Model-Based Design of Experiments: Where to go?

Robert J. Flassig [*] **René Schenkendorf** [**,***]

[*] *Max Planck Institute for Dynamics of Complex Technical Systems, Process Systems Engineering, Sandtorstraße 1, 39106 Magdeburg, Germany (e-mail: flassig@mpi-magdeburg.mpg.de)*
[**] *Braunschweig University of Technology, Institute of Energy and Process Systems Engineering, Franz-Liszt-Straße 35, 38106 Braunschweig, Germany (e-mail: r.schenkendorf@tu-braunschweig.de)*
[***] *Braunschweig University of Technology, Center of Pharmaceutical Engineering (PVZ), Franz-Liszt-Straße 35a, 38106 Braunschweig, Germany*

1. INTRODUCTION

Design of experiments (DoE) is a set of well-established and over 100 years evolved rational methodologies for validating and discovering relationships between controls and responses of an input-output system in a data efficient way. The philosophy behind DoE is that controls or factors affect the system's response. The response of a system to a specific control may be observed and thus by an appropriate set of applied controls (=DoE), one may gather information of the system's mechanism to disentangle the relationship between controls and responses. Responses may comprise system states but also observables or performance measures derived from the system states.

The use of mathematical models for analysing complex processes is a powerful tool to gain a deep system understanding. However, this approach requires realistic, predictive mathematical models. During the model development phase, scientists have to cope with numerous challenges, e.g., limited knowledge about the underlying mechanisms, lack or exorbitance of dynamic or static experimental data, large experimental and process variability. Given a specific model class, a plethora of many different methodologies to optimally identify a specific model class structure have been developed since the mid of 20th century. This includes on the one hand methods for discrimination of competing structures but also methods for parameter estimation. We would like to discuss, whether further methodologies in the direction of model-based design are still needed, and if yes, to what extent. Further, given the trend of gathering massive data of a system of interest we highlight the analogy of DoE for systems identification and big data analysis. Within the age of digitalization, analysis and modelling of big data have become an active field of DoE application. Big data typically comprise massive volume, heterogeneous and unexplored data collected in areas across science (e.g. structural biology, particle physics), health (e.g. genomics, predictive healthcare), economics (e.g. market analysis), ecology, business (e.g. process monitoring), Web 2.0 sources (e.g. social media, internet of things) and robotics (e.g. sensoring data) (Fan et al., 2014). To extract information, modelling big data with empirical (statistical) or mechanistic models with classical approaches is often note feasible and thus, several approaches from design of experiments have emerged to facilitate big data modelling. Specifically, model-based DoE supplies a rational for targeted sampling in divide-and-conquer algorithms or for sequential learning, which in classical DoE is known as sequential or multi-stage DoE (Box and Draper, 1986). The classic DoE based on statistical performance measures, e.g., A-, D-, E-, I-, T-optimality, have been complemented by probabilistic model-based performance measures. These measures include global sensitivities, information-based criteria and Bayesian inference based on the posterior calculation, which have been massively studied and applied in systems biology (Schenkendorf and Mangold, 2013; Flassig and Sundmacher, 2012).

2. DOE FOR BIG DATA ANALYSIS

Over the last decade, many research and engineering disciplines have become more and more data intense. Big data have arisen from innovative experiments, measurement and monitoring devices generating high-dimensional, massive sample sizes. Big data are therefore often difficult to analyse, and the extraction of information is notoriously laborious. In a sense, a big dataset can be understood as a complex system that is yet to be identified. Thus, the goal of modelling and analysing big data is similar to what is desired in complex systems identification: (i) understanding of the interdependencies of factors and responses that shape the dataset and (ii) accurate predictions of future outcomes. As in classical systems identification, the maxim of data efficiency is given. At first sight, this requirement seems awkward. It is the high-dimensional, massive sample property that generates several interesting emergent phenomena: scalability, storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity and measurement errors (Fan et al., 2014).

During the initial rise of big data, statistics and related disciplines of data analysis have failed to adequately address big data properties and related challenges (Wang et al., 2016). The situation has changed recently when big data challenges in many different application areas have naturally driven the development of new big data method-

ologies. Even though it is easy to see that model-based DoE methodologies for complex systems identification can be tailored to cope with the emerging phenomena in big data, studies and methodologies to DoE-based big data approaches have only recently been developed (Drovandi et al., 2017).

Initially, big data methods have considered the entire dataset, and thus scalability has been the focus. Scalability has been addressed by methods including 'divide-and-conquer' approaches (Guha et al., 2012), Bayesian inference based on a consensus Monte Carlo algorithm (Huang and Gelman, 2005; Scott et al., 2016), principle component analysis (Kettaneh et al., 2005), clustering approaches (Bouveyron and Brunet-Saumard, 2014), least angle regression (Efron et al., 2004), and sparsity assumptions (Hastie et al., 2015). In contrast to using the complete dataset, DoE-based methods have been recently developed following the paradigm that a well-chosen subset of the big dataset can deliver equivalent answers compared to the full dataset at considerably less effort (Drovandi et al., 2017). As in classical DoE for systems identification or in Bayesian optimization, exploration and exploitation are the pillars of optimally analysing big data. An advantage of the DoE-based approach to big data is the avoidance of pitfalls resulting from big data effects and classical, well-established statistics can be applied. However, the DoE itself needs to be well chosen.

3. OUTLOOK: WHERE TO GO?

Uncertainty quantification, meta-modelling and big data modelling are active fields of application of DoE. Whereas uncertainty quantification has advanced its methods to efficient non-linear transformations of random variables, we still need improvements when it comes to optimizing stochastic, distributed complex systems. The optimization of systems with stochastic spatio-temporal fluctuations in combination with distributed properties is a challenging task, either from the modelling but also from the optimization point of view. The current popularity of Bayesian optimization and machine learning algorithms should be used to foster cross-disciplinary research including classic DoE; sequential design, Bayesian optimization and adaptive learning are three sides of the same coin. A coalition between researchers from classical DoE, Bayesian optimization and machine learning community in combination with applications in the areas of big data applications (e.g. process monitoring, earth science, genomics, internet of things, robotics, social media), biotechnology, pharmaceutics and systems medicine will have a bride future in terms of scientific and socio-economic impact.

DoE-based big data analysis is in the need of further research in the direction of noise accumulation and spurious patterns in high dimensional data, improvement of computational and algorithmic efficiency and stability and mastering heterogeneity, experimental variations and statistical bias associated with combining data from different sources (Fan et al., 2014).

Finally, even though model-based DoE approaches have been very much advanced over the last decades, the hard work still needs to be done: given a specific problem, scientists and engineers still have to think critically about the problem. This also includes a keen awareness of strengths and weaknesses of their chosen tools. This statement may seem trivial, however, in the time of open source libraries, out of the box solutions, nearly limitless computing power and time pressure, superficial understanding of modelling and simulation methods can be disastrous. This implicates that we as the community have to provide access and support to well-documented open source implementations, tutorials and workshops. The recent MATHMOD Minisymposium *Model-Based Design of Experiments: Where to go?* is heading in this direction bringing experts from different fields together and taking up the viewpoints of the modelling and the big data community.

REFERENCES

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71, 52–78.

Box, G.E.P. and Draper, N.R. (1986). *Empirical Model-building and Response Surface*. John Wiley & Sons, Inc., New York, NY, USA.

Drovandi, C.C., Holmes, C., McGree, J.M., Mengersen, K., Richardson, S., and Ryan, E.G. (2017). Principles of experimental design for big data analysis. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 32(3), 385–404.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J.M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J., Stine, R., Turlach, B., Weisberg, S., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2), 407–499. Cited By 3701.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293–314.

Flassig, R.J. and Sundmacher, K. (2012). Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. *Bioinformatics*, 28(23), 3089–3096.

Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J., Xi, B., and Cleveland, W.S. (2012). Large complex data: divide and recombine (d&r) with rhipe. *Stat*, 1(1), 53–67.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Huang, Z. and Gelman, A. (2005). Sampling for bayesian computation with large datasets.

Kettaneh, N., Berglund, A., and Wold, S. (2005). Pca and pls with very large data sets. *Computational Statistics & Data Analysis*, 48(1), 69–85.

Schenkendorf, R. and Mangold, M. (2013). Online model selection approach based on Unscented Kalman Filtering. *Journal of Process Control*, 23(1), 44–57. doi: 10.1016/j.jprocont.2012.10.009.

Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A., George, E.I., and McCulloch, R.E. (2016). Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2), 78–88.

Wang, C., Chen, M.H., Schifano, E., Wu, J., and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and its interface*, 9(4), 399.