certain original variable for a given principal component to form a so-called biplot. Biplots are useful tools for multivariate exploratory data analysis not only for

detecting groups or outliers in the dataset, but also to visualize the correlation structure among variables in the dataset.

A PCA analysis can be used for some more quantitative analyses as well. One can use the distance between the actual data point projected onto the PCA space and the centre of the space (the so-called Hotelling's T^2 statistic but now based on the principal components instead of the original variables) to detect points that have a "normal" correlation structure but which are in a sense shifted. Alternatively, the Q statistic (also referred to as the Squared Prediction Error, SPE, statistic) provides useful information about points that do not comply with the "normal" correlation structure of the data considered (Ostyn et al., 2007).

Although the techniques needed for exploratory data analysis for high dimensional data are in a sense standard, they require some background in multivariate statistics, and the actual implementation in routine software asks for substantially more time than was the case for low dimensional data. It is seen in practice that the data exploration step is often overlooked, and that multivariate data analysis often comes down to a situation where "the data are crunched without having a clear feeling of them". Also the fact that the person that has set up the experiment and collected the data is not the same person that actually processes the data to draw conclusions is a dangerous one–a strong background in the data generation itself is of utmost importance.

Concluding, the increasing complexity of modern data makes the data exploration step less straightforward–where this step was almost trivial in low dimensional data, nowadays more sophisticated techniques should be used in order to have a good grip on them. *Techniques for multivariate data exploration should be part of any education for engineers and statistical practitioners in general* so that the person who set up the experiment is also capable of treating the data.

CALIBRATION AND VALIDATION FOR HIGH DIMENSIONAL DATA

High dimensional data collected from novel sensor technologies often have the property that the *correlation among the different variables is very high* ($r > 0.9$). This is for instance the case in spectral data that are found in a wide range of businesses.

Classical statistical methods tend to fail in most of such cases–the number of samples is often (much) lower than the number of variables considered. Instead, inverse methods such as Principal Component Regression (PCR) or Partial Least Squares (PLS) are mostly used for prediction purposes, whereas discriminant analysis or even machine learning techniques such as Support Vector Machines (SVM) and Neural Networks (NN) are widespread for classification. Those dedicated techniques that are fit to handle the high dimensional data are believed to gain more importance in the future, where the degree of automation will further increase.

PCR combines Principal Component Analysis (PCA, see higher) as a data reduction technique with classical multiple linear regression to predict the response. Both steps are widely described in the literature (e.g., Jolliffe, 2002; Kutner et al., 2004), and are in some cases part of master courses given to engineering students. PLS, however, is only rarely taught in university courses for engineers or statisticians. This is in some sense surprising, since it is probably the most widely used multivariate technique for prediction. Related to this aspect, it is also observed that textbooks on multivariate data analysis only scarcely mention PLS. On the contrary the technique is often mentioned in textbooks that handle spectral data, a discipline growing with high pace and that is often referred to as *chemometrics*.

The techniques used in chemometrics are often categorized as classical or inverse methods (see higher). The principal difference between these approaches is that in classical calibration the models are solved such that they are optimal in describing the measured responses (e.g., the spectral data) and can therefore be considered *optimal descriptors*, whereas in inverse methods the models are solved to be optimal in predicting the properties of interest (e.g., concentrations, *optimal predictors*). Inverse methods usually require less physical knowledge of the system, and at least in theory provide superior predictions from a mean squared error point of view. As a consequence, inverse approaches tend to be more frequently applied in multivariate calibration.

The fact that inverse methods do not require an understanding of the physical problem also poses important risks. Consider once more the example of optical spectra. Figure 1 (left) presents the average transmission spectrum together with pointwise 95 % confidence limits. Let us now alter the dataset by random permutation of the covariates, i.e. data vector $x_i = x_i(\lambda)$ is altered to

$x_i = x_i(\kappa)$ with κ being a random permutation of λ. Doing so, the natural ordering of the covariates along the corresponding wavelengths λ and, hence, spectral information, is completely destroyed. The permuted average spectrum and 95 % confidence limits are given in Figure 1 (right). However, using $x_i(\kappa)$ instead of $x_i(\lambda)$ yields exactly the same results if one would use an inverse prediction method such as PLS. One could conclude here that the inverse methods do not take advantage of the natural ordering of the covariates and are thus in a way sub-optimal. Not including the natural ordering of the covariates gives the model often too much flexibility, such that overfitting and/or lack of generality are potential threats (Saeys et al., 2008).

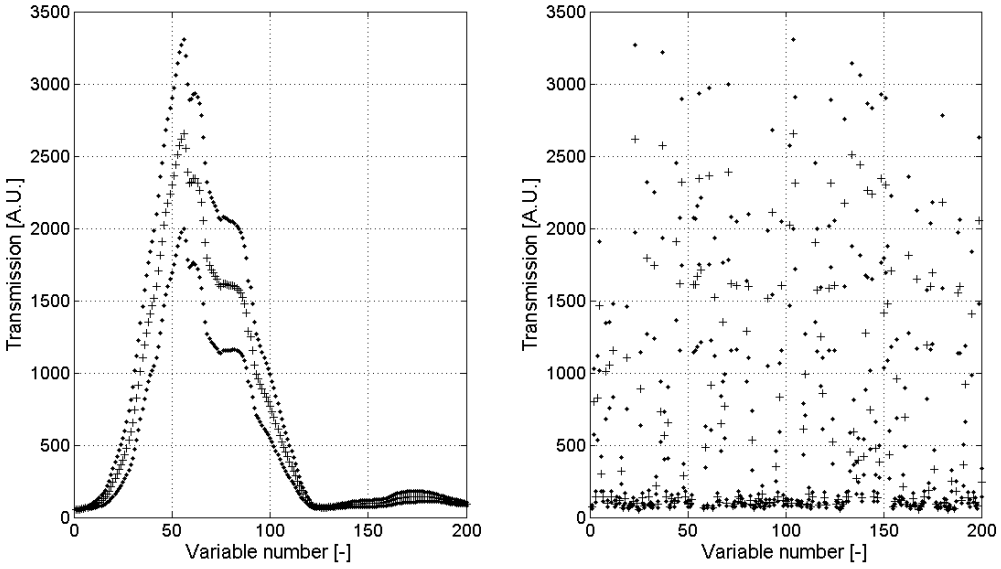

Figure 1. Visualization of the mean reflectance values (+) and their 95% confidence limits (.) in order of increasing wavelength (left) and in random order (right)
(both datasets yield exactly the same prediction accuracy when classical inverse methods are used)

Although powerful and available in most multivariate statistical software tools, inverse methods should be applied keeping in mind that overfitting is an important risk. Good statistical practice dictates that three different phases should be distinguished when using these models:

- The *calibration phase* where the parameters of the models are estimated;
- The *validation phase* to select the most appropriate model;
- The *test phase* to estimate the actual prediction error of the model

Given the importance of spectral data in modern production processes, teaching statisticians and engineers the art of model building for high dimensional data–with the different aspects mentioned above–seems indispensable. We strongly believe that *integration of physical knowledge of the measurement system or process into the algorithms used* will yield solutions with better performance.

MULTIVARIATE SPC FOR MONITORING PRODUCTION PROCESSES

In the case of spectral data mentioned above, the engineer can not only use the data for prediction purposes but may also use these data as a "fingerprint" of the current production to detect anomalies so that measures can be taken to bring it back to normal regime.

Monitoring of processes in order to detect anomalies is the subject of a field that is referred to as Statistical Process Control (SPC), although the term statistical process monitoring might be a better description since the actual control step (in an engineering sense) is not part of it. Key tools in SPC are *control charts*, first developed back in the 20's by Shewhart who concluded that while every process displays variation, some processes display controlled variation that is natural to the process (*common cause variation*), while others display uncontrolled variation that is not present in the process causal system at all times (*special cause variation*). Basically, a control chart is a graph of a measured process parameter or quality characteristic against the measurement time point. In

control charts, limits are formulated for the considered process or quality parameter based on the natural process variability. Any problem which causes unexpected process variation gives rise to the control chart crossing its limits (Montgomery, 2009).

A wide range of univariate control charts have been developed and they found their way into a broad class of industrial processes. The multivariate case, however, where a combination of variables is monitored simultaneously, is less developed and traditional approaches established a set of individual univariate control charts on each of the quality variables measured. Occasionally, also multivariate extensions of these charts based on Hotelling's T^2 statistic are employed on a subset of the most highly correlated variables.

It is surprising that for a long time multivariate control charts were not advocated by applied statistics and SPC groups. This is in sharp contrast to the current practice in the design of experiments (DOE) where statisticians have been quite successful in introducing the ideas of designed experiments in which many variables are changed simultaneously (McGregor, 1997). One can argue that the use of univariate SPC charts in multivariate situations is directly analogous to one factor at a time experimentation in DOE–the presence of variable interactions in DOE leads to the same difficulties in interpreting the results of one factor at a time experimentation as does the presence of correlation among variables in interpreting univariate SPC charts. This practice has to be avoided, and recent more research shows a rapid increase in publications that deal with SPC for high dimensional data. *In analogy with the data visualization step, also here dimension reduction techniques such as PCA are the main workhorses*. In a sense, also in the SPC context one is interested in finding those observations that show some special cause variation. Hotelling's T^2 and Q statistics together with appropriate limits can thus be used to construct these multivariate control charts. New research focusing on multivariate control charts able to detect small process changes, or capable of tracking time varying processes–an issue too often overlooked in SPC–seem to be most relevant for answering the growing needs from practice.

EFFICIENT ENGINEERING THROUGH DESIGN OF EXPERIMENTS

The SPC concepts described above are an important tool for detecting special cause variation and for keeping the process under control. Despite its clear benefits, however, statistical process control (SPC) is sometimes called a method of "counting dead bodies" (Hanrahan & Baltus, 1992): it provides little information on how to design a product or process to achieve quality goals and to maximize productivity. Designed experiments including systematic experimentation to determine the best combination of process inputs would then be a complement to SPC schemes in order to obtain high quality, stable production processes.

More than half a century ago, Box (1957) introduced the concept of EVOP–*Evolutionary Operation* as a way of systematic experimentation. The basic idea is to replace the static operation of a process by a continuous and systematic scheme of slight perturbations in the controllable variables (inputs to the process). The effect of these perturbations is evaluated and the process is shifted in the direction of improvement, a procedure that is sequentially repeated in order to keep the process at its (possibly time varying) optimum. Since only small perturbations are imposed, EVOP has as strength that it can be applied on the full scale process.

Box clearly understood the practical problem–when scaling up from small experimental conditions to full scale industrial processes, there are inevitably influences that are uncontrollable and that make the full scale process behaving sub-optimal. Although powerful classical experimental design helped to establish an optimal combination of the process inputs in the small scale (experimental) process, they are often very expensive, time consuming and require special training. Moreover, in most cases they interrupt production which is not feasible when moving to a full scale process. EVOP can thus be regarded as a tool in which a continuous investigative routine becomes the basic mode of operation for the plant and replaces normal static operation.

EVOP is based on some very basic principles and does not necessitate an advanced background in statistics. Back in the 50's, Box proposed some very simple score and analysis sheets that can be used to perform the calculations in case of 1 or 2 process variables. The practical applicability and the benefit to process engineers are huge, but surprisingly EVOP did not become a mainstream routine after Box proposed it. Reason for this might be the fact that industrial processes are often influenced by more than two variables and such cases make the manual score

sheets less appealing. Also, at that time a fast and reliable measurement of process inputs and outputs was less straightforward further complicating the technique from a practical point of view.

So many years later, there might be a revival of the original ideas proposed by Box. Not so much from a pure theoretical point of view, but from an applied statistics perspective. The same *problems* that Box described 50 years ago still exist–processes still need to be continuously checked and, whenever needed, updated, and upscaling is an important issue. However, the *possibilities* we have today with respect to the measurement of process variables and the analysis thereof using fast computers, have changed considerably. Modern processes are equipped with a vast amount of sensors measuring each and every process step with high accuracy and speed so that information rich data are available at virtually no cost. Data acquisition boards are able to merge such data and pre-process those in a format that dedicated software can be used to further analyze them. In order to do so, applied statisticians will have the challenge to fit existing design of experiments theory into the EVOP framework.

Think for instance about the progress that has been made during the last decades concerning computer aided design of experiments. These designs are specifically developed for cases where classical designs fail–mostly due to constraints imposed on the input space or the large amount of variables in the study. They rely on proven statistical methods to choose experimental points that show the influence of multiple variables with a minimum number of experimental trials. As such, they save the engineer valuable time in modeling, calculations and analysis of the results. Although software for performing computer aided designs does require some training for proper use, the best of the new programs are suitable for use by engineers so that they can solve most of the problems without resort to the services of a statistical expert. They allow the engineer to work as an engineer rather than as a statistician.

The higher mentioned constraints are typical for industrial production processes – they encompass a large number of settings (variables), some of those variables are hard-to-change, and several variable combinations are unfeasible or result in an unacceptable product(ion) quality so that they are to be avoided. However, the use of computer aided designs for improving full scale industrial processes is to our best knowledge hardly explored but deserves special attention.

CONCLUSION

Statisticians and engineers have a bright future ahead if they can turn the massive data generated using state-of-the-art sensor technologies into knowledge that enables the owner to further improve his processes. Since those data typically are labeled high dimensional, we strongly believe that multivariate data analysis in the broad sense of the word–including data exploration, model building, statistical process control as well as design of experiments–should be taught when forming the new generation of both engineers and statisticians.

REFERENCES

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*, 453-467.

Hanrahan, J. J., & Baltus, T. A. (1992). Efficient Engineering through Computer-Aided Design of Experiments. *IEEE Transactions on industry applications, 28*(2), 293-296.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd Ed). New York: Springer.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill.

McGregor, J. F. (1997). Using On-Line Process Data to Improve Quality: Challenges for Statisticians. *International Statistical Review, 65*(3), 309-323.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control* (6th Ed). Hoboken: Wiley.

Ostyn, B., Darius, P., De Baerdemaeker, J., & De Ketelaere, B. (2007). Statistical monitoring of a sealing process by means of multivariate accelerometer data. *Journal of Quality Engineering 19*, 299-310.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association, 91*, 1047–1061.

Saeys, W., De Ketelaere, B., & Darius, P. (2008). Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics, 22*(5-6), 335-344.