ROBUST PARAMETER DESIGN FOR AUTOMATICALLY CONTROLLED SYSTEMS AND NANOSTRUCTURE SYNTHESIS

A Thesis
Presented to
The Academic Faculty

by

Tirthankar Dasgupta

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2007

ROBUST PARAMETER DESIGN FOR AUTOMATICALLY CONTROLLED SYSTEMS AND NANOSTRUCTURE SYNTHESIS

Approved by:

Dr. C. F. Jeff Wu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kwok L. Tsui
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jye-chyi Lu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Soumen Ghosh
College of Management
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 8 June 2007

To my parents,

Malay Kumar Dasgupta and Mira Dasgupta,

and my beloved wife

Koyeli,

for their support, inspiration and encouragement

during this challenging journey.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my appreciation to all who have influenced, stimulated, expedited, and warmly supported my work in various ways.

First and foremost, I would like to express my deep and sincere gratitude to my advisor, Professor C. F. Jeff Wu, for his guidance, assistance, encouragement, and hearty support at all phases of my doctoral program. He took care of me both academically and personally in every conceivable way. He has not only been my academic advisor, but also a great mentor for my graduate student life in the Uniter States.

I am extremely thankful to Dr. Roshan Joseph Vengazhiyil, who has been a friend, guide and philosopher and has rendered an overwhelming support during my studies. His constant inspiration helped me immensely at tiny moments of distress. His calm, yet determined and optimistic approach to life oozed a tranquility that helped me overcome many problems.

I would like to thank Dr. J. C. Lu, Dr. Alexander Shapiro and Dr. Brani Vidacovic for having extended their support to my research and having guided and inspired me in various ways. I am also thankful to Dr. Kwok Leung Tsui and Dr. Soumen Ghosh for serving on my dissertation committee and for their valuable comments and suggestions.

I would like to express my gratitude to Prof Rahul Mukerjee of the Indian Institute of Management and Prof Debasis Sengupta of the Indian Statistical Institute, whose inspiration and guidance helped me achieve this milestone.

I am very thankful to my lab members Dr. Abhyuday Mandal, Dr. Zhiguang Qian, Ying Hung, Xinwei Deng, Lulu Kang and Nagesh Adiga who shared time,

space and knowledge with me at Georgia Tech. Each of them helped me selflessly and contributed immensely to create the wonderful force of synergy and cohesion that prevails in our lab. I consider myself very fortunate to be able to spend a lot of time with these outstanding people.

I would like to thank all the staff members of ISyE, Georgia Tech, especially Pamela Morrison, Anita Race, Cheryl Wilkerson, Valerie DuRant-Modeste, Mark Reese, Eric Mungai and Harry Sharp, for their kind support and help at each and every stage of my graduate life at Georgia Tech.

Last, but by no means the least, my heartfelt appreciation and gratitude goes to my family, especially my parents and my beloved wife, for their constant support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Robust parameter design is an innovative and cost-effective quality engineering technique that is widely used in industries for developing new products and processes or for improving the existing ones. Application of robust parameter design to a process improves its average level of performance as well as its consistency and makes the process less sensitive to the effect of uncontrollable or noise variables. The technique needs modification and a tailor-made framework depending on the type of problem. This research focuses on developing comprehensive frameworks for developing robust parameter design methodology for dynamic systems with automatic control and for synthesis of nanostructures. The two areas, which are technologically quite different, pose plenty of unique statistical and engineering challenges in experimental planning, modeling, and optimization.

Apart from robust parameter design, a commonly used methodology for mitigating the effect of noise on the output of dynamic processes is on-line control. There are two variants of on-line control, namely feedforward and feedback control. In many processes, the optimal feedback control law depends on the parameter design solution and vice-versa. The need for an integrated approach that combines the two methods is therefore evident. A parameter design methodology in the presence of feedback control is developed for processes of long duration under the assumption that experimental noise factors are uncorrelated over time. Systems that follow a pure-gain dynamic model are considered and the best proportional-integral and minimum mean squared error control strategies are developed by using robust parameter design. The proposed method is illustrated using a simulated example and a case study in a urea

packing plant. This idea is also extended to cases with on-line noise factors, relaxing the assumption that the experimental noise factors are correlated. The possibility of integrating feedforward control with a minimum mean squared error feedback control scheme is explored.

Nanostructures, by virtue of their novel physical, chemical and biological properties, are building blocks in nanoscience and nanotechnology. To meet the needs of large scale, controlled and designed synthesis of nanostructures, it is critical to systematically find experimental conditions under which the desired nanostructures are synthesized reproducibly, at large quantity and with controlled morphology. This research focuses on an extensive application of statistical design, modeling and optimization for achieving the above goal. The problems encountered in the synthesis of nanostructures pose challenges that cannot be solved by existing experimental design techniques. Therefore, the primary objective of this research is to develop and apply novel experimental design and modeling techniques in order to find optimal and robust processing conditions for growing pure and high-quality nanostructures under time and cost constraints.

The first part of the research in this area focuses on modeling and optimization of existing experimental data. Cadmium Selenide (CdSe) has been found to exhibit one-dimensional morphologies of nanowires, nanobelts and nanosaws, often with the three morphologies being intimately intermingled within the as-deposited material. A slight change in growth condition can result in a totally different morphology. In order to identify the optimal process conditions that maximize the yield of each type of nanostructure and, at the same time, make the synthesis process robust (i.e., less sensitive) to variations of process variables around set values, a large number of trials were conducted with varying process conditions. Here, the response is a vector whose elements correspond to the numbers of appearance of different types of nanostructures. The fitted statistical models would enable nano-manufacturers to

identify the probability of transition from one nanostructure to another when changes, even of the slightest order, are made in one or more process variables. Inferential methods associated with the modeling procedure help in judging the relative impact of the process variables and their interactions on the growth of different nanostructures. Owing to the presence of internal noise, i.e., variation around the set value, each predictor variable is a random variable. Using Monte-Carlo simulations, the mean and variance of transformed probabilities are expressed as functions of the set points of the predictor variables. The mean is then maximized to find the optimum set values of the process variables, with the constraint that the variance is under control.

The second part of the research deals with development of an experimental design methodology, tailor-made to address the unique phenomena associated with nanostructure synthesis. A sequential space filling design called Sequential Minimum Energy Design (SMED) for exploring best process conditions for synthesis of nanowires. The SMED is a novel approach to generate designs that are model independent, can quickly "carve out" regions with no observable nanostructure morphology, allow for the exploration of complex response surfaces, and can be used for sequential experimentation. A unique feature of this technique lies in the fact that it originates from a combination of statistical theory and fundamental laws of physics. The basic idea has been developed into a practically implementable algorithm for deterministic functions, and guidelines for choosing the parameters of the design have been proposed. Performance of the algorithm has been studied using experimental data on nanowire synthesis as well as the modified Branin function. A modification of the algorithm based on Bayesian estimation has been proposed for random functions.

CHAPTER I

INTRODUCTION

1.1 Introduction to Robust Parameter Design

Robust Parameter Design (also known as parameter design) is a quality improvement technique proposed by Genichi Taguchi (1986,1987). Parameter design is a cost-effective approach for reducing variation in products and processes. As summarized in Nair (1992), Taguchi classifies the inputs to the system into two groups - control factors $\mathbf{x}$ and noise factors $\mathbf{z}$. The former can be easily controlled and manipulated, whereas the latter are difficult, expensive or impossible to control. Let y denote the response, which is actually some quality characteristic that measures the output of the system. Variation in $\mathbf{z}$ during the manufacturing process causes variation in y (see Figure 1). There could be many combinations or settings of $\mathbf{x}$ at which the system can produce the desired level of y (also called the target) on an average. Out of these, there will be some settings at which the system is insensitive to the effect of the noise variables $\mathbf{z}$. The basic idea of parameter design is to select the control factor settings in such a way that the performance of the system is robust, or insensitive to the noise variation $\mathbf{z}$. This is done by exploiting interactions between control factors and noise factors. Various aspects associated with the planning and implementation of robust parameter design can be found in Taguchi (1987), Phadke (1989), Nair (1992) and Wu and Hamada (2000).

1.2 Key statistical issues in robust parameter design

The major statistical issues in the context of robust parameter design can be broadly classified as follows

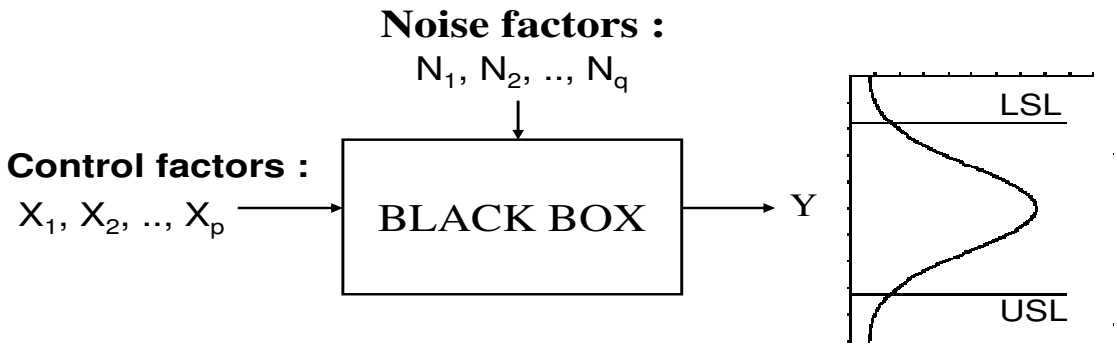


Figure 1: The black-box model for robust parameter design

1. **Performance measure and its optimization**: In the context of robust parameter design, a performance measure is a function (or summarized version) of the experimental data which needs to be optimized with respect to the control factors to obtain the most robust setting. Taguchi classified parameter design problems into different categories and defined a performance measure called "signal-to-noise" (SN) ratio, for each category. For example, for simple-target systems (see Hamada and Wu 2000, Ch. 10), where there is a fixed target T for nominal-the-best (NTB) characteristic $\mathbf{y}$, Taguchi used the SN ratio $10\log_{10} E(y^2)/var(y)$ as the appropriate performance measure. He then suggested a *two-step optimization procedure* where the first step is to maximize the SN ratio.

   Leon, Kacker and Shoemaker (1987) showed that under specific model assumptions, maximization of Taguchi's SN ratios leads to minimization of the expected quadratic loss. The SN ratios take advantage of the existence of special control variables called *adjustment factors*. When these parameters exist, use of SN ratio allows the parameter design optimization procedure to be conveniently decomposed into two smaller optimization steps, the first of which is maximization

of the SN ratio. They proposed a type of performance measure that takes advantage of adjustment parameters and is more general than Taguchi's SN ratios and called these measures "Performance Measures independent of adjustment" (PERMIA).

Different classes of engineering systems pose unique challenges . Thus, deriving appropriate performance measure is a research problem for each class of problem. This will be elaborated in the next sub-section.

2. **Modeling strategies and data analysis**: As discussed in Nair (1992), broadly there are three possible modeling strategies for parameter design experiments. One is to compute and model directly the performance measure. This is called the *performance measure modeling* (see Hamada and Wu 2000, Chap. 10, 11). A second method is to separately model the mean and log-standard deviation and combine them to optimize the performance measure. This is referred to as the *location-dispersion modeling* approach. The third method, known as the *response modeling* approach, is to model the raw data (response), including the significant control-noise interactions in the fitted model (Welch et al. 1990; Shoemaker, Tsui and Wu 1991). From this response model, the performance measure can be directly computed and optimized.

   Several alternative data analysis methods for robust design experiments have been suggested in literature. Quite a few statisticians have suggested data transforms as better alternatives to Taguchi's SN ratios (Box 1988, Nair and Pregibon 1988). Nelder and Lee (1991) recommended the use of Generalized Linear Models (GLM) for mean-variance modeling.

3. **Experimental strategy and planning techniques**: Taguchi's experimental strategy, in general, was to pick the optimum factor combination from a one-shot

experiment and then conduct a small follow-up experiment (called a confirmatory experiment). Wu, Mao and Ma (1990) investigated the deficiencies and some remedial measures for Taguchi's strategy for confirmatory experiments. Vining and Myers (1990) recommended a sequential experimental strategy and combined the robust design approach with the response surface methodology proposed by Box and Wilson (1951).

In parameter design experiments, a *control array* (called inner array by Taguchi) refers to a design matrix for the control factors and a *noise array* (called outer array by Taguchi) refers to a design matrix for the noise factors. Each level combination in the control array is crossed with all the level combinations of the noise array. The resultant array, called a *cross array* then consists of all the level combinations between those in the control array and those in the noise array.

There are two experimental formats for parameter design experiments: *cross arrays* and *single arrays*. In the latter approach, which often requires a much smaller run size, a single array for both the control and noise factors is used. However, a drawback of the method is, location-dispersion modeling or performance measure modeling cannot be employed with a single array experimental design. Wu and Zhu (2000) provided guidelines for optimal selection of single arrays in parameter design experiments.

Usually fractional factorials and orthogonal arrays are chosen for the control array and for the noise array. However, when the number of levels of control and/or noise factors is large, the run size of orthogonal array for noise factors may be prohibitively large. An alternative is to use space-filling designs like Latin hypercube designs (Koehler and Owen 1996) and uniform designs (Fang and Wang 1994) may be used. Such designs are also useful to model complex response surfaces.

1.3 Robust parameter design for different engineering systems

The problems and statistical issues associated with parameter design get more complicated as the complexity of the underlying engineering system increases. For example, there are systems in which the response takes on different values as a result of changes in a specified factor, called a *signal factor*. The signal-response relationship is of primary importance to the performance of a system. Such systems are called *signal-response systems* (see Wu and Hamada, Ch. 11), or a system with dynamic characteristics in Taguchi's terminology. Measurement systems, multiple target systems and control systems all belong to the class of signal-response systems. The parameter design strategy for such systems was originally proposed by Taguchi (1987) and have been studied by a few researchers. Leon, Shoemaker and Kacker (1987) developed PERMIA for different classes of signal-response systems. Miller and Wu (1996) developed a general framework for experimental planning techniques and modeling of data for signal-response systems. A comprehensive framework for modeling and optimization in parameter design for multiple-target systems were developed by Joseph and Wu (2002). Joseph and Wu (2002) rigorously investigated the operating window technique For systems whose outputs are evaluated on the basis of failure or defect rate. They also developed an information maximization method named Failure Amplification Method (Joseph and Wu 2004) for such systems. It is thus evident the problem of parameter design poses unique and new challenges for different types of engineering systems and applications.

1.4 Robust parameter design for dynamic systems with automatic control

Online adjustment, also called adaptive control, has been a popular strategy adopted by process engineers to reduce the variation in output due to noise factors. There are

many systems (e.g., continuous chemical processes) that cannot be made insensitive to the effect of noise by using robust parameter design, owing to strong autocorrelated noises. The use of control is inevitable for such processes. There are two types of adjustment strategies - feedforward and feedback control. In the former, a noise is measured online and a compensation provided through a suitable adjustment variable. The latter involves measurement of output at regular intervals and compensation for the effect of the uncontrollable disturbance through a controllable adjustment variable.

It may be noted that the success of robust parameter design depends on the existence of control $\times$ noise interactions. whereas the control solution does not require any such assumption and therefore has a wider applicability (Joseph 2003). However, since parameter design is a one-time strategy is usually much more economic than online control which is a continuous activity. Thus, it is not prudent to implement control systems straightaway without exploring the opportunities for robustness. In systems which require control, the opportunities of reducing variation through a parameter design strategy are often overlooked. As pointed out by Joseph (2003), a cost-effective strategy would be to use robust parameter design to make the process as robust as possible, and then use a control system to further improve its performance. However, a two-step approach for quality improvement by first using robust design and then a control system may not always work well. In fact, such an approach often yields a sub-optimal solution under certain model assumptions (See Dasgupta and Wu 2006). It is therefore necessary to develop a comprehensive framework that involves experimental planning, development of a suitable performance measure, modeling and optimization strategies for conducting parameter design experiments for systems with control. Joseph (2003) developed such a framework for systems with feedforward control. However, in his approach, the dynamics of the process was not considered. Part of the current research focusses on developing a complete framework for systems

with feedback control and the results achieved are presented in Chapter II.

1.5 Robust Parameter Design for Synthesis of Nanostructures

Nanotechnology is the construction and use of functional structures designed from atomic or molecular scale with at least one characteristic dimension measured in nanometers (one nanometer = 10^{-9} meter, which is about 1/50,000 of the width of human hair). The size of these nanostructures allows them to exhibit novel and significantly improved physical, chemical, and biological properties, phenomena, and processes. Nanotechnology can provide unprecedented understanding about materials and devices and is likely to impact many fields. By using structure at nanoscale as a tunable physical variable, scientists can greatly expand the range of performance of existing chemicals and materials. Alignment of linear molecules in an ordered array on a substrate surface (self-assembled monolayers) can function as a new generation of chemical and biological sensors. Switching devices and functional units at nanoscale can improve computer storage and operation capacity by a factor of a million. Entirely new biological sensors facilitate early diagnostics and disease prevention of cancers. Nanostructured ceramics and metals have greatly improved mechanical properties, both in ductility and strength.

To meet the needs of large scale, controlled and designed synthesis of nanowires, it is critical to systematically find experimental conditions under which the desired nanostructures are synthesized reproducibly, at large quantity and with controlled morphology. Of these, reproducibility is the biggest issue. Owing to a large number of uncontrollable factors affecting the synthesis process, drastic change of results are observed even under no apparent changes in process conditions. Therefore, there are tremendous opportunities for improvement by using robust parameter design.

However, there are challenges that would require innovative statistical methods for design and analysis of parameter design experiments in this area. They are as

follows:

1. Intermingled nanostructures.

2. Nominal and categorical responses.

3. Strong internal noises, i.e., variation of control variables around their set values.

4. Complete disappearance of morphology with slight changes in process conditions.

5. Complex response surfaces making exploration of optimal settings very difficult.

Therefore, the primary objective is to develop and apply novel experimental design and modeling techniques in order to find optimal and robust processing conditions for growing pure and high-quality nanostructures under time and cost constraints. In Chapter III, a new modeling and optimization strategy is proposed. In Chapter IV, an experimental design methodology called Sequential Minimum Energy Designs (SMED), tailor-made to address the unique phenomena associated with nanostructure synthesis is developed.

1.6 References

Box, G.E.P. (1988), "Signal-to-Noise Ratios, Performance Criteria, and Transformations," *Technometrics*, 30, 1–17.

Fang, K.T. and Wang, Y. (1994), *Number-Theoretic Methods in Statistics*, London: Chapman and Hall.

Joseph, V.R. (2003), "Robust Parameter Design With Feed-Forward Control," *Technometrics*, 45, 284-291.

Joseph, V.R. and Wu, C.F.J. (2002), "Robust Parameter Design of Multiple Target Systems," *Technometrics*, 44, 338-346.

Joseph, V.R. and Wu, C.F.J. (2002), "Operating Window Experiments: A Novel Approach to Quality Improvement," *Journal of Quality Technology* 34, 345-354.

Joseph, V.R. and Wu, C.F.J. (2004), "Failure Amplification Method: An Information Maximization Approach to Categorical Response Optimization," (with discussion), *Technometrics* 46, 1-31.

Koehler, J. and Owen, A. (1996), "Computer Experiments," in *Handbook of Statistics: Design and Analysis of Experiments*, (Ed. C.R. Rao), Amsterdam: Elsevier Science B.V., pp. 261–308.

León, R.V., Shoemaker, A.C., and Kacker, R.N. (1987), "Performance Measures Independent of Adjustment," *Technometrics*, 29, 253–285.

Miller, A. and Wu, C.F.J. (1996), "Parameter Design for Signal-Response Systems: A Different Look at Taguchi's Dynamic Parameter Design" *Statistical Science*, 11, 122–136.

Nair, V.N. (Ed.) (1992), "Taguchi's Parameter Design: A Panel Discussion," *Technometrics*, 34, 127–161.

Nair, V.N. and Pregibon, D. (1988), "Analyzing Dispersion Effects from Replicated Factorial Experiments," *Technometrics*, 30, 247–257.

Nelder, J.A. and Lee, Y. (1991), "Generalized Linear Models for the Analysis of Taguchi-type Experiments," *Applied Stochastic Models and Data Analysis*, 7, 107–120.

Phadke, M.S. (1989), *Quality Engineering Using Robust Design*, Engelwood Cliffs, NJ: Prentice Hall.

Shoemaker, A.C., Tsui, K.L., and Wu, C.F.J. (1991), "Economical Experimentation Methods for Robust Design," *Technometrics*, 33, 415–427.

Taguchi, G. (1986), *Introduction to Quality Engineering*, Tokyo, Japan: Asian Productivity Organization.

Taguchi, G. (1987), *System of Experimental Design*, White Plains, NY: Unipub/Kraus International Publications.

Vining, G.G. and Myers, R.H. (1990), "Combining Taguchi and Response Surface Philosophies: A Dual Response Approach," *Journal of Quality Technology*, 22, 38–45.

Welch, W.J., Yu, T.K., Kang, S.M., and Sacks, J. (1990), "Computer Experiments for Quality Control by Parameter Design," *Journal of Quality Technology*, 22, 15–22.

Wu, C.F.J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Wu, C.F.J. and Zhu, Y. (2000), "Optimal Selection of Single Arrays in Parameter Design Experiments," Technical Report, Department of Statistics, University of Michigan.

CHAPTER II

ROBUST PARAMETER DESIGN WITH FEEDBACK CONTROL

2.1 Introduction

There are many processes of long duration (e.g., continuous chemical processes) that cannot be made insensitive to the effect of noise by using robust parameter design (or, briefly, parameter design). The use of control is inevitable in these situations. Feedback control involves measurement of the output at regular intervals and compensation for the effect of the uncontrollable disturbance through a controllable process parameter. To understand the combined role of parameter design and feedback control in reducing process variation, consider a simple model

$$Y_t = -2 + 2x - N_t + 0.5xN_t + 2C_{t-1} + z_t, \tag{1}$$

where x is a control factor that is not changed during production, N is a noise factor and C is a control factor that is adjusted to compensate for the unobservable disturbance z. Changing C by one unit at time $t-1$ produces a 2 units change in Y at time point t. We assume that N and z are random variables with mean 0 and variance 1. Suppose the target value of Y is 10. Clearly, if we can set $x = 4$ and $C = 2$, then the target is achieved on average, and we have $Var(Y) = 2$. Instead, if we set x to 2 and C to 4, then the effect of N on Y is removed, and the target is still achieved with a much lower variance of 1.

Now suppose that instead of being a white noise process, z_t is a non-stationary disturbance which makes the output Y unstable. In such a case, one can set C to an initial value $C_0 = 4$, and keep on adjusting C_t with a view to compensate for the disturbance z, make the process stable and consequently minimize the variation of Y

around the target. In actual practice, this can be achieved by obtaining a forecast of z_t at time point $t-1$ from the past observations and adjusting C based on such a forecast.

Thus, there are two objectives to be fulfilled - one is to find a robust setting for x, and the other is to find the optimum control law. One way of achieving this may be to fix C at a certain level and conduct a parameter design experiment to find the optimum setting of x, and then fix x at its optimal level and determine the optimal control law. This is referred to as a *two-stage approach*. Such an approach, though not found in the robust design or the control theory literature, is not very difficult to implement. However, such a two-stage approach for quality improvement may not always work well. For example, if we have a model of the form

$$Y_t = -2 + 2x - N_t + 0.5xN_t + (2 - 0.75x)C_{t-1} + z_t, \tag{2}$$

where z_t is an autoregressive process of order 1 such that $z_t = a_t + \phi z_{t-1}$ and $var(a_t) = \sigma^2(x) = (1-0.5x)^2$, then obviously the choice of x would have an impact on the effect of N on Y (i.e., robustness of the process) as well as on the control law. Thus the control law would depend on the parameter design solution and vice-versa and a two-stage approach may yield a sub-optimal solution.

Joseph (2003) developed a general parameter design methodology for systems with feedforward control. In this chapter, we propose an integrated approach to conduct a parameter design experiment for systems with feedback control. In Section 2.2, we describe an industrial scenario as a motivating example. In Section 2.3 we give an overview of some common process inertia models and feedback control schemes. In Section 2.4, a framework for parameter design with feedback control is proposed for a specific class of process inertia models (pure gain) and the discrete proportional-integral (PI) control scheme. In this Section, we also define the two-stage approach and compare it with the proposed single-stage approach. In Section 2.5 the proposed methodology is demonstrated through a simulation experiment. In Section 2.6 we

discuss the extension of our proposed framework to minimum mean squared error (MMSE) feedback control scheme. Section 2.7 illustrates the proposed approach with an example from a packing plant. Section 2.8 contains concluding remarks and future research directions.

2.2 Motivating Example

As a motivating example we consider the packing experiment described by Dasgupta, Sarkar and Tamankar (2002). The paper describes an automated packing process in which the input material flows into the machine from a hopper. The target weight can be pre-set. There are several control factors $\mathbf{X}$, which are set at the beginning of production and usually not altered.

Let Y denote the response (weight of packed bag) and T denote the target weight. When a bag is packed, the material flows into the bag in two stages, viz. main (coarse) feed stage (when the material flows into the bag thick and fast) and dribble (fine) feed stage (when the material just trickles down into the bag). In-flight material compensation C determines how early the main feed will be cut off. The main-feed cut-off value is $T - (C+$Dribble feed quantity). For example, if C is set to zero, and the target weight is 50 lb, and dribble feed quantity = 12 lb, then the main feed will be cut-off at (50-12)=38 lb. But after the main feed is cut off, there will still be some material flow, which will result in Y being greater than 50 lb. If C is now increased to 1 lb, then the main feed will be cut off at (50-12-1) = 37 lb, and Y will consequently be reduced. C is therefore used as an on-line adjustment parameter to compensate for the effect of noise. The noise is strong and is a manifestation of a multitude of small effects, none of which can be measured individually. However, an off-line noise factor that can be controlled to some extent for experimental purposes is the material composition (course/fine/lumpy).

Among the set of control factors $\mathbf{X}$, some are likely to interact with noise and/or

with C. Further, the variance of the unobservable noise is also expected to depend on some of the control factors. This is thus a case of robust parameter design with feedback control. The actual experiment and analysis of experimental data will be discussed in Section 2.7.

2.3 Feedback control schemes, models for process inertia and role of DOE

2.3.1 Feedback control schemes and process inertia

Suppose the response Y has a target T. Corresponding to time t, let Y_t denote the value of Y, $e_t = Y_t - T$ denote the deviation of the response from the target and C_t denote the value of the adjustment factor. Further, assume that Y and C are linked by the following transfer function

$$Y_t = \beta(Y_{t-1}, Y_{t-1}, \ldots, C_t, C_{t-1}, \ldots) + z_t, \tag{3}$$

where z_t is the unobservable disturbance. In any feedback control scheme, a correction is given to C_t on the basis of the observed output error e_t through a control equation $C_t = f(e_t, e_{t-1}, \ldots)$.

There is a vast literature on feedback control schemes (e.g., Astrom (1970); Davis and Vinter (1985); Box, Jenkins and Reinsel (1994); Seborg, Edgar and Mellichamp (1989); Del Castillo (2002)). Among various control schemes, the discrete proportional-integral (PI) control schemes have received particular attention because of their simple structure and ease of implementation. In a discrete PI control scheme, the control equation is of the form

$$-C_t = k_0 + k_p e_t + k_I \sum_{i=1}^{t} e_i, \tag{4}$$

where k_p and k_I are positive constants that determine the amount of proportional and integral control. In the example cited in Section 2.2, the controller is a special case of the discrete PI controller with $k_p = 0$ (integral control).

Another commonly used feedback control scheme is the minimum mean squared error (MMSE) scheme. Under certain model assumptions and choice of parameters, the discrete PI control scheme and MMSE schemes can be shown to be equivalent (Box and Luceno 1997). However, in general, the PI schemes are seen to be quite efficient over a broad range of the parameter space. Furthermore, as shown by Tsung, Wu and Nair (1998), the PI schemes are more robust to model misspecification than MMSE schemes.

The transfer function in (3) can also be of various types. A simple first-order dynamic model that characterizes many processes of practical interest is given by the following equation

$$Y_t = \alpha + \delta Y_{t-1} + g(1-\delta)C_{t-1} + z_t, \tag{5}$$

where $0 < \delta < 1$

A further simplicification of (5) can be achieved by assuming that essentially all the change induced by C will occur in a single time interval. This corresponds to setting $\delta = 0$ in (5), i.e.,

$$Y_t = \alpha + gC_{t-1} + z_t. \tag{6}$$

This is called the *pure-gain model*. Box & Kramer (1992) considered primarily the pure gain model in their discussion on feedback control.

In Section 2.4, while developing a framework for robust parameter design with feedback control, we shall restrict attention to the pure-gain dynamics and the integral control scheme.

2.3.2 Choice of control scheme parameters and role of DOE

It is clear that under the discrete PI control scheme, the control can be poor or unstable if the constant k_I is incorrectly chosen. One way of selection of k_I is to study the nature of the underlying time series model for z_t and use this information

for optimum selection of k. For example, if z_t is an ARIMA(0,1,1) process with parameter λ then under model (5), $k_p = 0$ and $k_I = \lambda/g$ result in minimum output variation (Box, Jenkins and Reinsel (1994), Chapter 13).

Suppose a controller has been hooked up to a system and is approximately of right design but is mistuned. One may tune it by formally identifying and fitting models for the process disturbance and dynamics. However, such an approach may be too tedious for routine use. Different experimental approaches for tuning of controller parameters have found place in control theory and chemical engineering literature. These methods were originally based on the trial and error approach (Ziegler and Nichols (1942)), but later, methods based on a single experiment were proposed (e.g., Cohen and Coon (1953), Yuwana and Soberg (1982)). However, it was felt that a sequential approach would be more appropriate for exploring the optimal values of the controller parameters (Carpenter and Sweeny (1965)). As Box and Kramer (1992) point out, to avoid upsetting the system, experimental runs may be made in the evolutionary operation mode, and the response surface methodology may be used to explore and optimize the important factors. Nakano and Jutan (1994) first used the response surface methodology with integral of the squared error (ISE) as the objective function to tune PI controllers. This idea was extended to track dynamic optima by Edwards and Jutan (1997) and Jiang and Jutan (2000).

However, in case a controller has to be set up from scratch, or we have a controller whose basic design is inappropriate, one has to design and conduct a more elaborate experiment to identify the appropriate models for process disturbance and dynamics. We shall consider both these situations in our proposed framework described in the following Section.

2.4 A framework for robust design with first-order pure-gain dynamic models and discrete PI control scheme

2.4.1 Framework and statistical model

Figure 2 depicts a model for feedback control in the presence of control and noise factors. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ denote the set of control factors that can only be changed at the process set-up. Let $\mathbf{U} = (U_1, U_2, \ldots, U_q)'$ denote the set of uncontrollable noise factors. Components of $\mathbf{U}$ interact with the components of $\mathbf{X}$. We can write $\mathbf{U} = \{\mathbf{N}, \mathbf{Z}\}$, where $\mathbf{N}$ denotes the set of noise factors that can be deliberately varied during the experiment and $\mathbf{Z}$ denotes the set of remaining noise factors that cannot be identified or controlled during experimentation. To develop the framework, we assume that all the components of $\mathbf{N}$ are white noise. We also assume that none of the noise factors are measurable online during production.

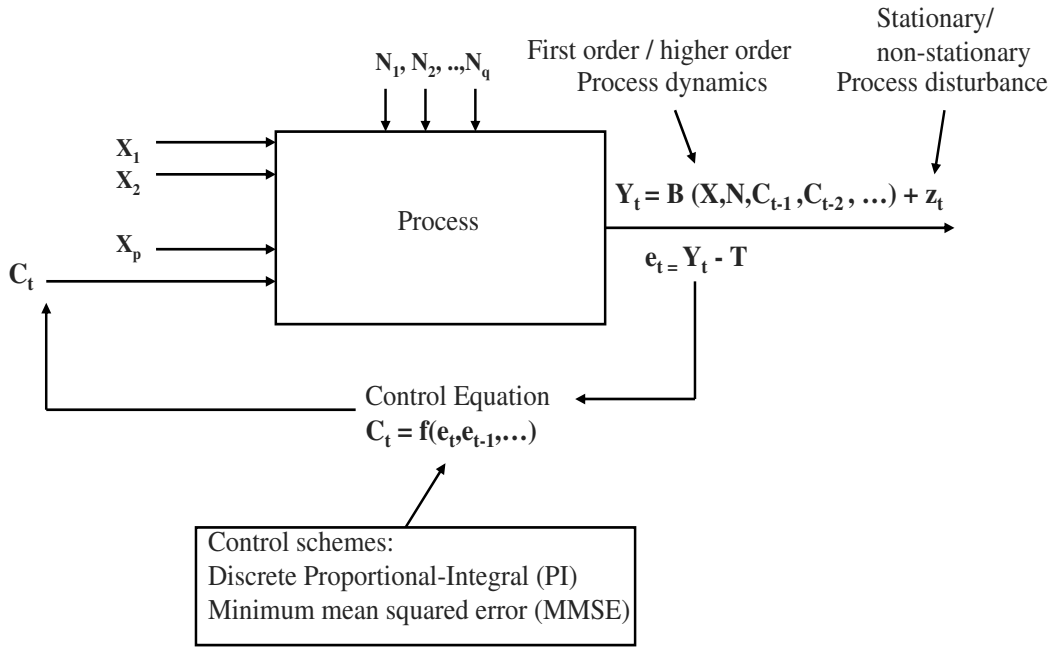


Figure 2: Feedback Control With Control and Noise Factors

In addition, we have a control factor C that is adjusted *on-line* during production. This control factor, also called *adjustment* factor in robust design literature (Wu and Hamada 2000, Chapter 10), is such that it affects the mean of the response but not

its variance. In other words, it does not interact with any of the noise factors.

The response Y and control factor C are linked by a transfer function of the form

$$Y_t = \beta(\mathbf{X}, \mathbf{N}, \mathbf{Z}, C_{t-1}, C_{t-2}, \ldots),$$

which can be rewritten as

$$Y_t = \beta(\mathbf{X}, \mathbf{N}, C_{t-1}, C_{t-2}, \ldots) + z_t,$$

where $\{z_t\}$ is the disturbance due to unobservable and uncontrollable noise factors $\mathbf{Z}$ and may be stationary or non-stationary. Since $\mathbf{Z}$ interacts with $\mathbf{X}$ but not with C (by definition of adjustment factor), the variance of z_t may be assumed to depend on $\mathbf{X}$, but not on C.

At time t, a correction is given to C_t on the basis of the observed output error e_t through a control equation $C_t = f(e_t, e_{t-1}, \ldots)$. As discussed in Section 2.3.1, we shall consider the following forms for the functions β and $f(e_t, e_{t-1}, \ldots)$.

$$\begin{aligned} \beta(\mathbf{X}, \mathbf{N}, C_{t-1}, C_{t-2}, \ldots) &= \alpha_0(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_{t-1}, \\ f(e_t, e_{t-1}, \ldots) &= -k_0 - k_I \sum_{i=1}^{t} e_i. \end{aligned}$$

We thus postulate the following first-order pure-gain dynamic model

$$Y_t = \alpha_0(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_{t-1} + z_t. \tag{7}$$

If T denotes the target, and $e_t = Y_t - T$ denotes the deviation from the target, then,

$$e_t = \alpha(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_{t-1} + z_t, \tag{8}$$

where $\alpha(\mathbf{X}, \mathbf{N}) = \alpha_0(\mathbf{X}, \mathbf{N}) - T$.

Note that this is essentially the same as (6), the added aspect being the dependence of the dynamics on $\mathbf{X}$ and $\mathbf{N}$. It is thus imperative that, if an integral control scheme is employed for such a process, k_I necessarily has to be a function of $\mathbf{X}$ and $\mathbf{X}$ has to be such that the output is least sensitive to the effect of $\mathbf{N}$.

2.4.2 Performance measure and its optimization

Let us assume that $z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$, where $\{a_t\}$ is a white noise process with zero mean and variance $\sigma^2(\mathbf{X})$. Recalling that in an integral control scheme C_{t-1} is set to $-k_0 - k_I \sum_{i=1}^{t-1} e_i$, we have from (8),

$$\begin{aligned} e_t &= \alpha(\mathbf{X},\mathbf{N}) - g(\mathbf{X})\Big(k_0 + k_I \sum_{i=1}^{t-1} e_i\Big) + z_t \\ &= \Big(\alpha(\mathbf{X},\mathbf{N}) - g(\mathbf{X})k_0\Big) - g(\mathbf{X})k_I \sum_{i=1}^{t-1} e_i + \sum_{j=0}^{\infty} \psi_j a_{t-j}. \end{aligned} \tag{9}$$

Clearly, the objective is to select $\mathbf{X}$ and the control law in such a way that the variance of e_t is minimized. For the type of processes considered here, the duration of the control session will be long. Thus, it is reasonable to consider $PM(\mathbf{X}, k_I) = Var(e_t)$ as the appropriate performance measure, provided the output e_t is asymptotically stable with mean zero.

We have

$$\begin{aligned} Var(e_t) &= Var_a E_{\mathbf{N}}(e_t|a) + E_a Var_{\mathbf{N}}(e_t|a) \\ &= Var(u_t) + \pi(\mathbf{X}), \end{aligned} \tag{10}$$

where

$$\begin{aligned} u_t &= E_{\mathbf{N}}(e_t|a) \\ &= \Big(\alpha(\mathbf{X}) - g(\mathbf{X})k_0\Big) - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + \sum_{j=0}^{\infty} \psi_j a_{t-j} \end{aligned}$$

and $\alpha(\mathbf{X}) = E_{\mathbf{N}}(\alpha(\mathbf{X},\mathbf{N})), \quad \pi(\mathbf{X}) = Var_{\mathbf{N}}(\alpha(\mathbf{X},\mathbf{N}))$.

It is seen that if model (8) and the distribution of $\mathbf{N}$ is known, then we can compute the performance measure $PM(\mathbf{X}, k_I)$ by substituting $Var(u_t)$ in (10). For a given process disturbance model (i.e., given the weights ψ_j) $Var(u_t)$ can be obtained through routine but tedious derivations (Tsung, Wu and Nair, 1996). $Var(u_t)$ will be a function of $\mathbf{X}$ and k_I, and will be of the form $V(\mathbf{X}, k_I)\sigma^2(\mathbf{X})$. Clearly, $V(\mathbf{X}, k_I)$

depends on $\mathbf{X}$ through the function $g(\mathbf{X})$. For example, if z_t is an ARIMA(0,1,1) process with parameter λ, one can obtain the following expression for $V(\mathbf{X}, k_I)$ by slightly extending the result of Box and Kramer (1992),

$$\begin{aligned} V(\mathbf{X}, k_I) &= \frac{1+\theta^2-2\phi(\mathbf{X},k_I)\theta}{1-\phi^2(\mathbf{X},k_I)}, \quad \text{if} -1 < \phi(\mathbf{X},k_I) < 1, \\ &= \infty \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise,} \end{aligned} \tag{11}$$

where $\phi(\mathbf{X}, k_I) = 1 - g(\mathbf{X})k_I$ and $\theta = 1 - \lambda$.

We thus have

$$PM(\mathbf{X}, k_I) = V(\mathbf{X}, k_I)\sigma^2(\mathbf{X}) + \pi(\mathbf{X}). \tag{12}$$

Stability of the output e_t can be ensured by choosing k_I within appropriate stability region $\kappa(\mathbf{X})$, which depends on the nature of the disturbance z_t.

From (8), we can write for $t = 1$,

$$e_1 = \alpha(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_0 + z_1.$$

Noting that the starting value of C_0 is k_0, we have

$$E(e_1) = \alpha(\mathbf{X}) + g(\mathbf{X})k_0.$$

Thus, to ensure that the output is asymptotically stable around zero, we must have $k_0 = -\alpha(\mathbf{X})/g(\mathbf{X})$ and $k_I \in \kappa(\mathbf{X})$.

Next, the following *two-step optimization* is performed :

1. Minimize $PM(\mathbf{X}, k_I)$ subject to $\mathbf{X} \in [\mathbf{X^L}, \mathbf{X^U}]$ and $k_I \in \kappa(\mathbf{X})$, where $[\mathbf{X^L}, \mathbf{X^U}]$ denotes the experimental range for the control variables. Let $\mathbf{X}^*$ and k_I^* denote optimum $\mathbf{X}$ and k_I respectively.

2. Obtain the optimal k_0 as $k_0^* = -\alpha(\mathbf{X}^*)/g(\mathbf{X}^*)$.

To illustrate the computation of the performance measure, let us consider the following example :

$$e_t = -2x - N + 0.5xN + (2 - 0.75x)C_{t-1} + z_t.$$

We assume that $z_t = \phi z_{t-1} + a_t$ is an AR(1) process with $\phi = 0.7$, $E(a_t) = 0$, $Var(a_t) = (2 - 0.8x)^2$ and N is a random variable with mean 0 and variance 1.

Extending the results of Tsung, Wu and Nair (1996) to the current model, it can easily be seen that the stability region here is $\kappa(\mathbf{X}) = \{k_I : |1 - g(\mathbf{X})| \leq 1\}$. Also, an appropriate expression for $V(\mathbf{X}, k_I)$ under this model is given by

$$\begin{aligned} V(\mathbf{X}, k_I) &= \frac{2}{(1 + \eta(\mathbf{X}))(1 - \phi\eta(\mathbf{X}))(1 + \phi)}, \quad \text{where} \\ \eta(\mathbf{X}) &= 1 - g(\mathbf{X})k_I. \end{aligned} \tag{13}$$

Substituting appropriate expressions in (13) and (12), we get

$$PM(\mathbf{X}, k_I) = \frac{1.1765(2 - 0.8x)^2}{\Big(1 + \{1 - k_I(2 - 0.75x)\}\Big)\Big(1 - 0.7\{1 - k_I(2 - 0.75x)\}\Big)} + (1 - 0.5x)^2.$$

If the experimental region is $0 \leq x \leq 5$, the problem is to

$$\text{minimize } PM(\mathbf{X}, k_I) \text{ subject to}$$

$$0 \leq x \leq 5,$$

$$k_I \in \{k_I : |1 - k_I(2 - 0.75x)| < 1\}.$$

Using a simple non-linear constrained optimization function from MATLAB, we find that the minimum is obtained at $x^* = 2.2990$, $k_I^* = 2.8497$ and the corresponding value of the performance measure is $PM(x^*, k_I^*) = 0.0374$. Assuming $E(N) = 0$ and $var(N) = 1$, and substituting $x^* = 2.299$, we obtain the optimal k_0 as $k_0^* = -\alpha(\mathbf{X}^*)/g(\mathbf{X}^*) = 16.6745$.

The performance of the system without control may be evaluated by substituting $k_I = 0$ in the expression for performance measure, which gives $PM(\mathbf{X}, 0) = \frac{(2-0.8x)^2}{(1-\phi^2)} + (1-0.5x)^2$. Here $PM(\mathbf{X}^*, 0) = 0.0730 > PM(\mathbf{X}^*, k_I^*)$, demonstrating that the control will be effective.

2.4.3 Comparison with a two-stage approach

The interesting aspect of the proposed approach is that, the optimal settings of the control variables $\mathbf{X}$ and the controller parameters k_0 and k_I are obtained by conducting a single experiment. Let us now compare this *single-stage* approach with a two-stage approach where at the first stage robust level of $\mathbf{X}$ is decided by conducting a traditional parameter design experiment and next the optimal control law is obtained.

From (8), we have,

$$\begin{aligned} e_1 &= \alpha(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_0 + z_1 \\ &= \alpha(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_0 + a_1, \quad [\text{assuming } \psi_0 = 1.] \end{aligned}$$

Setting $C_0 = C_0^*$, one can perform a traditional parameter design experiment to find out the optimum setting of $\mathbf{X}$ by minimizing $Var(e_1) = \pi(\mathbf{X}) + \sigma^2(\mathbf{X})$. Here, unlike traditional robust parameter design, the second-step of optimization (adjusting the mean to target by proper choice of $\mathbf{X}$) will not be done, since C is considered to be the only adjustment factor here. Let $\widetilde{\mathbf{X}}$ denote the optimum $\mathbf{X}$ obtained this way.

The second stage of the optimization consists of obtaining the optimal k_I for the chosen setting $\mathbf{X} = \widetilde{\mathbf{X}}$. This is equivalent to obtaining

$$\widetilde{k_I} = \arg \min_{k_I \in \kappa(\widetilde{\mathbf{X}}), \mathbf{X} = \widetilde{\mathbf{X}}} Var(e_t) = \arg \min_{k_I \in \kappa(\widetilde{\mathbf{X}})} V(\widetilde{\mathbf{X}}, k_I).$$

If $\mathbf{X}^*$ and k_I^* denote the optimum $\mathbf{X}$ and k_I obtained using the proposed single stage approach, then

$$\begin{aligned} PM(\mathbf{X}^*, k_I^*) &= \min_{\mathbf{X} \in (X^L, X^U), k_I \in \kappa(\mathbf{X})} V(\mathbf{X}, k_I)\sigma^2(\mathbf{X}) + \pi(\mathbf{X}) \\ &\leq PM(\widetilde{\mathbf{X}}, \widetilde{k_I}). \end{aligned}$$

Thus, in general, the asymptotic variance of the output obtained using the proposed single-stage approach will be less than or equal to that for the two-stage approach. To demonstrate this numerically, consider the example in Section 2.4.2. Here,

the two-stage approach would consist of the following :

- Stage I :

$$\begin{aligned}\widetilde{x} &= \arg\min_{0<x<5}\left(\sigma^2(x)+\pi(x)\right)\\ &= \arg\min_{0<x<5}\left((1-0.5x)^2+(2-0.8x)^2\right)=2.3596.\end{aligned}$$

- Stage II

$$\widetilde{k_I}=\arg\min_{k_I\in\kappa(\widetilde{x})} V(\widetilde{x},k_I)=3.4117.$$

The proposed single-stage approach would yield the following result :

$$(x^*,k_I^*)=\arg\min_{x\in(0,5),k_I\in\kappa(x)} V(x,k_I)\sigma^2(x)+\pi(x)=(2.2990,2.8497).$$

We have, $0.0397 = PM(\widetilde{x},\widetilde{k_I}) > PM(x^*,k_I^*) = 0.0374$, which indicates that the two-stage approach yields a sub-optimal solution.

However, it may be noted that if $\min_{k_I\in\kappa(\mathbf{X})} V(\mathbf{X},k_I)=1$ for all $\mathbf{X}$, then the two procedures would lead to identical results. Physically this means that through the control mechanism, it is possible to reduce the variation of the output to the variance of the underlying white noise. This will hold good for the ARIMA(0,1,1) model, where by choosing $k_I=\frac{\lambda}{g(\mathbf{X})}$, $V(\mathbf{X},k_I)$ can be reduced to 1.

Another drawback of the two-stage approach is the improper selection of k_0, since it will be impossible to distinguish $\alpha(\mathbf{X},\mathbf{N})$ from $g(\mathbf{X})$ in the first-stage model obtained by keeping C fixed. This will give rise to instability in the process during the initial phase of production.

2.4.4 Design of experiments and analysis of data

In order to express PM as a suitable function of $\mathbf{X}$ and k_I, one may conduct a suitable open-loop experiment with $\mathbf{X},\mathbf{N}$ and C as experimental factors to estimate model (8), fit an appropriate time series model for z_t and then obtain an expression

for PM by considering an appropriate distribution of $\mathbf{N}$. This approach is known as *response modeling* (Wu and Hamada 2000, Chaps 10 and 11). An alternative procedure is to directly model PM as a function of $\mathbf{X}$ and k_I by treating k_I as an experimental factor and conducting the experiment with the control loop. This is called *performance measure modeling*. In the following two subsections, we discuss the design and analysis of experiments under these two approaches.

2.4.4.1 Response modeling approach

Recall that in Section 2.4.2, we had assumed the dependence of σ_a^2 (the variance of a_t) on $\mathbf{X}$. Thus, the response modeling may be thought of as a two-step approach:

1. Fitting a transfer function of the form $e_t = \alpha + gC_{t-1} + z_t$ for various combinations of $\mathbf{X}$ and $\mathbf{N}$.

2. Modelling α as a function of $(\mathbf{X}, \mathbf{N})$; g and σ_a^2 as functions of $\mathbf{X}$.

To achieve this objective, we may use a cross array design between $\mathbf{X}$ and $\mathbf{N}$ and nest all the levels of C within each $\mathbf{X}, \mathbf{N}$ combination. Thus, for each combination of $\mathbf{X}$ and $\mathbf{N}$, a time series in Y_t and hence e_t will be obtained by changing the levels of C. As in the simulated experiment described on p. 442 of Box et al. (1994), each level of C may be held constant for a fixed time, and τ observations may be generated. Instead of employing a cross array design, we may use a single array as well, ensuring that all the interactions between $\mathbf{X}$ and $\mathbf{N}$ are estimable. Details of cross array and optimal single array designs may be found in Chapter 10 of Wu and Hamada (2000).

As indicated above, the analysis consists of the following two broad stages:

1. For each combination of $\mathbf{X}$ and $\mathbf{N}$,

   (a) Identification of the form of the disturbance $z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$ using autocorrelation function (ACF) plots and partial autocorrelation function (PACF) plots;

(b) Estimating the parameters (α, g, ψ_j's, σ_a^2) using a constrained iterative nonlinear least squares algorithm (Box et al. 1994, Chapter 7).

2. Treating α, g and σ_a^2 as three different responses, we identify significant control factors and control×noise interactions and fit $\widehat{\alpha} = \alpha(\mathbf{X}, \mathbf{N})$, $\widehat{g} = g(\mathbf{X})$, $\widehat{\sigma_a^2} = \sigma_a^2(\mathbf{X})$.

Note that if we assume that σ_a^2 does not depend on $\mathbf{X}$, i.e., all the parameters associated with the disturbance z_t are free of $\mathbf{X}$, then the experiment can be considerably simplified. In such a case we may estimate the time series parameters from a single experimental run for fixed $\mathbf{X} = \mathbf{X}^*$ and $\mathbf{N} = \mathbf{N}^*$. Next, we may conduct a cross array design $D(\mathbf{X}) \bigotimes D(\mathbf{N}) \bigotimes D(C)$ and estimate model (8) directly from the experimental data.

Although the response modeling approach provides an in-depth understanding of the underlying phenomena, it is clear that there is a possibility that the experiment will be very large and will involve intensive computation. Further, it is obvious, that this experiment has to be run with an open loop. For systems in which controllers have already been installed, industrial personnel would usually be reluctant to run open-loop experiments. Thus, when the objective is to achieve robustness of a system that already has a feedback controller, the performance measure modeling approach discussed in the following section will be appropriate.

2.4.4.2 Performance measure modeling

Since this approach can be thought of as modeling the performance measure as a function of $\mathbf{X}$ and k_I, it would be reasonable to use a cross array design between $\mathbf{X}$ and k_I. The various noise combinations may be nested within each $(\mathbf{X}, k_I)$ combination. Thus, for each combination of $\mathbf{X}$ and k_I, a time series in Y_t and hence e_t will be obtained by changing the levels of $\mathbf{N}$. In order to reduce the run size, $\mathbf{X}$ and k_I can be accommodated in a single array. In this case, one must ensure that all or most of

the interactions between k_I and $\mathbf{X}$ are estimable.

Note that, in this set-up, the adjustment factor C need not be included in the experiment as it will be automatically changed during the course of this closed loop experiment. However, k_0 corresponds to the initial value of C (at time 0) and may be treated as 'another' control factor.

Selection of levels for k_I is a very important aspect, regarding which the experimenter has to be careful. At least three levels should be chosen for k_I since for any given setting, the variation in the output is approximately a quadratic function of k_I and the quadratic effect of k_I should be important. However, keeping in mind the fact that grossly improper choice of k_I may make the output unstable and upset the entire process, some amount of caution would usually be exercised in the selection of its levels. Thus, it may not be possible to hit the optimum with a single experiment and additional runs may be added later in the evolutionary operation mode as suggested by Box and Kramer (1992).

Let Y_{ijkt} be the t^{th} measured value of the characteristic at the i^{th} level of $\mathbf{X}$, j^{th} level of k_I, and k^{th} level of $\mathbf{N}$ ($i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$, $t = 1, 2, \ldots, \tau$). Let $e_{ijkt} = Y_{ijkt} - T$. Then we compute the estimated value of the performance measure corresponding to the i^{th} level of $\mathbf{X}$ and j^{th} level of k_I as

$$\widehat{PM_{ij}} = \frac{1}{K\tau - 1} \sum_{t=1}^{K\tau} \left(e_{ijkt} - \overline{e_{ij..}} \right)^2,$$

where

$$\overline{e_{ij..}} = \tfrac{1}{K\tau} \textstyle\sum_{k=1}^{K} \sum_{t=1}^{\tau} e_{ijkt}.$$

Next, fit the linear regression model

$$\ln \widehat{PM} \;\; = \;\; f(\mathbf{X}, k_I), \tag{14}$$

and we determine optimum values of $\mathbf{X}$ and k_I by optimizing the fitted function $f(\mathbf{X}, k_I)$.

The above analysis is based on the assumption that all the factor-level combinations would produce a stationary output around T thereby ensuring the finiteness of $V(e_t)$. As seen in Section 2.4.2, this may not be the case if k_I is chosen at a level beyond the range within which it is capable of producing a stationary output. If it is found that for some level combinations (i, j, k), $\overline{e_{ij.}}$ is largely different from zero, it would be pragmatic to use the sample mean squared error $m_{ij}^2 = \frac{1}{K\tau - 1} \sum_{t=1}^{K\tau} \left(Y_{ijkt} - T \right)^2$ instead of $\widehat{PM_{ij}}$.

2.5 A simulation study

Let us revisit the example discussed in Section 2.4.2 for a simulation study:

$$e_t = -2X_1 - N + 0.5X_1N + (2 - 0.75X_1)C_{t-1} + z_t.$$

We assume that $z_t = \phi z_{t-1} + a_t$ is an AR(1) process with $\phi = 0.7$, $E(a_t) = 0$, $Var(a_t) = (2 - 0.8X_1)^2$ and N is a random variable with mean 0 and variance 1. Also assume that besides X_1, there is another control factor X_2, which the experimenter will consider for experimentation.

The experimenter's objective is to choose robust settings of X_1, X_2 and the optimal integral control law $C_t = -k_0 - k_I \sum_{i=1}^{t} e_i$ so that the deviation of e is minimal around zero.

2.5.1 Response modeling

Three levels for each of the two control factors X_1 and X_2, two levels for the noise factor N and three levels for the adjustment factor C. The levels of the factors are shown in Table 1. The control array is a 3^2 design, and each of the 9 combinations of the control factors are crossed with the two levels of the noise factor N. For each of the 18 control-noise combinations, each level of C is held constant till 10 observations are generated. From these 30 observations in each control-noise cell, a transfer function of the form $e_t = \alpha + gC_{t-1} + z_t$ is fit using weighted least squares, and an appropriate

Table 1: Factors and Levels for Response Modeling Simulation Experiment

Factor		Level 1	2	3
X_1.	First control factor	2.2	2.3	2.4
X_2.	Second control factor	3	5	7
N.	Noise factor	-2	2	–
C.	Adjustment factor	10	15	20

Table 2: Summarized Data from Response Modeling Simulation Experiment

Control factor		N = -2			N = 2		
X_1	X_2	α	g	σ^2	α	g	σ^2
2.2	3	-4.625	0.3511	0.0616	-4.182	0.3654	0.0618
2.2	5	-4.578	0.3571	0.0550	-4.178	0.3330	0.0535
2.2	7	-4.595	0.3504	0.0613	-4.238	0.3397	0.0603
2.3	3	-4.909	0.2753	0.0308	-4.309	0.2752	0.0254
2.3	5	-4.909	0.2718	0.0203	-4.330	0.2832	0.0242
2.3	7	-4.887	0.2848	0.0274	-4.287	0.2820	0.0248
2.4	3	-5.200	0.1977	0.0064	-4.366	0.1920	0.0061
2.4	5	-5.207	0.1989	0.0055	-4.391	0.1976	0.0054
2.4	7	-5.220	0.2013	0.0070	-4.405	0.1991	0.0065

time series model is fit to the residuals. The summarized experimental data are shown in Table 2.

It is found that the residuals constitute an AR(1) process with estimated parameter $\hat{\phi} = 0.61$. As described in Section 2.4.4, treating α, g and σ^2 as three different responses, using the methodology described in Wu and Hamada (2000), it is found that the significant factors affecting α are X_1 and N, and g and σ^2 are affected by X_1 only.

The following models are obtained:

$$\begin{aligned} \alpha(\mathbf{X}, \mathbf{N}) &= -1.99X_1 - 1.061N + 0.527X_1N, \\ g(\mathbf{X}) &= 2.02 - 0.76X_1, \\ \sigma^2(\mathbf{X}) &= 4.36 - 3.51x + 0.71X_1^2. \end{aligned}$$

Consequently, the complete fitted model is

$$e_t = -1.99X_1 - 1.061N + 0.527X_1N + (2.02 - 0.76X_1)C_{t-1} + z_t,$$

where $z_t = \phi z_{t-1} + a_t$ is an AR(1) process with $\phi = 0.61$, and $Var(a_t) = 4.36 - 3.51x + 0.71X_1^2$.

Substituting all the estimated parameters of the model in (13) and (12), we get

$$\begin{aligned} PM(\mathbf{X}, k_I) &= f_1(X_1)f_2(X_1), \quad \text{where,} \\ f_1(X_1) &= \frac{1.2422}{\Big(1 + \{1 - k_I(2.02 - 0.76X_1)\}\Big)\Big(1 - 0.61\{1 - k_I(2.02 - 0.76X_1)\}\Big)}, \\ f_2(X_1) &= (4.36 - 3.51X_1 + 0.71X_1^2) + (1.061 - 0.527X_1)^2. \end{aligned}$$

Minimizing the above function subject to the constraints $2.2 \leq X_1 \leq 2.4$ and $0 \leq k_I \leq \frac{2}{2.02 - 0.76X_1}$, we get $X_1^* = 2.3569$ and $k_I^* = 2.9739$. These are fairly good estimates of the true optimal values $(X_1^*, k_I^*) = (2.2990, 2.8497)$. The optimal value of k_0 is obtained as $k_0^* = 20.5$.

2.5.2 Performance measure modeling

For this experiment, we choose three levels for each of the factors X_1, X_2, k_0, k_I and two levels for the noise factor N. The levels of the control factors X_1, X_2, noise factor N and the control equation parameters k_0 and k_I are shown in Table 3. A 3^4 design is used, in which, for each combination of X_1, X_2, k_0, k_I, 100 observations are generated with N at level -1 and 100 observations with N at level $+1$. The mean squared error of these 200 observations are computed. The experimental data, i.e., the mean squared errors corresponding to the 81 runs are shown in Table 4. Finding out the significant linear and quadratic effects following the methodology described in Chapter 5 of Wu and Hamada (2000) and fitting a second-order regression equation relating log(MSE) to the significant variables, the following performance measure

model is obtained :

$$\begin{aligned}\log(\widehat{PM}) &= 97.81 - 104.3X_1 + 27.98X_1^2 + 2.047k_0 + 0.00576k_0^2 \\ &+ 5.489k_I + 0.7828k_I^2 - 0.9895X_1k_0 - 4.447X_1k_I.\end{aligned}$$

The problem is thus

$$\text{minimize } \log(\widehat{PM})$$

$$\text{subject to}$$

$$2.2 \leq X_1 \leq 2.4, 12 \leq k_0 \leq 24, 1.5 \leq k_I \leq 4.5.$$

The minimum is obtained at $X_1 = 2.32, k_0 = 12, k_I = 3.2$; which is reasonably close to the true optimum, considering the fact that in this approach we are trying to approximate a complex response surface by a simple second-order polynomial function. A sequential approach may be adopted to explore the response surface and the settings may be fine tuned.

With this simple model involving one control variable, the response modeling and performance measure modeling approach give almost identical results (except for k_0, which is, of course, not as important as $\mathbf{X}$ and k_I). However, as the underlying model becomes increasingly complex with addition of more control variables, the simple second-order polynomial function obtained from the response modeling method may not be good enough to model the performance measure properly. In such cases, it may be better to conduct a response modeling experiment at the first stage to get a fairly good idea about the significant variables, their interactions with noise and then use a closed-loop performance measure modeling experiment to fine-tune the settings.

2.6 Robust design with the MMSE control scheme and the pure-gain model

Once again, let us consider model (8) with the same assumptions for the process disturbance z_t as mentioned at the beginning of Section 2.4.2. Under the MMSE

Table 3: Factors and Levels for Performance Measure Modeling Simulation Experiment

Factor		Level 1	2	3
X_1.	First control factor	2.2	2.3	2.4
X_2.	Second control factor	3	5	7
k_0.	Control equation constant	12	16	24
k_I.	Control equation constant	1.5	3.0	4.5

Table 4: Summarized Data from Performance Measure Modeling Simulation Experiment

Factor levels				MSE from 200 observations	Factor levels				MSE from 200 observations
X_1	X_2	k_0	k_I	MSE	X_1	X_2	k_0	k_I	MSE
2.2	3	12	1.5	0.0705	2.3	5	18	3.0	0.0270
2.2	3	12	3.0	0.0704	2.3	5	18	4.5	0.0300
2.2	3	12	4.5	0.1165	2.3	5	24	1.5	0.0516
2.2	3	18	1.5	0.0741	2.3	5	24	3.0	0.0495
2.2	3	18	3.0	0.0759	2.3	5	24	4.5	0.0443
2.2	3	18	4.5	0.1173	2.3	7	12	1.5	0.0528
2.2	3	24	1.5	0.1829	2.3	7	12	3.0	0.0429
2.2	3	24	3.0	0.1590	2.3	7	12	4.5	0.0520
2.2	3	24	4.5	0.2227	2.3	7	18	1.5	0.0367
2.2	5	12	1.5	0.0720	2.3	7	18	3.0	0.0269
2.2	5	12	3.0	0.0720	2.3	7	18	4.5	0.0307
2.2	5	12	4.5	0.0963	2.3	7	24	1.5	0.0667
2.2	5	18	1.5	0.0962	2.3	7	24	3.0	0.0496
2.2	5	18	3.0	0.0806	2.3	7	24	4.5	0.0521
2.2	5	18	4.5	0.2000	2.4	3	12	1.5	0.0958
2.2	5	24	1.5	0.1469	2.4	3	12	3.0	0.0567
2.2	5	24	3.0	0.1269	2.4	3	12	4.5	0.0536
2.2	5	24	4.5	0.2304	2.4	3	18	1.5	0.0399
2.2	7	12	1.5	0.0669	2.4	3	18	3.0	0.0289
2.2	7	12	3.0	0.0653	2.4	3	18	4.5	0.0238
2.2	7	12	4.5	0.1142	2.4	3	24	1.5	0.0150
2.2	7	18	1.5	0.0804	2.4	3	24	3.0	0.0114
2.2	7	18	3.0	0.0946	2.4	3	24	4.5	0.0124
2.2	7	18	4.5	0.1168	2.4	5	12	1.5	0.0904
2.2	7	24	1.5	0.1616	2.4	5	12	3.0	0.0622
2.2	7	24	3.0	0.1435	2.4	5	12	4.5	0.0510
2.2	7	24	4.5	0.2268	2.4	5	18	1.5	0.0421
2.3	3	12	1.5	0.0492	2.4	5	18	3.0	0.0244
2.3	3	12	3.0	0.0390	2.4	5	18	4.5	0.0206
2.3	3	12	4.5	0.0485	2.4	5	24	1.5	0.0164
2.3	3	18	1.5	0.0357	2.4	5	24	3.0	0.0127
2.3	3	18	3.0	0.0315	2.4	5	24	4.5	0.0112
2.3	3	18	4.5	0.0386	2.4	7	12	1.5	0.0851
2.3	3	24	1.5	0.0611	2.4	7	12	3.0	0.0557
2.3	3	24	3.0	0.0441	2.4	7	12	4.5	0.0508
2.3	3	24	4.5	0.0528	2.4	7	18	1.5	0.0408
2.3	5	12	1.5	0.0639	2.4	7	18	3.0	0.0272
2.3	5	12	3.0	0.0403	2.4	7	18	4.5	0.0258
2.3	5	12	4.5	0.0525	2.4	7	24	1.5	0.0146
2.3	5	18	1.5	0.0312	2.4	7	24	3.0	0.0123
–	–	–	–	–	2.4	7	24	4.5	0.0120

control scheme, C_{t-1} is to be set in such a way so that the one-step ahead forecast of Y_t is equal to the target T (or equivalently, the forecast of e_t is zero). An elaborate discussion can be found in Box and Luceno (1997).

Thus, under the MMSE scheme for the pure gain model, C_{t-1} should be such that

$$\hat{e}_{t-1}(1) = \alpha(\mathbf{X}, \mathbf{N}) + g(\mathbf{X})C_{t-1} + \hat{z}_{t-1}(1) = 0. \tag{15}$$

Since $\alpha(\mathbf{X}, \mathbf{N})$ will not be known in reality, the MMSE control equation can be obtained from the above by taking the expectation over $\mathbf{N}$, i.e.,

$$C_{t-1} = \frac{1}{g(\mathbf{X})}\Big(- \alpha(\mathbf{X}) - \hat{z}_{t-1}(1)\Big). \tag{16}$$

Substituting (14) into (8), we get

$$e_t = \Big(\alpha(\mathbf{X}, \mathbf{N}) - \alpha(\mathbf{X})\Big) + (z_t - \hat{z}_{t-1}(1)). \tag{17}$$

Thus, we have

$$\begin{aligned} Var(e_t) &= Var_a E_{\mathbf{N}}(e_t|a) + E_a Var_{\mathbf{N}}(e_t|a) \\ &= Var\Big(z_t - \hat{z}_{t-1}(1)\Big) + \pi(\mathbf{X}) \\ &= \sigma^2(\mathbf{X}) + \pi(\mathbf{X}), \end{aligned} \tag{18}$$

where $\pi(\mathbf{X})$ is defined as before.

The last step of (18) follows from the properties of the MMSE forecast (Box et al. 1994, Chapter 5). The fact that under the ARIMA(0,1,1) disturbance, the discrete PI scheme with $k_I = \frac{\lambda}{g(\mathbf{X})}$ is the same as the MMSE scheme can be easily seen by substituting $k_I = \frac{\lambda}{g(\mathbf{X})}$ in (11) and observing that $Var(e_t)$ reduces to the form given by (18).

It is evident that with the MMSE control scheme, only the response modeling approach would work, since it is not possible to establish the control equation and observe the output error under control without estimating model (8). Since MMSE

control schemes reduce the variance of the output error to the variance of the underlying white noise, by the argument in Section 2.4.2, one may use the two-stage approach in this case. However, as discussed before, a two-stage approach will lead to improper choice of the component $\alpha(\mathbf{X})$ in the control equation.

2.7 A case study

In Section 2.2 a study on optimization of a control scheme of the packing process of a urea manufacturing plant (Dasgupta et al., 2002) was mentioned as a motivating example. The underlying control scheme was a discrete PI scheme, slightly different from the classical one (see Appendix of Dasgupta et al. 2002). Among the 16 factors listed in the original case study, D (auto compensation proportional constant) and K (in-flight material compensation - start) correspond to k_I and k_0 respectively. The recoded factors and levels of the experiment are given in Table 5. The packing system is shown in Figure 3.

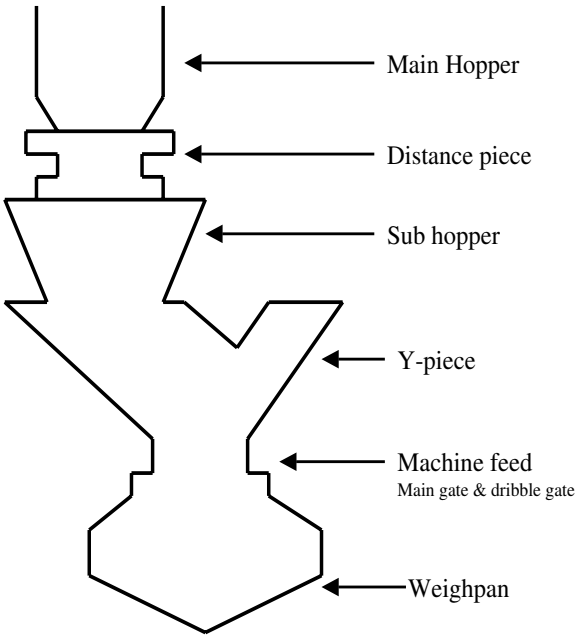


Figure 3: The Packing Process

The original experiment was conducted and analyzed somewhat superficially just

Table 5: Factors and Levels, Packing Plant Experiment

		Level		
	Control Factor	1	2	3
X_1.	Sample frequency	10	20	–
X_2.	Sample number	3	5	–
X_3.	Sample frequency timer	240	300	–
X_4.	Main feed blanking timer (sec)	0.5	0.9	–
X_5.	Dribble feed blanking timer (sec)	0.5	0.9	–
X_6.	Discharge timer (sec)	0.3	0.6	0.9
X_7.	Dribble feed time correction constant (sec)	0.1	0.8	–
X_8.	Gate allowance timer (sec)	0.4	2.0	–
X_9.	Feed delay timer (sec)	0.3	0.7	–
X_{10}.	Dribble feed quantity (start) (Kg)	8	12	–
X_{11}.	Discharge cut-off value (Kg)	20	30	–
X_{12}.	Overweight tolerance (Kg)	0.15	0.20	–
X_{13}.	Underweight tolerance (Kg)	0.10	0.15	–
X_{14}.	Dribble feed time (sec)	1.2	1.4	1.6
		Level		
	PI control scheme parameters	1	2	3
k_I.	Auto compensation proportional constant	0.2	0.3	0.4
k_0.	In-flight material compensation (start)	2.0	3.0	–
		Level		
	Noise factors	1	2	3
N.	Composition of material	N_1	N_2	N_3

like "another" robust design exercise, not recognizing the specific roles of factors k_I and k_0. It may also be noted that the original experiment does not explicitly consider any noise factor. However, since we have three replicates, we can consider them to correspond to three levels of the noise factor material composition N.

An L_{32} orthogonal array (OA) with an idle column was used to design the experiment. The idle column method is a technique to generate three level columns by collapsing two columns in a two-level orthogonal array (Taguchi 1987; Grove & Davis 1991; Huwang, Wu & Yen 2002). Besides the 16 main effects (19 degrees of freedom), provisions were kept for estimation of five interaction effects, viz. $X_1 \times X_2$, $X_6 \times X_{11}$, $X_6 \times X_9$, $X_7 \times X_{10}$ and $X_2 \times k_I$.

Sixty bags were filled for each of the $32 \times 3 = 96$ level combinations. The target weight was set at $T = 50.5$ kg.

As mentioned in Section 2.2, in-flight material composition is the adjustment factor C in this case. $\mathbf{X} = \{X_1, \ldots, X_{14}\}$ denotes the set of control factors. Note that we can use equation (7) to model Y_t, the weight of a packed bag at time t, where z_t in (7) is the disturbance due to other uncontrollable noise factors such as temperature

Table 6: Data from the Packing Experiment

Control Factor														PI		Var over 3 noise levels
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	k_I	k_0	$\ln(s^2)$
+	−	+	+	+	0	−	+	−	+	+	−	+	0	1	−	-5.6821
−	−	−	+	−	0	−	−	−	−	+	+	−	1	0	+	-4.2966
+	+	−	−	+	0	+	+	−	+	−	+	−	0	1	+	-4.0499
−	+	+	−	−	0	+	−	−	−	−	−	+	1	0	−	-4.6484
+	+	+	−	+	0	+	−	+	−	+	−	−	1	1	−	-7.1007
−	+	−	−	−	0	+	+	+	+	+	+	+	0	0	+	-6.7234
+	−	−	+	+	0	−	−	+	−	−	+	+	1	1	+	-4.4138
−	−	+	+	−	0	−	+	+	+	−	−	−	0	0	−	-7.0905
−	+	+	+	−	2	−	−	−	−	+	+	+	0	1	−	-7.4974
+	+	−	+	+	2	−	+	−	+	+	−	−	1	0	+	-4.4772
−	−	−	−	−	2	+	−	−	−	−	−	−	0	1	+	-4.4792
+	−	+	−	+	2	+	+	−	+	−	+	+	1	0	−	-6.3351
−	−	+	−	−	2	+	+	+	+	+	+	−	1	1	−	-7.5304
+	−	−	−	+	2	+	−	+	−	+	−	+	0	0	+	-3.8414
−	+	−	+	−	2	−	+	+	+	−	−	+	1	1	+	-4.5632
+	+	+	+	+	2	−	−	+	−	−	+	−	0	0	−	-7.7693
+	−	−	−	−	1	−	+	−	−	+	−	−	0	1	−	-8.2079
−	−	+	−	+	1	−	−	−	+	+	+	+	2	2	+	-4.3529
+	+	+	+	−	1	+	+	−	−	−	+	+	0	1	+	-4.6972
−	+	−	+	+	1	+	−	−	+	−	−	−	2	2	−	-8.6886
+	+	−	+	−	1	+	−	+	+	+	−	−	2	1	−	-7.9619
−	+	+	+	+	1	+	+	+	−	+	+	+	0	2	+	-5.4851
+	−	+	−	−	1	−	−	+	+	−	+	+	2	1	+	-3.7536
−	−	−	−	+	1	−	+	+	−	−	−	−	0	2	−	7.5840
−	+	−	−	+	2	−	−	−	+	+	+	+	0	1	−	-7.3038
+	+	+	−	−	2	−	+	−	−	+	−	−	2	2	+	-3.5010
−	−	+	+	+	2	+	−	−	+	−	−	−	0	1	+	-4.2572
+	−	−	+	−	2	+	+	−	−	−	+	+	2	2	−	-7.6150
−	−	−	+	+	2	+	+	+	−	+	+	−	2	1	−	-8.0317
+	−	+	+	−	2	+	−	+	+	+	−	+	0	2	+	-3.7810
−	+	+	−	+	2	−	+	+	−	−	−	+	2	1	+	-6.9418
+	+	−	−	−	2	−	−	+	+	−	+	−	0	2	−	-8.0519

of the flowing material, moisture content in the hopper, etc. and has a certain time series structure.

The response modelling approach cannot be employed because it would require an open-loop experiment as discussed in Section 2.4.4.1. Here, since the experiment was conducted with a closed loop with k_0 (starting value of the adjustment factor C) and k_I as two of the experimental factors, the performance measure modelling approach needs to be used. Recall also that in this approach it is not necessary to fit a time series model to z_t; rather, one has to express the performance measure (variance or mean squared error of data collected over time) as a function of the control factors, k_0 and k_I.

Note that in performance measure modelling, for each experimental combination

we would generate a time series over the different levels of noise. Such was, however, not the case in the original experiment. For purpose of illustration, we assume that the 180 observations corresponding to each of the 32 experimental runs constitute a single time series. The sample standard deviation of the 180 observations corresponding to each run is shown in Table 6.

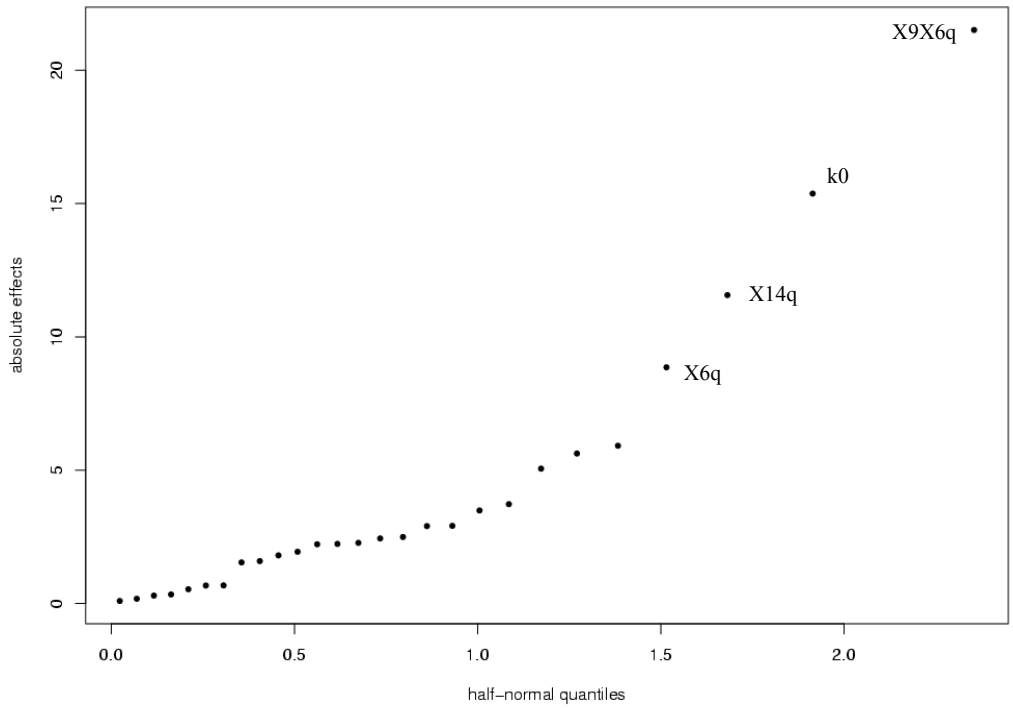


Figure 4: Half-normal Plot

Using half-normal plots for the 27 treatment effects (Figure 4), we find that four effects viz. X_9X_{6q}, k_0, X_{14q} and X_{6q} stand apart from the rest. Using Lenth's method (Wu and Hamada 2000, Chapter 4) as a formal test of significance it is seen that all of these 4 effects are significant at 1% level. Note that for any three-level factor X, we denote the linear and quadratic effects by X_l and X_q respectively (Wu and Hamada 2000, Chapter 5).

We notice that neither the main effect of k_I nor the $X_2 \times k_I$ interaction (which is estimable) is significant. Another factor that is suspected to interact with k_I is

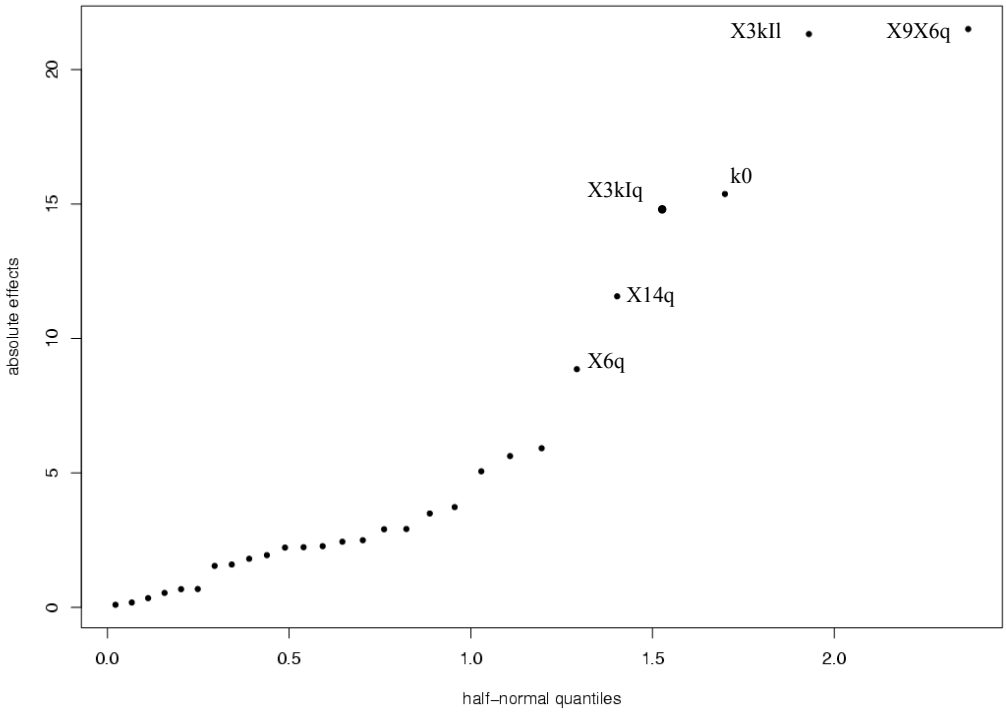


Figure 5: Modified Half-normal Plot (with $X_3 \times k_I$ interaction)

X_3, and we explore the possibility of incorporating the $X_3 \times k_I$ interaction into our analysis by replacing some insignificant effects in the preliminary model. Using the table of interactions for the L_{32} OA, we find that the $X_3 \times k_I$ interaction can be estimated from columns 8 and 9 of the OA. Out of these two columns, 8 corresponds to factor X_{11} which is seen to be insignificant and 9 is a free column to which no other main effect or significant interaction is assigned. Thus we re-perform the analysis by including the $X_3 \times k_I$ interaction (with two degrees of freedom) instead of X_{11}.

The half-normal plot (Figure 5) now identifies 6 effects standing above the rest. Apart from the four that were already seen to be significant, the interaction effects $X_3 k_{I_l}$ and $X_3 k_{I_q}$ turn out to be significant at 1% level.

From the plots of significant main effects (Figure 5) and the $X_6 \times X_9$ interaction (Figure 6), we choose the optimal settings of k_0, X_9, X_6 and X_{14} as $k_0 = -1$,$X_9 = 1$,$X_6 = 1$ and $X_{14} = 2$. Note that all of these factors (or interactions involving them)

were found significant in the original analysis and the same levels (although different notations were used) were selected as the optimum ones.

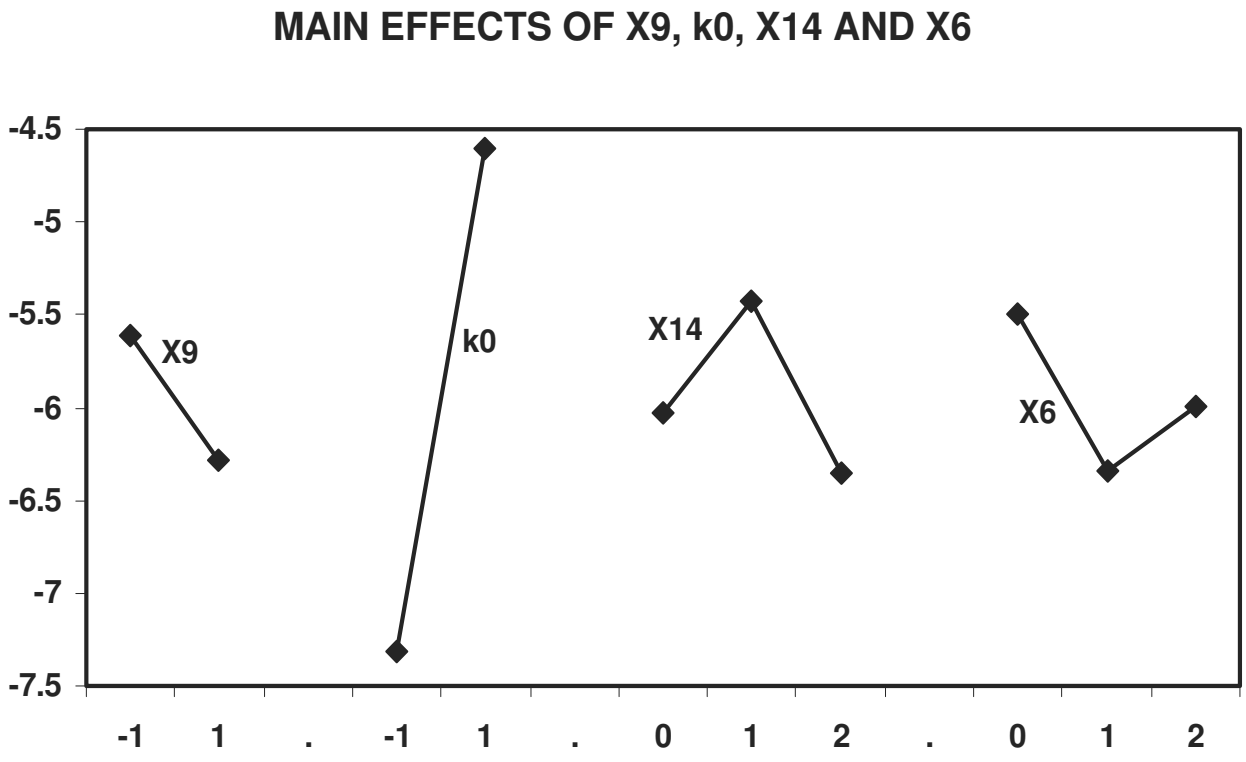


Figure 6: Main Effects Plots

The interaction plot of $X_3 \times k_I$ (Figure 6) suggests that at neither of the two levels of X_3, the optimum k_I could be reached. Corresponding to $X_3 = 1$, $k_I^*(X_3) \leq 0$ while corresponding to $X_3 = -1$, $k_I^*(X_3) \geq 2$. We also note that the curve corresponding to $X_3 = 1$ is slightly convex as expected, whereas that corresponding to $X_3 = -1$ exhibits a slight concavity, which can be attributed to sampling error or effect of higher order interactions. Since this difference in convexity results in significance of the quadratic component of the interaction, we may only consider the linear component of the interaction while modeling $\ln(P\hat{M})$. From the interaction plot, we choose $X_3 = -1$ and $k_I = 2$ as the optimal settings.

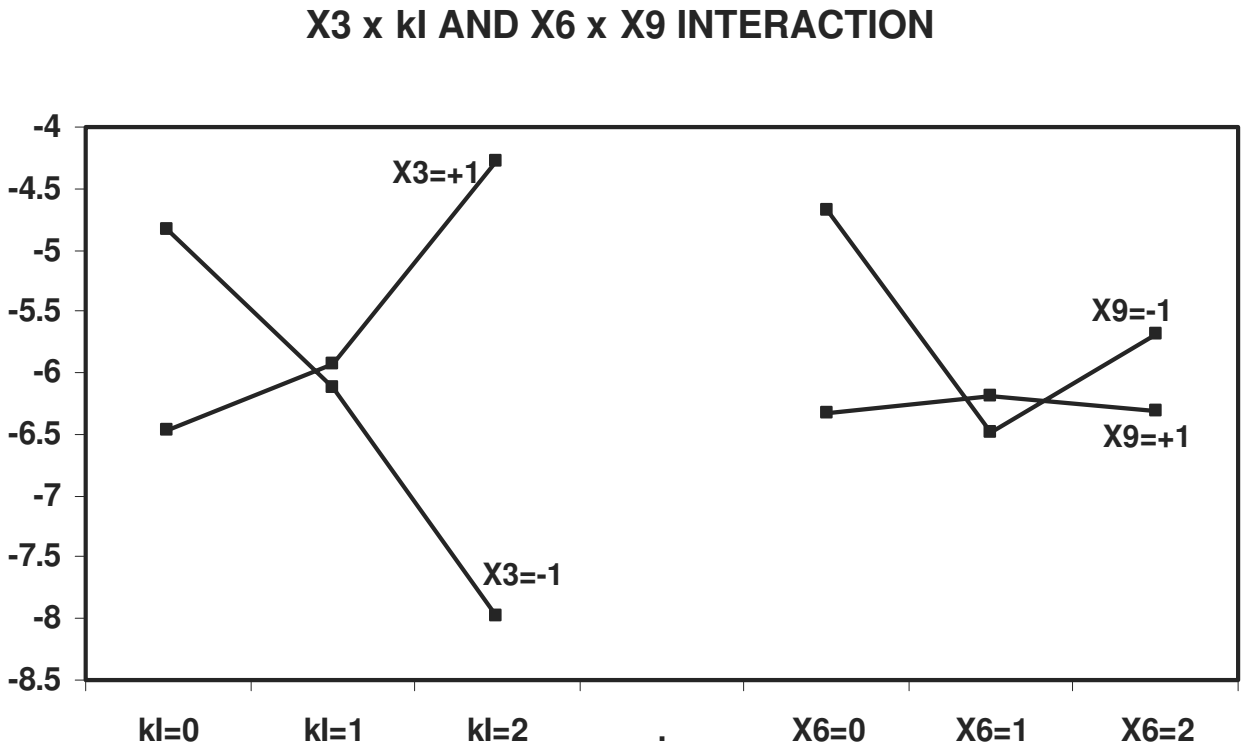


Figure 7: Interaction Plots

The following model is thus obtained:

$$\widehat{PM} = -5.7438 + 1.2408x_{k_0} - 0.832x_6 + 0.3344x_6^2 + 1.0725x_{14} - 0.5863x_{14}^2 + 0.2362x_3x_{k_I} - 0.1240x_9x_6^2. \quad (19)$$

Substituting the optimal settings of the control factors in the above model, we get the optimal value of the performance measure as $\widehat{PM^*} = -8.2788$, which corresponds to a standard deviation of 0.0159. The original experiment was able to reduce the output standard deviation drastically to 0.031 from the existing value of 0.121. We find from this re-analysis that with the newly recommended settings, it might have been possible to reduce the variation to almost 50% of what had been achieved.

This analysis also points out the importance of adopting a sequential approach for the performance modeling experiment. It is clear that with wider choices of levels of k_I, it might have been possible to reduce the output variance even further.

2.8 Robust Parameter Design with on-line noise factors

One important assumption in the framework proposed so far in this chapter is that the noise factors (that appear in the experiment) are uncorrelated over time. There are many examples of noise factors, e.g., material composition in the case study, which are uncorrelated or weakly correlated over time and may be well approximated by a white noise disturbance. However, there are also many noise factors associated with dynamic systems which do not satisfy this assumption. In this section we discuss the case where we have some correlated noise factors (some of them may be measurable on-line), develop performance measures for systems containing such factors, and give an outline of the related optimization procedure.

Extending the framework described in Section 2.4.1, let $\mathbf{N} = \{\mathbf{R}, \mathbf{Q}\}$, where $\mathbf{R}$ denotes the set of off-line noise factor and $\mathbf{Q}$ the set of on-line noise factors. For simplicity, we consider a single on-line noise factor Q. The following forms are considered for the functions β and $f(e_t, e_{t-1}, \ldots)$ described in Section 2.3.1:

$$\begin{aligned} \beta(\mathbf{X}, \mathbf{N}, C_{t-1}, C_{t-2}, \ldots) &= \alpha_0(\mathbf{X}, \mathbf{R}) + g(\mathbf{X})C_{t-1} + \beta(\mathbf{X})Q_t \\ f(e_t, e_{t-1}, \ldots) &= -k_0 - k_I \sum_{i=1}^{t} e_i. \end{aligned}$$

We thus postulate the following first-order pure-gain dynamic model

$$Y_t = \alpha_0(\mathbf{X}, \mathbf{R}) + g(\mathbf{X})C_{t-1} + \beta(\mathbf{X})Q_t + z_t. \tag{20}$$

If T denotes the target, and $e_t = Y_t - T$ denotes the deviation from the target, then

$$e_t = \alpha(\mathbf{X}, \mathbf{R}) + g(\mathbf{X})C_{t-1} + \beta(\mathbf{X})Q_t + z_t, \tag{21}$$

where $\alpha(\mathbf{X}, \mathbf{R}) = \alpha_0(\mathbf{X}, \mathbf{R}) - T$.

Note that this is essentially the same as (8), the added aspect being the dependence of the dynamics on Q. If an integral control scheme is employed to a process described by (21), k_I necessarily has to be a function of $\mathbf{X}$ and $\mathbf{X}$ has to be such that the output is least sensitive to the effects of $\mathbf{R}$ and Q.

2.8.1 Performance measure for integral control

Let us assume that z_t and Q_t are independent linear processes defined as

$$\begin{aligned} z_t &= \sum_{j=0}^{\infty} \psi_j a_{t-j}, \\ Q_t &= \sum_{j=0}^{\infty} \phi_j w_{t-j}, \end{aligned}$$

where $\{a_t\}$ and $\{w_t\}$ are white noise processes with zero mean and variances $\sigma_a^2(\mathbf{X})$ and $\sigma_w^2(\mathbf{X})$ respectively. Recalling that in an integral control scheme C_{t-1} is set to be $-k_0 - k_I \sum_{i=1}^{t-1} e_i$, we have from (21),

$$\begin{aligned} e_t &= \alpha(\mathbf{X},\mathbf{R}) - g(\mathbf{X})\Big(k_0 + k_I \sum_{i=1}^{t-1} e_i\Big) + \beta(\mathbf{X})Q_t + z_t \\ &= \Big(\alpha(\mathbf{X},\mathbf{R}) - g(\mathbf{X})k_0\Big) - g(\mathbf{X})k_I \sum_{i=1}^{t-1} e_i + \beta(\mathbf{X}) \sum_{j=0}^{\infty} \phi_j w_{t-j} + \sum_{j=0}^{\infty} \psi_j a_{t-j}. \end{aligned} \tag{22}$$

As in Section 2.4.2, we consider $Var(e_t)$ as the appropriate performance measure, assuming that the controller output is asymptotically stable around zero. $Var(e_t)$ can be obtained as follows:

$$\begin{aligned} Var(e_t) &= Var_a E_w\Big[\Big(E_{\mathbf{R}}(e_t|a,w)\Big)|a\Big] \\ &+ E_a Var_w\Big[\Big(E_{\mathbf{R}}(e_t|a,w)\Big)|a\Big] \\ &+ E_a E_w\Big[\Big(Var_{\mathbf{R}}(e_t|a,w)\Big)|a\Big] \\ &= Var_a E_w(u_t|a) + E_a Var_w(u_t|a) + \pi(\mathbf{X}) \\ &= Var_a(v_t) + Var_w(u_t|a) + \pi(\mathbf{X}). \end{aligned} \tag{23}$$

where

$$\begin{aligned} u_t &= E_{\mathbf{R}}(e_t|a,w) \\ &= \left(\alpha(\mathbf{X}) - g(\mathbf{X})k_0\right) - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + \beta(\mathbf{X}) \sum_{j=0}^{\infty} \phi_j w_{t-j} + \sum_{j=0}^{\infty} \psi_j a_{t-j}, \quad (24) \end{aligned}$$

$$\begin{aligned} v_t &= E_w(u_t|a) \\ &= \left(\alpha(\mathbf{X}) - g(\mathbf{X})k_0\right) - g(\mathbf{X})k_I \sum_{i=1}^{t-1} v_i + \sum_{j=0}^{\infty} \psi_j a_{t-j}, \quad (25) \end{aligned}$$

and

$$\begin{aligned} \alpha(\mathbf{X}) &= E_{\mathbf{R}}(\alpha(\mathbf{X},\mathbf{R})), \\ \pi(\mathbf{X}) &= Var_{\mathbf{R}}(\alpha(\mathbf{X},\mathbf{R})). \end{aligned}$$

The first two terms in (23) are functions of $\mathbf{X}$ and k_I. Define $V_a(\mathbf{X},k_I) = Var_a(v_t)$ and $V_w(\mathbf{X},k_I) = V_w(u_t|a)$. Then, the performance measure is given by

$$PM(\mathbf{X},k_I) = V_a(\mathbf{X},k_I) + V_w(\mathbf{X},k_I) + \pi(\mathbf{X}). \quad (26)$$

Clearly, the forms of $V_a(\mathbf{X},k_I)$ and $V_w(\mathbf{X},k_I)$ depend on the time series structures of z_t and Q_t respectively. Also, $V_a(\mathbf{X},k_I)$ and $V_w(\mathbf{X},k_I)$ depend on $\mathbf{X}$ through the functions $g(\mathbf{X})$ and $\beta(\mathbf{X})$ respectively. Therefore if model (21), the time series structures associated with z_t and Q_t, and the distribution of $\mathbf{R}$ are completely known, then we can compute the performance measure from (26) by deriving and substituting the expressions for the three terms. A couple of examples are given below:

EXAMPLE 1. If z_t and Q_t are ARIMA(0,1,1) processes with parameters λ_z and λ_Q respectively, then

$$V_a(\mathbf{X},k_I) = \begin{cases} \frac{1+\theta_z^2-2\phi(\mathbf{X},k_I)\theta_z}{1-\phi^2(\mathbf{X},k_I)}\sigma_a^2(\mathbf{X}) & ,|\phi(\mathbf{X},k_I)| \leq 1 \\ \infty & ,\text{otherwise} \end{cases},$$

$$V_w(\mathbf{X}, k_I) = \begin{cases} \frac{1+\theta_Q^2-2\phi(\mathbf{X},k_I)\theta_Q}{1-\phi^2(\mathbf{X},k_I)}\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X}) & , |\phi(\mathbf{X}, k_I)| \leq 1 \\ \infty & , \text{otherwise} \end{cases},$$

where $\phi(\mathbf{X}, k_I) = 1 - g(\mathbf{X})k_I$, $\theta_z = 1 - \lambda_z$ and $\theta_Q = 1 - \lambda_Q$.

The expression for $V_a(\mathbf{X}, k_I)$ follows from the same argument as in Section 2.4.2. To derive the expression for $V_w(\mathbf{X}, k_I)$, note that for fixed a's, we can write from (24)

$$\begin{aligned} u_t &= constant - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + \beta(\mathbf{X})\big(w_t + \lambda_Q \sum_{i=1}^{t-1} w_i\big) \\ &= constant - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + w_t' + \lambda_Q \sum_{i=1}^{t-1} w_i', \end{aligned}$$

where w_t's are iid with zero mean and variance $\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X})$. Following the lines of proof given by Box and Kramer (1992), the result follows.

EXAMPLE 2. If z_t and Q_t are AR(1) processes with parameters ϕ_z and ϕ_Q respectively, then

$$V_a(\mathbf{X}, k_I) = \begin{cases} \frac{2}{(1+\eta(\mathbf{X},k_I))(1-\phi_z\eta(\mathbf{X},k_I))(1+\phi_z)}\sigma_a^2(\mathbf{X}) & , |\phi(\mathbf{X}, k_I)| \leq 1 \\ \infty & , \text{otherwise} \end{cases}.$$

$$V_w(\mathbf{X}, k_I) = \begin{cases} \frac{2}{(1+\eta(\mathbf{X},k_I))(1-\phi_Q\eta(\mathbf{X},k_I))(1+\phi_Q)}\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X}) & , |\phi(\mathbf{X}, k_I)| \leq 1 \\ \infty & , \text{otherwise} \end{cases}.$$

The expression for $V_w(\mathbf{X}, k_I)$ is obtained by noting that for fixed a's, from (24) we have

$$\begin{aligned} u_t &= constant - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + \beta(\mathbf{X})\big(\phi_Q Q_{t-1} + w_t\big) \\ &= constant - g(\mathbf{X})k_I \sum_{i=1}^{t-1} u_i + \sum_{j=0}^{\infty} \phi_Q^j w_{t-j}'. \end{aligned}$$

where w_t's are iid with zero mean and variance $\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X})$. The expression for $V_w(\mathbf{X}, k_I)$ is now easily obtained by using the results of Tsung et al. (1996) in the same way as in Section 2.4.2.

2.8.2 MMSE control scheme

Recall that under the MMSE control scheme, C_{t-1} is to be set in such a way so that the one-step ahead forecast of Y_t is equal to the target T (or equivalently, the forecast of e_t is zero). Thus, under the MMSE scheme for the pure gain model given by (21) with only feedback control, C_{t-1} should be such that

$$\hat{e}_{t-1}(1) = \alpha(\mathbf{X}, \mathbf{R}) + g(\mathbf{X})C_{t-1} + \beta(\mathbf{X})Q_t + \hat{z}_{t-1}(1) = 0.$$

Since $\alpha(\mathbf{X}, \mathbf{R})$ will not be known in reality in the MMSE control equation, it can be replaced by its expectation over $\mathbf{R}$. Further, since at time $t-1$, Q_{t-1} and not Q_t will be known, the control equation will be

$$C_{t-1} = \frac{1}{g(\mathbf{X})}\Big(- \alpha(\mathbf{X}) - \beta(\mathbf{X})Q_{t-1} - \hat{z}_{t-1}(1)\Big). \tag{27}$$

Substituting (27) into (21), we get

$$e_t = \big(\alpha(\mathbf{X}, \mathbf{R}) - \alpha(\mathbf{X})\big) + \beta(\mathbf{X})\big(Q_t - Q_{t-1}\big) + (z_t - \hat{z}_{t-1}(1)). \tag{28}$$

Thus, we have

$$\begin{aligned} Var(e_t) &= Var_R\big(\alpha(\mathbf{X}, \mathbf{R})\big) + \beta^2(\mathbf{X})Var_w\big(Q_t - Q_{t-1}\big) + Var_a\big(z_t - \hat{z}_{t-1}(1)\big) \\ &= \pi(\mathbf{X}) + \beta^2(\mathbf{X})\left(\phi_0^2 + \sum_{j=1}^{\infty}(\phi_j - \phi_{j-1})^2\right)\sigma_w^2(\mathbf{X}) + \sigma_a^2(\mathbf{X}). \end{aligned} \tag{29}$$

Clearly, if model (21), the time series model associated with Q_t, $\sigma_a^2(\mathbf{X})$ and the distribution of $\mathbf{R}$ are known, then we can compute $Var(e_t)$ from (29). Subsequently, the optimal $\mathbf{X}$ can be found by minimizing $Var(e_t)$.

If we want to introduce a feedforward control loop along with the feedback control mechanism (see Figure 8), then Q_{t-1} in equations (27) and (28) will be replaced by

the one-step ahead MMSE forecast $\hat{Q}_{t-1}(1)$ and consequently,

$$\begin{aligned} Var(e_t) &= Var_R\big(\alpha(\mathbf{X},\mathbf{R})\big) + \beta^2(\mathbf{X}) Var_w\big(Q_t - \hat{Q}_{t-1}(1)\big) + Var_a\big(z_t - \hat{z}_{t-1}(1)\big) \\ &= \pi(\mathbf{X}) + \beta^2(\mathbf{X})\sigma_w^2(\mathbf{X}) + \sigma_a^2(\mathbf{X}). \end{aligned} \tag{30}$$

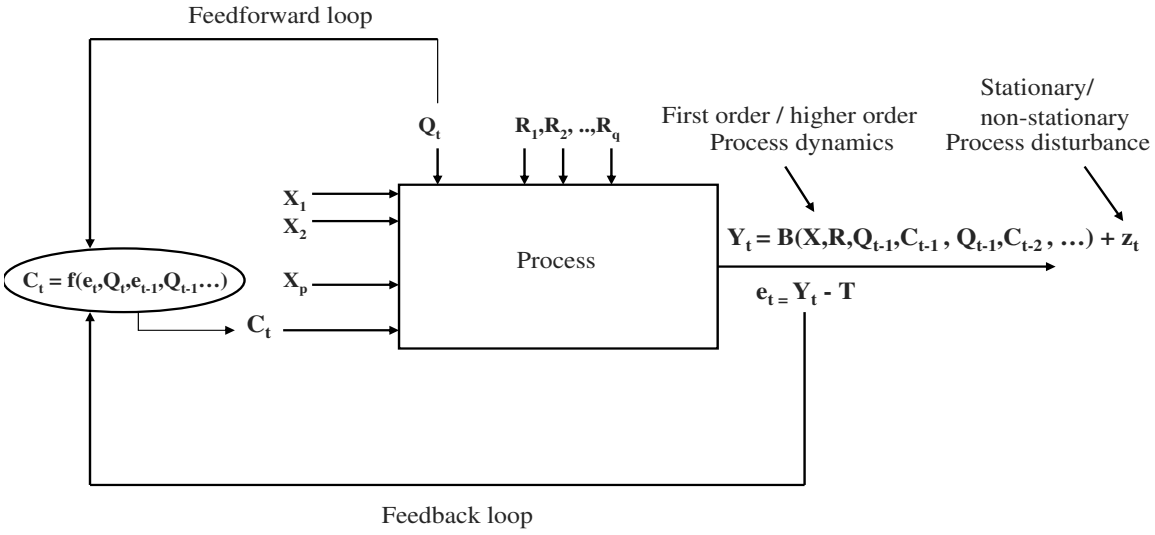


Figure 8: Parameter design with feedback and feedforward control

The last step of (30) follows from the properties of the MMSE forecast (Box et al. 1994, Chapter 5). It is clear from equations (29) and (30) that adding a feedforward loop to the system will be beneficial if

$$G_Q(\phi) = \phi_0^2 + \sum_{j=1}^{\infty}(\phi_j - \phi_{j-1})^2 > 1.$$

For example, if Q_t is white noise with mean 0 and variance 1, then $G_Q(\phi) = 1$, which means, introducing feedforward control will not be beneficial. The reduction in variance that can be achieved by introduction of feedforward control is $G_Q(\phi)\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X})$. Two examples are discussed below.

Example 1: If Q_t is a stationary AR(1) process with parameter ϕ such that $|\phi| < 1$, then

$$G_Q(\phi) = 1 + \sum_{j=1}^{\infty}(\phi^j - \phi^{j-1})^2 = 2/(1+\phi).$$

Thus, if $Q_t = 0.5Q_{t-1} + w_t$, adding an MMSE feedforward controller to the feedback system will reduce the asymptotic variance of the output by $1.33\beta^2(\mathbf{X})\sigma_w^2(\mathbf{X})$.

If Q_t is a nonstationary AR(1) process with parameter ϕ such that $|\phi| > 1$, $G_Q(\phi) = \infty$, which means without feedforward control, the variance of the output will explode. Therefore if an MMSE feedback controller is used for such a system, it is mandatory to also have a feedforward controller.

Example 2: If Q_t is an ARIMA(0,1,1) process with parameter λ, then $G_Q(\phi) = 1 + (1-\lambda)^2$, and the feedforward loop will be beneficial if $\lambda \neq 1$. Note that $\lambda = 1$ makes Q_t a random walk process.

2.9 Concluding remarks

In this chapter we have developed a framework for robust parameter design of systems with feedback control. The suggested approach can be used to obtain the optimal control law and robust parameter design solution at a single stage. Appropriate performance measures have been developed and the design and analysis of experiments for estimation and optimization of these performance measures have been discussed. The advantages of using the proposed approach as compared to a two-stage approach have been discussed. The benefits of using the proposed method have been demonstrated using a simulation study and an example from a packing plant.

Although we have considered the pure-gain dynamic model and primarily the discrete integral control scheme, it should be possible to extend the proposed methodology to a much more generic class of models and other control schemes.

Our proposed framework mostly deals with noise factors (that appear in the experiment) that are uncorrelated over time. Even if this assumption is not true, the framework developed in Sections 2.4-2.6 can still handle correlated noise factors by absorbing the effect of such noise factors in the disturbance term z_t by virtue of the important assumption that $var(z_t)$ depends on the control factors. For example, in

the context of the packing experiment, temperature of the flowing material is an example of such a correlated noise factor, which can hardly be controlled and thus cannot be treated as a experimental factor in the study. However, its effect can easily be absorbed into the disturbance term z_t. However, if such noise factors can be measured on-line, then one can think of augmenting a feedforward loop to the system. Such situations call for an integration of robust parameter design with feedback and feedforward control and has been briefly discussed in the context of MMSE control in Section 2.8.

The performance measure modeling approach is a simple alternative to closed-loop estimation. However, as already discussed before, if the number of significant process variables is large, then a sequential approach may be necessary to obtain the optimal solution. Estimation of the true model from a closed-loop parameter design experiment is another worthy topic for future research.

2.10 References

Astrom, K.J. (1970), *Introduction to Stochastic Control Theory*, New York: Academic Press.

Box, G.E.P, Jenkins G.M., and Reinsel G.C. (1994), *Time Series Analysis - Forecasting and Control*, New Jersey: Prentice Hall.

Box, G.E.P. and Kramer, T. (1992),"Statistical Process Monitoring and Feedback Adjustment - A Discussion," (with discussion) *Technometrics*, 34, 251-285.

Box, G.E.P. and Luceno, A. (1997), *Statistical Control by Monitoring and Feedback Adjustments*, New York: Wiley.

Carpenter, B.H. and Sweeny, H.C. (1965), "Process Improvement with Simplex Self Directing Evolutionary Operation," *Chemical Engineering*, 72(14), 117-126.

Cohen, G.H. and Coon, G.A. (1953), "Theoretical Investigations of retarded control," *Transactions of ASME*, 75, 827-834.

Dasgupta, T., Sarkar, N.R., and Tamankar, K.G.T. (2002), "Using Taguchi methods to improve a control scheme by adjustment of changeable settings," *Total Quality Management*, 13, 863-876.

Davis, M.H.A., and Vinter, R.B. (1985), *Stochastic Modeling and Control*, London: Chapman and Hall.

Del Castillo, E. (2002), *Statistical Process Adjustment for Quality Control*, John Wiley & Sons.

Edwards, I.M., and Jutan, A. (1997), "Optimization and control using response surface methods," *Computers and Chemical Engineering*, 21(4), 441-453.

Grove, D.M. and Davis, T.P. (1991),"Taguchi's Idle Column Method," *Technometrics*, 33, 349-353.

Huwang, L., Wu, C. F. J. and Yen, C. H.(2002), "The idle column method: construction, properties and comparisons," *Technometrics*, 44, 347-355.

Jiang, A. and Jutan, A. (2000), "Response Surface tuning methods in Dynamic Matrix Control of a multivariable pressure tank," *Industrial and Engineering Chemistry Research*, 39, 3835-3843.

Joseph, V.R. (2003), "Robust Parameter Design With Feed-Forward Control," *Technometrics*, 45, 284-291.

Nakano, E. and Jutan, A. (1994), "Application of Response Surface Methodology in Controller Fine-Tuning," *ISA Transactions*, 33, 353-366.

Seborg, D.E., Edgar, T.F., and Mellichamp, D.A. (1989), *Process Dynamics and Control*, New York: Wiley.

Taguchi, G. (1987), *System of Experimental Design - Engineering Methods to Optimize Quality and Minimize Costs*, UNIPUB, Kraus International Publications.

Tsung, F, Wu, H, and Nair, V. N. (1998), "On the Efficiency and Robustness of Discrete Proportional-Integral Control Schemes," *Technometrics*, 40, 214-222.

Tsung, F, Wu, H, and Nair, V. N. (1996), "Efficiency and Robustness of Discrete Proportional-Integral Control Schemes," Technical Report 272, University of Michigan, Dept. of Statistics.

Wu, C.F.J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Yuwana, M. and Seborg, D.E. (1982), "A New Method for On-line Controller Tuning," *AIChE Journal*, 28, 434-440.

Ziegler, J.G., and Nichols, N.B., (1942) "Optimum Settings for Automatic Controllers," *Transactions of ASME*, 64, 759-768.

CHAPTER III

ROBUST SYNTHESIS OF NANOSTRUCTURES

3.1 Introduction

Nanotechnology is the construction and use of functional structures designed from atomic or molecular scale with at least one characteristic dimension measured in nanometers (one nanometer = 10^{-9} meter, which is about 1/50,000 of the width of human hair). The size of these nanostructures allows them to exhibit novel and significantly improved physical, chemical, and biological properties, phenomena, and processes. Nanotechnology can provide unprecedented understanding about materials and devices and is likely to impact many fields. By using structure at nanoscale as a tunable physical variable, scientists can greatly expand the range of performance of existing chemicals and materials. Alignment of linear molecules in an ordered array on a substrate surface (self-assembled monolayers) can function as a new generation of chemical and biological sensors. Switching devices and functional units at nanoscale can improve computer storage and operation capacity by a factor of a million. Entirely new biological sensors facilitate early diagnostics and disease prevention of cancers. Nanostructured ceramics and metals have greatly improved mechanical properties, both in ductility and strength.

Current research by nanoscientists typically focuses on novelty, discovering new growth phenomena and new morphologies. However, within the next five years there will likely be a shift in the nanotechnology community towards controlled and large-scale synthesis with high yield and reproducibility. This transition from laboratory-level synthesis to large scale, controlled and designed synthesis of nanostructures necessarily demands systematic investigation of the manufacturing conditions under

which the desired nanostructures are synthesized *reproducibly*, in *large* quantity and with *controlled* or isolated morphology. Application of statistical techniques can play a key role in achieving these objectives. This chapter reports a systematic study on the growth of 1D CdSe nanostructures through statistical modeling and optimization of the experimental parameters required for synthesizing desired nanostructures. This work is based on the experimental data presented in this paper and research published in Ma and Wang (2005). Some general statistical issues and research opportunities related to the synthesis of nanostructures are discussed in the concluding section.

Research in synthesizing semiconducting nanostructures is a forefront area in nanotechnology due to their applications in nanoelectronics, photonics, data storage, and sensing (Tolbert and Alivisatos 1994; Ma, Moore, Ding, Li and Wang 2004; Tran, Goldman, Anderson, Mauro and Mattoussi 2002). In particular, one-dimensional (1D) nanostructures present the ability to experimentally address the fundamental issues of reduced dimensionality and quantum confinement in one dimension (Lieber 1998; Alivisatos, Levinos, Steigerwald and Brus 1988). Cadmium selenide (CdSe) has been investigated over the past decade for applications in optoelectronics (Hodes, Albu-Yaron, Decker and Motisuke 1987), luminescent materials (Bawendi, Kortan, Steigerwald and Brus 1989), lasing materials (Ma, Ding, Moore, Wang and Wang 2004) and biomedical imaging. It is the most extensively studied quantum-dot material and is therefore regarded as the model system for investigating a wide range of nanoscale processes. CdSe is found to exhibit one-dimensional morphologies of nanowires, nanobelts and nanosaws (Ma and Wang 2005), often with the three morphologies being intimately intermingled within the as-deposited material. Images of these three nanostructures obtained using scanning electron microscope are shown in Figure 9.

In this experiment, the response is a vector whose elements correspond to the

numbers of appearance of different types of nanostructures and hence is a multinomial random variable. Thus a multinomial generalized linear model (GLM) is the appropriate tool for analyzing the experimental data and expressing the multinomial logits as functions of the predictor variables (McCullagh and Nelder 1989; Faraway 2006). A new iterative algorithm for fitting multinomial GLM that has certain advantages over the existing methods is proposed and implemented. The probability of obtaining each nanostructure is expressed as a function of the predictor variables. Owing to the presence of inner noise, i.e., variation around the set value, each predictor variable is a random variable. Using Monte-Carlo simulations, the expectation and variance of transformed probabilities are expressed as functions of the set points of the predictor variables. The expectation is then maximized to find the optimum set values of the process variables, ensuring at the same time that the variance is under control. The idea is thus similar to the two-step robust parameter design for larger-the-better responses (Wu and Hamada 2000, chap. 10).

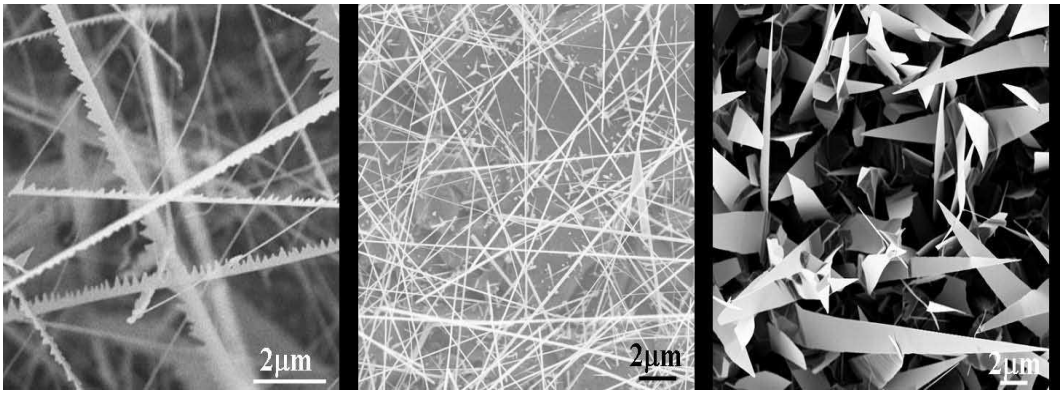


Figure 9: Nanosaws, Nanobelts and Nanowires

The chapter is organized as follows. In Section 3.2, we give a brief account of the synthesis process. In Section 3.3, the experimental design and collection of data are described. Section 3.4 is devoted to fitting of appropriate statistical models to the experimental data. This section consists of two subsections. In Section 3.4.1 a preliminary analysis using a binomial GLM is shown. Estimates of the parameters

obtained here are used as initial estimates in the iterative algorithm for multinomial GLM, which is developed and described in Section 3.4.2. In Section 3.5, we study the optimization of the process variables to maximize the expected yield of each nanostructure. Some general statistical issues and challenges in nanostructure synthesis and opportunities for future research are discussed in Section 3.6.

3.2 The synthesis process

The CdSe nanostructures were synthesized through a thermal evaporation process in a single zone horizontal tube furnace (Thermolyne 79300). A 30-inch polycrystalline Al_2O_3 tube (99.9% purity) with an inner diameter of 1.5 inches was placed inside the furnace. Commercial grade CdSe (Alfa Aesar, 99.995% purity, metal basis) was placed at the center of the tube as use for a source material. Single-crystal silicon substrates with a 2-nanometer thermally evaporated non-continuous layer of gold were placed downstream of the source in order to collect the deposition of the CdSe nanostructures. Water-cooled aluminum endcaps were used to seal the system as a mechanical roughing pump purged the system of oxygen. After the chamber had maintained a pressure of 2×10^{-2} torr for an hour, the system temperature was raised to a designated set point at a rate of 20^0 C/min and a nitrogen carrier gas was sent through the system at a rate of 50 sccm. Although the primary function of the carrier gas was to transport the sublimated vapor to cooler regions of the furnace, the secondary function of the gas was to build up the initial pressure of the system as well as controlling the partial pressure of the vaporized source material. This ensured that the pressure of the system was constant throughout the entire synthesis process. The system was held at the set temperature and pressure for a period of 60 minutes and cooled to room temperature afterwards. The as-deposited products were characterized and analyzed by scanning electron microscopy (SEM) (LEO 1530 FEG), transmission electron microscopy (TEM) (Hitachi HF-2000 FEG at 200 kV).

180 individual nanostructures were counted from the deposition on each substrate.

A schematic diagram of the synthesis process is shown in Figure 10.

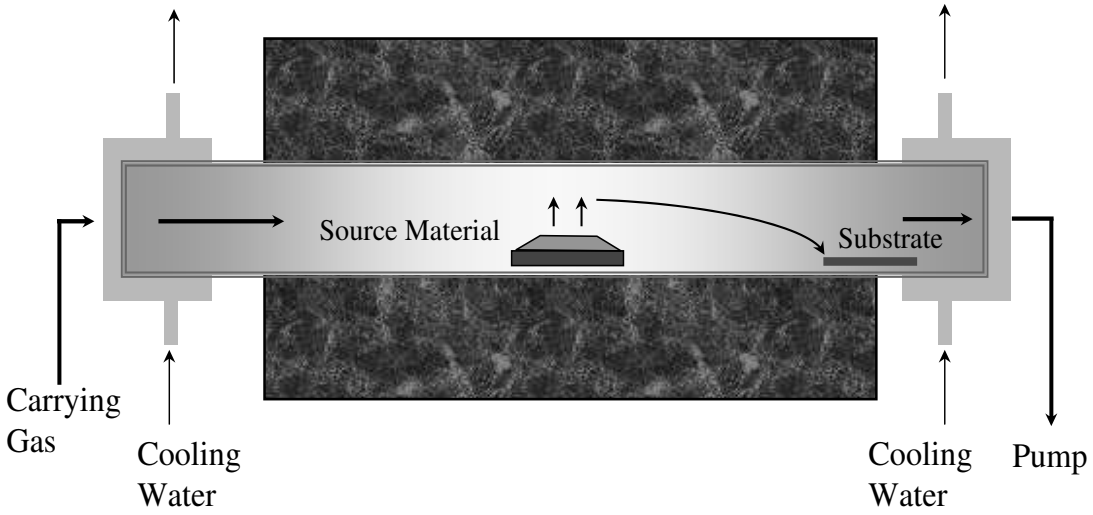


Figure 10: The Synthesis Process

3.3 Design of experiment and data collection

The two key process variables affecting morphology of CdSe nanostructures are temperature and pressure. A 5×9 full factorial experiment was conducted with five levels of source temperature (630, 700, 750, 800, 850^0 C) and nine levels of pressure (4, 100, 200, 300, 400, 500, 600, 700, 800 mbar). For a specific combination of source temperature and pressure, 4-6 substrates were placed downstream of the source to collect the deposition of nanostructures. The distance of the mid-point of the substrate from the source was measured and treated as a covariate.

Three experimental runs were conducted with each of the 45 combinations of temperature and pressure. However, these three runs cannot be considered to be replicates, since the number and location of substrates were not the same in the three runs. Consider, for example, the three runs performed with a temperature of 630^0 C and pressure of 4 mb. In the first run, six substrates were placed at distances of

1.9, 4.2, 4.9, 6.4, 8.1, 10.2 cm from the source. In the second run, four substrates were placed at distances of 1.7, 4.6, 7.1, 8.9 cm from the source. Seven substrates were placed at distances of 2.0, 4.3, 4.9, 6.4, 8.5, 10.6, 13.0 cm from the source in the third run. Therefore 17 (=6+4+7) individual substrates were obtained with the temperature and pressure combination of (630^0 C, 4 mb). Each of these 17 substrates constitute a row in Table 1. The total number of substrates obtained from the 135 ($=45 \times 3$) runs was 415. Note that this is not a multiple of 45 owing to an unequal number of substrates corresponding to each run.

Considering each of the 415 substrates as an experimental unit, the design matrix can thus be considered to be a 415×3 matrix, where the three columns correspond to source temperature ($TEMP$), pressure($PRES$) and distance from the source ($DIST$). Each row corresponds to a substrate, on which a deposition is formed with a specific combination of $TEMP, PRES$ and $DIST$ (see Table 1).

Recall that from the deposition on each substrate, 180 individual nanostructures were counted using SEM images. The response was thus a vector $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$, where Y_1, Y_2, Y_3, and Y_4 denote respectively the number of nanosaws, nanowires, nanobelts and no morphology, with $\sum_{j=1}^{4} Y_j = 180$. For demonstration purposes, the first 29 rows of the complete data are shown in Table 7. These rows correspond to the temperature-pressure combinations (630,4) and (630,100). The complete data can be downloaded from www.isye.gatech.edu/∼roshan.

It was observed that, at a source temperature of 850^0 C, almost no morphology was observed. Therefore, results obtained from the 67 experimental units involving this level of temperature were excluded and the data for the remaining 348 units were considered for analysis.

Henceforth, we shall use the suffixes 1,2,3 and 4 to represent quantities associated with nanosaws, nanowires, nanobelts and no growth respectivly.

Table 7: Partial data (29 rows out of 415) obtained from the nano-experiment

Temperature	Pressure	Distance	Nanosaws	Nanowires	Nanobelts	No growth
630	4	12.4	0	0	0	180
630	4	14.7	74	106	0	0
630	4	15.4	59	121	0	0
630	4	16.9	92	38	50	0
630	4	18.6	0	99	81	0
630	4	20.7	0	180	0	0
630	4	12.2	50	94	36	0
630	4	15.1	90	90	0	0
630	4	17.6	41	81	58	0
630	4	19.4	0	121	59	0
630	4	12.5	49	86	45	0
630	4	14.8	108	72	0	0
630	4	15.4	180	0	0	0
630	4	16.9	140	40	0	0
630	4	19.0	77	47	56	0
630	4	21.1	0	88	92	0
630	4	23.5	0	0	0	180
630	100	12.1	0	0	0	180
630	100	15.8	92	74	0	14
630	100	18.4	0	180	0	0
630	100	20.1	0	0	0	180
630	100	12.3	0	92	88	0
630	100	15.0	14	144	22	0
630	100	17.1	31	113	36	0
630	100	19.5	0	0	0	180
630	100	12.1	85	59	36	0
630	100	15.4	65	74	41	0
630	100	18.0	0	180	0	0
630	100	19.9	0	0	0	180

3.4 Model fitting

3.4.1 Individual modeling of the probability of obtaining each nanostructure using binomial GLM

Here, the response is considered binary, depending on whether we get a specific nanostructure or not. Let p_1, p_2 and p_3 denote respectively the probabilities of getting a nanosaw/nanocomb, nanowire and nanobelt. Then, for $j = 1, 2, 3$, the marginal distribution of Y_j is binomial with $n = 180$ and probability of success p_j. The log-odds ratio of obtaining the j^{th} type of morphology is given by

$$\zeta_j = \log \frac{p_j}{1 - p_j}.$$

Our objective is to fit a model that expresses the above log-odds ratios in terms $TEMP$, $PRES$ and $DIST$.

From the main effects plot of $TEMP$, $PRES$ and $DIST$ against observed proportions of nanosaws, nanowires and nanobelts (Figure 11), we observe that a quadratic

model should be able to express the effect of each variable on p_j adequately. The interaction plots (not shown here) give a preliminary impression that all the three two-factor interactions are likely to be important. We therefore decide to fit a quadratic response model to the data.

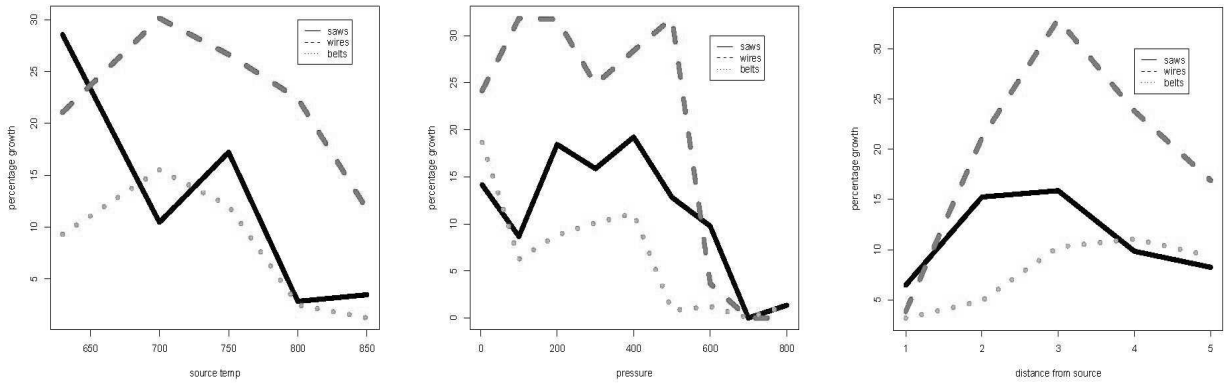


Figure 11: From left - growth vs temperature, growth vs pressure, growth vs distance

Each of three process variables are scaled to [-1,1] by appropriate transformations. Let T, P and D denote the scaled variables obtained by transforming $TEMP$, $PRES$ and $DIST$ respectively.

Using a binomial GLM with a logit link (McCullagh and Nelder, 1989), we obtain the following models that express the log-odds ratios of getting a nanosaw/nanocomb, nanowire and nanobelt as functions of T, P, D :

$$\begin{aligned}\hat{\zeta}_1 = \; & - \; 0.99 - 0.29 \;\; T - 1.52 \;\; P - 2.11 \;\; D - 0.95 \;\; T^2 - 1.30 \;\; P^2 - 5.64 \;\; D^2 \\ & - \; 0.18 \;\; TP - 1.03 \;\; PD + 4.29 \;\; TD, \end{aligned} \tag{1}$$

$$\begin{aligned}\hat{\zeta}_2 = \; & - \; 0.56 + 0.82 \;\; T - 2.53 \;\; P - 1.59 \;\; D - 0.58 \;\; T^2 - 2.04 \;\; P^2 - 2.62 \;\; D^2 \\ & + \; 1.17 \;\; TP - 1.44 \;\; PD + 0.87 \;\; DT, \end{aligned} \tag{2}$$

$$\hat{\zeta}_3 = - 1.68 + 0.19\ T - 1.88\ P - 0.58\ D - 1.69\ T^2 - 0.34\ P^2 - 3.20\ D^2$$
$$+ 0.87\ TP - 0.94\ PD - 2.58\ TD. \quad (3)$$

All the terms are seen to be highly significant. The residual plots for all the three models do not exhibit any unusual pattern.

3.4.2 Simultaneous modeling of the probability vector using multinomial GLM

Denoting the probability of not obtaining any nanostructure by p_4, we must have $\sum_{j=1}^{4} p_j = 1$. Although the results obtained by using the binomial GLM are easily interpretable and useful, the method suffers from the inherent drawback that, for specific values of T, P and D, the fitted values of the probabilities may be such that $\sum_{j=1}^{3} p_j > 1$. This is due to the fact that the *correlation* structure of $\mathbf{Y}$ is completely ignored in this approach.

A more appropriate modeling strategy is to utilize the fact that the response vector $\mathbf{Y}$ follows a multinomial distribution with $n = 180$ and probability vector $\mathbf{p} = (p_1, p_2, p_3, p_4)$. In this case, one can express the multinomial logits $\eta_j = \log(\frac{p_j}{p_4}), j = 1, 2, 3$ as functions of T, P and D. Note that η_j can be easily interpreted as the log-odds ratio of obtaining the j^{th} morphology as compared to no nanostructure, with $\eta_4 = 0$.

Methods for fitting multinomial logistic models by maximizing the multinomial likelihood have been discussed by several authors (McCullagh and Nelder 1989, Aitkin, Anderson, Francis and Hinde 1989, Agresti 2002, Faraway 2006, Long and Freese 2006). These methods have been implemented in several software packages like R/S-plus (multinom function), STATA (.mlogit function), LIMDEP (Mlogit$ function), SAS (CATMOD function) and SPSS (Nomreg function). All of these functions use some algorithm for maximization of the multinomial likelihood (e.g., the multinom

function in R/S-plus uses the neural network based optimizer provided by Venebles and Ripley (2002)). They produce more or less similar outputs, the default output generally consisting of the model coefficients, their standard errors and z-values, and model deviance.

Another popular algorithm to indirectly maximize the multinomial likelihood is to create a pseudo factor with a level for each data point, and use a Poisson GLM with log link. This method, although appropriate for small data sets, becomes cumbersome when the number of data points is large. In the presence of a large number of levels of the pseudo factor, a large part of the output generated by standard statistical softwares like R becomes redundant, because only the terms involving interaction between the categories and the predictor variables are of interest. Faraway (2006) points out some practical inconveniences of using this method. Its application to the current problem clearly becomes very cumbersome owing to the large number (348) of data points.

We propose a new iterative method of fitting multinomial logit models. The method is based on an iterative application of binomial GLMs. Besides the intuitive extension of binomial GLMs to a multinomial GLM, the method has certain advantages over the existing methods which are described towards the end of the section.

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i4})$ denote the response vector corresponding to the i^{th} data point, $i = 1$ to N. Let $n_i = \sum_{j=1}^{4} Y_{ij}$. Here, $N = 348$ and $n_i = n = 180$ for all i. We have,

$$P(Y_{i1} = y_{i1}, \ldots, Y_{i4} = y_{i4}) = \frac{n_i!}{y_{i1}! \ldots y_{i4}!} p_{i1}^{y_{i1}} \ldots p_{i4}^{y_{i4}}.$$

Thus the likelihood function is given by

$$\begin{aligned} L(\mathbf{Y}_1, \ldots, \mathbf{Y}_N) &= \prod_{i=1}^{N} \frac{n_i!}{y_{i1}! \ldots y_{i4}!} p_{i1}^{y_{i1}} \ldots p_{i4}^{y_{i4}} \\ &= \prod_{i=1}^{N} \frac{n_i!}{y_{i1}! \ldots y_{i4}!} \prod_{j=1}^{3} \left(\frac{p_{ij}}{p_{i4}}\right)^{y_{ij}} p_{i4}^{\sum_{j=1}^{4} y_{ij}}. \end{aligned}$$

Defining $\eta_{ij} = \log \frac{p_{ij}}{p_{i4}}$, we have

$$p_{ij} = \frac{\eta_{ij}}{1 + \sum_{j=1}^{3} \exp(\eta_{ij})} \quad j = 1, 2, 3, \tag{4}$$

and

$$p_{i4} = \frac{1}{1 + \sum_{j=1}^{3} \exp(\eta_{ij})}. \tag{5}$$

Therefore the log-likelihood can be written as

$$\begin{aligned} \log(L) &= \sum_{i=1}^{N} \left(\log n_i! - \sum_{j=1}^{4} \log y_{ij}! + \sum_{j=1}^{3} y_{ij} \log \frac{p_{ij}}{p_{i4}} + n_i \log p_{i4} \right) \\ &= \sum_{i=1}^{N} \left(\log n_i! - \sum_{j=1}^{4} \log y_{ij}! + \sum_{j=1}^{3} y_{ij} \eta_{ij} - n_i \log \left(1 + \sum_{j=1}^{3} \exp(\eta_{ij})\right) \right). \end{aligned} \tag{6}$$

Let $\mathbf{x}_i = (1, T_i, P_i, D_i, T_i^2, P_i^2, D_i^2, T_i P_i, P_i D_i, T_i D_i)', i = 1, \ldots, N$. The objective is to express the η's as functions of $\mathbf{x}$. Substituting $\eta_{ij} = \mathbf{x}_i' \boldsymbol{\beta}_j$ in (6) and successively differentiating with respect to each $\boldsymbol{\beta}_j$, we get the maximum likelihood (ML) equations as

$$\sum_{i=1}^{N} \mathbf{x}_i \left(y_{ij} - n_i \frac{\exp(\eta_{ij})}{1 + \sum_{j=1}^{3} \exp(\eta_{ij})} \right) = \mathbf{0}, \quad j = 1, 2, 3, \tag{7}$$

$$\sum_{i=1}^{N} \mathbf{x}_i \left(y_{14} - n_i \frac{1}{1 + \sum_{j=1}^{3} \exp(\eta_{ij})} \right) = \mathbf{0}, \tag{8}$$

where $\mathbf{0}$ denotes a vector of zeros having length 10. Writing $\exp(\gamma_{il}) = \left(1 + \sum_{l \neq j} \exp(\eta_{ij})\right)^{-1}$, we obtain from (7)

$$\sum_{i=1}^{N} \mathbf{x}_i \left(y_{ij} - n_i \frac{\exp(\eta_{ij} + \gamma_{ij})}{1 + \exp(\eta_{ij} + \gamma_{ij})} \right) = \mathbf{0}, \ \ j = 1, 2, 3. \tag{9}$$

Note that each equation in (9) is the maximum likelihood (ML) equation of a binomial GLM with logit link. Thus, if some initial estimates of $\boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are available, and consequently γ_{i1} can be computed, then $\boldsymbol{\beta}_1$ can be estimated by fitting a binomial GLM of Y_1 on $\mathbf{x}$. Similarly, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ can be estimated. The following algorithm is thus proposed.

Binomial GLM-based iterative algorithm for fitting a multinomial GLM :

Let $\boldsymbol{\beta}_j^{(k)}$ be the estimate of $\boldsymbol{\beta}_j, j = 1, 2, 3$, at the end of the k^{th} iteration.

Step 1. Using $\boldsymbol{\beta}_2^{(k)}$ and $\boldsymbol{\beta}_3^{(k)}$, compute $\eta_{i2}^{(k)} = \mathbf{x}_i'\boldsymbol{\beta}_2^{(k)}$ and $\eta_{i3}^{(k)} = \mathbf{x}_i'\boldsymbol{\beta}_3^{(k)}$ for $i = 1, \ldots, n$.

Step 2. Compute $\gamma_{i1}^{(k)} = \log \frac{1}{1+\exp(\eta_{i2}^{(k)})+\exp(\eta_{i3}^{(k)})}, \quad i = 1, \ldots, n.$

Step 3. Treating Y_1 as the response and using the same design matrix, fit a binomial GLM with logit link. The vector of coefficients thus obtained is $\boldsymbol{\beta}_1^{(k+1)}$.

Step 4. Repeat steps 1-3 by successively updating γ_{i2} and γ_{i3} and estimating $\boldsymbol{\beta}_2^{(k+1)}$ and $\boldsymbol{\beta}_3^{(k+1)}$

Repeat steps 1-4 until convergence. The convergence of the algorithm is established by Theorem 1 stated later. Note that we use the 'offset' command in statistical software R to separate the coefficients associated with η_1 from those with γ_1.

To obtain the initial estimates $\widehat{\eta}_{i2}^{(0)}$ and $\widehat{\eta}_{i3}^{(0)}$, we use the results obtained from the binomial GLM as described in Section 4.1. Let

$$\log \frac{\widehat{p}_{ij}}{1 - \widehat{p}_{ij}} = \mathbf{x}_i'\widehat{\boldsymbol{\delta}}_j, \tag{10}$$

where $\widehat{\boldsymbol{\delta}}_j$ is obtained by using binomial GLM. Recalling the definition of η_{ij}, the initial estimates are obtained as

$$\widehat{\eta}_{ij}^{(0)} = \log \frac{\widehat{p}_{ij}}{1 - \sum_{l=1}^{3} \widehat{p}_{il}}, \quad j = 2, 3, \tag{11}$$

where $\widehat{p}_{il}, l = 1, 2, 3$ are estimated from (10). It is possible, however, that for some i, $\sum_{l=1}^{3} \widehat{p}_{il} = \pi_i \geq 1$. For those data points, we provide a small correction as follows:

$$\widehat{p}_{il}^{c} = \begin{cases} \frac{\widehat{p}_{il}}{\pi_i}(1 - \frac{1}{2n_i}) & , l = 1, 2, 3 \\ \frac{1}{2n_i} & , l = 4 \end{cases} .$$

where $\widehat{p}_{il}^{c}$ denotes the corrected estimated probability. To justify the correction, we note that it is a common practice to give a correction of $\frac{1}{2n_i}$ (Cox 1970, chap. 3) in estimation of probabilities from binary data. The correction given to category 4 is adjusted among the other three categories in the same proportion as the estimated probabilities. This ensures that $\widehat{p}_{il} > 0$ for all i and $\sum_{l=1}^{4} \widehat{p}_{il} = 1$

In this example, there were 18 data points (out of 348) corresponding to which we had $\sum_{l=1}^{3} \widehat{p}_{il} \geq 1$. Following the procedure described above to obtain the initial estimates, the following models were obtained after convergence:

$$\begin{aligned} \widehat{\eta_1} = \quad & 0.42 - 0.12\ T - 3.08\ P - 3.68\ D - 1.84\ T^2 - 1.52\ P^2 - 9.09\ D^2 \\ + \quad & 0.60\ TP - 2.31\ PD + 5.75\ TD, \end{aligned} \tag{12}$$

$$\begin{aligned} \widehat{\eta_2} = \quad & 0.54 + 0.88\ T - 3.85\ P - 3.13\ D - 1.21\ T^2 - 2.28\ P^2 - 5.26\ D^2 \\ + \quad & 1.83\ TP - 2.62\ PD + 2.07\ TD, \end{aligned} \tag{13}$$

$$\begin{aligned} \widehat{\eta_3} = \ - \quad & 0.10 + 0.39\ T - 3.67\ P - 2.51\ D - 2.51\ T^2 - 1.12\ P^2 - 7.07\ D^2 \\ + \quad & 1.72\ TP - 2.38\ PD + 4.47\ TD. \end{aligned} \tag{14}$$

Inference for the proposed method:

To test the significance of the terms in the model, one can use the asymptotic normality of the maximum likelihood estimates. Let $\mathbf{H}_\beta$ denote the 30×30 matrix consisting of the negative expectations of second-order partial derivatives of the log-likelihood function in (6), the derivatives being taken with respect to the components

of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. Denoting the final estimator of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^*$, the estimated asymptotic variance-covariance matrix of the estimated model coefficients is given by $\boldsymbol{\Sigma}_{\beta^*} = \mathbf{H}_{\beta^*}{}^{-\mathbf{1}}$. For a specific coefficient β_l, the null hypothesis $H_0 : \beta_l = 0$ can be tested using the test statistic $z = \widehat{\beta}_l / s(\widehat{\beta}_l)$, where $s^2(\widehat{\beta}_l)$ is the l^{th} diagonal element of $\boldsymbol{\Sigma}_{\beta^*}$. The following theorem (proof in Appendix A) is useful in this context:

Theorem 3.1. *Let $L_M(\boldsymbol{\beta}, \mathbf{y})$ and $L_B(\boldsymbol{\beta}, \mathbf{y})$ denote the multinomial and binomial likelihood respectively. Let $\boldsymbol{\beta}^{(k)}$ denote the estimator of $\boldsymbol{\beta}$ after the kth iteration and $\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} log L_M(\boldsymbol{\beta}, \mathbf{y})$. Also, let*

$$\boldsymbol{\Sigma}_{\beta^{(\mathbf{k})}} = -\mathbf{E}\Big(\frac{\partial^{\mathbf{2}} \log \mathbf{L_B}(\boldsymbol{\beta},\mathbf{y})}{\partial \boldsymbol{\beta}^{\mathbf{2}}}\Big)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(\mathbf{k})}} \text{ and } \boldsymbol{\Sigma}_{\beta^*} = -\mathbf{E}\Big(\frac{\partial^{\mathbf{2}} \log \mathbf{L_M}(\boldsymbol{\beta},\mathbf{y})}{\partial \boldsymbol{\beta}^{\mathbf{2}}}\Big)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

Then,

(a) $\boldsymbol{\beta}^{(k)} \longrightarrow \boldsymbol{\beta}^*$.

(b) $\boldsymbol{\Sigma}_{\beta^{(\mathbf{k})}} \longrightarrow \boldsymbol{\Sigma}_{\beta^*}$

Thus, as the parameter estimates converge to the maximum likelihood estimates, their standard errors also converge to the standard error of the MLE. This property of the proposed algorithm ensures that one does not have to spend any extra computational effort in judging the significance of the model terms. The binomial GLM function in R used in every iteration automatically tests the significance of the model terms, and the p-values associated with the estimated coefficients after convergence can be used for inference. Thus, the inferential procedures and diagnostic tools of the binomial GLM can easily be used in the multinomial GLM model. This is clearly an advantage of the proposed algorithm over existing methods. Further, the three models for nanosaws, nanobelts and nanowires can be compared using these diagnostic

tools. Such facilities are not available in the current implementation of other software packages.

In the fitted models given by (12)-(14), all the 30 coefficients are seen to be highly significant with p values of the order 10^{-6} or less. To check the model adequacy, we use the generalized R^2 statistic derived by Naglekerke (1991) defined as $R^2 = \big(1-\exp\left((D - D_{null})/n\right)\big)/\big(1-\exp\left(-D_{null}/n\right)\big)$, where D and D_{null} denote the residual deviance and the null deviance respectively. The R^2 associated with the models for nanosaws, nanowires and nanobelts are obtained as 61%, 50% and 76% respectively. This shows that the prediction error associated with the model for nanowires is the largest. This finding is consistent with the observation made by Ma and Wang (2005) that growth of nanowires is less restrictive compared to nanosaws and nanowires, and can be carried out over wide ranges of temperature and pressure.

However, the small p values may also arise from the fact that some improper variance is used in testing. This overdispersion may be attributed either to some correlation among the outcomes from a given run in the experiment or some unexplained heterogeneity within a group owing to the effect of some unobserved variable. A multitude of external noise factors in the system make the second reason a more plausible one. We re-perform the testing by introducing three dispersion parameters $\sigma_1^2, \sigma_2^2, \sigma_3^2$, which are estimated from the initial binomial fits using $\hat{\sigma}_j^2 = \chi_j^2/(N-10)$, where χ_j^2 denotes Pearson's χ^2 statistic for the jth nanostructure, $j = 1, 2, 3$ and $N - 10 = 338$ is the residual degrees of freedom. The estimated standard errors of the coefficients and the corresponding p values are shown in Table 8.

From Table 8 we find that the linear effect of temperature is not significant for nanosaws and nanobelts. However, since the quadratic term T^2 and the interactions involving T are significant, we prefer to retain T in the models.

Note that although techniques for analyzing overdispersed binomial data are well

Table 8: Computed values of the test statistic for each estimated coefficient

	Nanosaws ($\widehat{\eta_1}$)			Nanowires ($\widehat{\eta_2}$)			Nanobelts($\widehat{\eta_3}$)		
Term	$\hat{\beta}$	S.E.	p-val.	$\hat{\beta}$	S.E.	p-val.	$\hat{\beta}$	S.E.	p-val.
Intercept	0.42	0.24	0.0807	0.54	0.25	0.0343	-0.10	0.24	0.6763
T	-0.12	0.30	0.6855	0.88	0.28	0.0020	0.39	0.38	0.3125
P	-3.08	0.41	0.0000	-3.85	0.50	0.0000	-3.67	0.51	0.0000
D	-3.69	0.67	0.0000	-3.13	0.57	0.0000	-2.51	0.66	0.0001
T^2	-1.84	0.34	0.0000	-1.21	0.27	0.0000	-2.51	0.34	0.0000
P^2	-1.52	0.44	0.0006	-2.28	0.52	0.0000	-1.12	0.54	0.0381
D^2	-9.09	0.99	0.0000	-5.26	0.66	0.0000	-7.08	0.77	0.0000
TP	0.60	0.42	0.1515	1.83	0.41	0.0000	1.72	0.53	0.0011
PD	-2.31	0.83	0.0053	-2.62	0.70	0.0000	-2.38	0.84	0.0043
TD	5.75	0.80	0.0000	2.07	0.45	0.0000	4.47	0.69	0.0000

known (e.g., Faraway 2006, Ch. 2), methods for handling overdispersion in multinomial logit models are not readily available. The proposed algorithm provides us with a very simple heuristic way to do this and thereby has an additional advantage over the existing methods.

3.5 Optimization of the synthesis process

In the previous subsections, the three process variables have been treated as non-stochastic. However, in reality, none of these variables can be controlled precisely and each of them exhibits certain fluctuations around the set (nominal) value. Such fluctuation is a form of noise, called internal noise (Wu and Hamada 2000, chap. 10) associated with the synthesis process and needs to be considered in performing optimization.

It is therefore reasonable to consider $TEMP$, $PRES$ and $DIST$ as random variables. Let μ_{TEMP}, μ_{PRES}, μ_{DIST} denote the set values of $TEMP$, $PRES$ and $DIST$ respectively. Then we assume

$$TEMP \sim N(\mu_{TEMP}, \sigma^2_{TEMP}), PRES \sim N(\mu_{PRES}, \sigma^2_{PRES}),\ DIST \sim N(\mu_{DIST}, \sigma^2_{DIST}).$$

where σ^2_{TEMP}, σ^2_{PRES}, σ^2_{DIST} are the respective variances of $TEMP$, $PRES$ and $DIST$ around their set values and are estimated from process data (Section 5.1). The task now is to determine the optimal nominal values μ_{TEMP}, μ_{PRES} and μ_{DIST} so

that the expected yield of each nanostructure is maximized subject to the condition that the variance in yield is acceptable.

3.5.1 Measurement of internal noise in the synthesis process

Some surrogate process data collected from the furnace were used for estimation of the above variance components. Temperature and pressure were set at specific levels (those used in the experiment), and their actual values were measured repeatedly over a certain period of time. The range of temperature and pressure corresponding to each set value was noted. The variation in distance, which is due to repeatability and reproducibility errors associated with the measurement system, was assessed separately. The summarized data in Table 3 show the observed ranges of $TEMP$, $PRES$ and $DIST$ against different nominal values. Under the assumption of normality, the range can be assumed to be approximately equal to six times the standard deviation.

Table 9: Fluctuation of process parameters around set values

Temperature		Pressure		Distance	
Nominal value (μ_T)	Observed range ($\approx \pm 3\sigma_T$)	Nominal value (μ_P)	Observed range ($\approx \pm 3\sigma_P$)	Nominal value (μ_D)	Observed range ($\approx \pm 3\sigma_D$)
630	±7	4	±10	11	±0.02
700	±7	100	±10	13	±0.02
750	±6	200	±20	15	±0.02
800	±6	300	±20	17	±0.02
850	±6	400	±20	19	±0.02
		500	±40	21	±0.02
		600	±40		

We observe from Table 3 that, for the process variable $DIST$, the range of values around the nominal μ_{DIST} is a constant ($2 \times 0.02 = 0.04$ mm) and independent of μ_{DIST}. Equating this range to $6\sigma_{DIST}$, we obtain an estimate of σ_{DIST} as $0.04/6 = 0.067$ mm.

Similarly, for $TEMP$, the range can be taken to be almost a constant. Equating the mean range of 12.8 ($=2 \times (2 \times 7 + 3 \times 6)/5$) degrees to $6\sigma_{TEMP}$, an estimate of σ_{TEMP} is obtained as $12.8/6 = 2.13$ degree C.

The case of $PRES$ is, however, different. The range, and hence σ_{PRES} is seen to be

an increasing function of μ_{PRES}. Corresponding to each value of μ_{PRES}, an estimate of σ_{PRES} is obtained by dividing the range by 6. Using these values of σ_{PRES}, the following regression line is fitted through the origin to express the relationship between σ_{PRES} and μ_{PRES}

$$\sigma_{PRES} = 0.025\mu_{PRES}. \tag{15}$$

Recall that all the models are fitted with the transformed variables T, P, D. The means μ_T, μ_P, μ_D and the variances $\sigma_T^2, \sigma_p^2, \sigma_D^2$ can easily be expressed in terms of the respective means and variances of the original variables.

3.5.2 Obtaining the mean and variance functions of p_1, p_2, p_3

From (4), we have the estimated probability functions as $\widehat{p_j} = \exp(\widehat{\eta_j})/\big(1+\sum_{j=1}^{3}\exp(\widehat{\eta_j})\big)$, where $\widehat{\eta_j}$ are given by (12)-(14).

Expressing $E(p_j)$ and $Var(p_j)$ in terms of μ_T, μ_P, μ_D is not a straightforward task. To do this, we use Monte Carlo simulations. For each of the 180 combinations of $\mu_{TEMP}, \mu_{PRES}, \mu_{DIST}$ ($\mu_{TEMP} = 630, 700, 750, 800$; $\mu_{PRES} = 4, 100, 200, \ldots, 800$; $\mu_{DIST} = 12, 14, 16, 18, 20$) the following are done:

1. μ_T, μ_P and μ_D are obtained by appropriate transformation.

2. 5000 observations on T, P and D are generated from the respective normal distributions and η_j is obtained using equation (12), (13) or (14).

3. From the η_j values thus obtained, p_j's are computed using (4) and transformed to $\zeta_j = log\frac{p_j}{1-p_j}$.

4. The mean and variance of those 5000 ζ_j values (denoted by $\overline{\zeta_j}$ and $s^2(\zeta_j)$ respectively) are computed.

5. Using linear regression, $\bar{\zeta_j}$ and $\log s^2(\zeta_j)$ are expressed in terms of μ_T, μ_P and μ_D.

3.5.3 Maximizing the average yield

Since the response here is of larger-the-better type, maximizing the mean is more important than minimizing the variance in the two-step optimization procedure (Wu and Hamada 2000, chap. 10) associated with robust parameter design.

The problem can thus be formulated as :

$$\text{maximize } \overline{\zeta_j} \qquad \text{subject to}$$

$$-1 \leq \mu_T \leq 1, \;\; -1 \leq \mu_P \leq 1, \;\; -1 \leq \mu_D \leq 1 \quad \text{for} \quad j = 1, 2, 3.$$

Physically, this would mean maximizing the average log-odds ratio of getting a specific morphology.

The following models are obtained from the simulated data:

$$\begin{aligned} \overline{\zeta_1} = \; & - \; 0.75 + 0.20\mu_T - 1.02\mu_P - 1.39\mu_D - 1.50\mu_T^2 - 3.54\mu_P^2 - 11.02\mu_D^2 \\ & + \; 1.58\mu_T\mu_P - 2.22\mu_P\mu_D + 8.41\mu_T\mu_D, \end{aligned} \tag{16}$$

$$\begin{aligned} \overline{\zeta_2} = \; & - \; 0.40 + 0.80\mu_T - 2.96\mu_P - 1.43\mu_D - 0.98\mu_T^2 - 2.45\mu_P^2 - 6.05\mu_D^2 \\ & + \; 1.87\mu_T\mu_P - 3.41\mu_P\mu_D + 2.13\mu_T\mu_D, \end{aligned} \tag{17}$$

$$\begin{aligned} \overline{\zeta_3} = \; & - \; 1.25 + 0.26\mu_T - 2.6\mu_P - 0.42\mu_D - 2.36\mu_T^2 - 1.24\mu_P^2 - 8.03\mu_D^2 \\ & + \; 1.74\mu_T\mu_P - 3.32\mu_P\mu_D + 4.57\mu_T\mu_D. \end{aligned} \tag{18}$$

Maximizing these three functions using the *optim* command in R, we get the optimal conditions for maximizing the expected yield of nanosaws, nanowires and nanobelts in terms of μ_T, μ_P and μ_D. These optimal values are transformed to the original units (i.e., in terms of μ_{TEMP}, μ_{PRES}, and μ_{DIST}) and are summarized in Table 4.

Contour plots of average and variance of the yield probabilities of nanosaws, nanowires and nanobelts against temperature and pressure (at optimal distances)

Table 10: Optimal process conditions for maximizing expected yield of nanostructures

Nanostructure	Temperature	Pressure	Distance
Nanosaws	630	307	15.1
Nanowires	695	113	19.0
Nanobelts	683	4	17.0

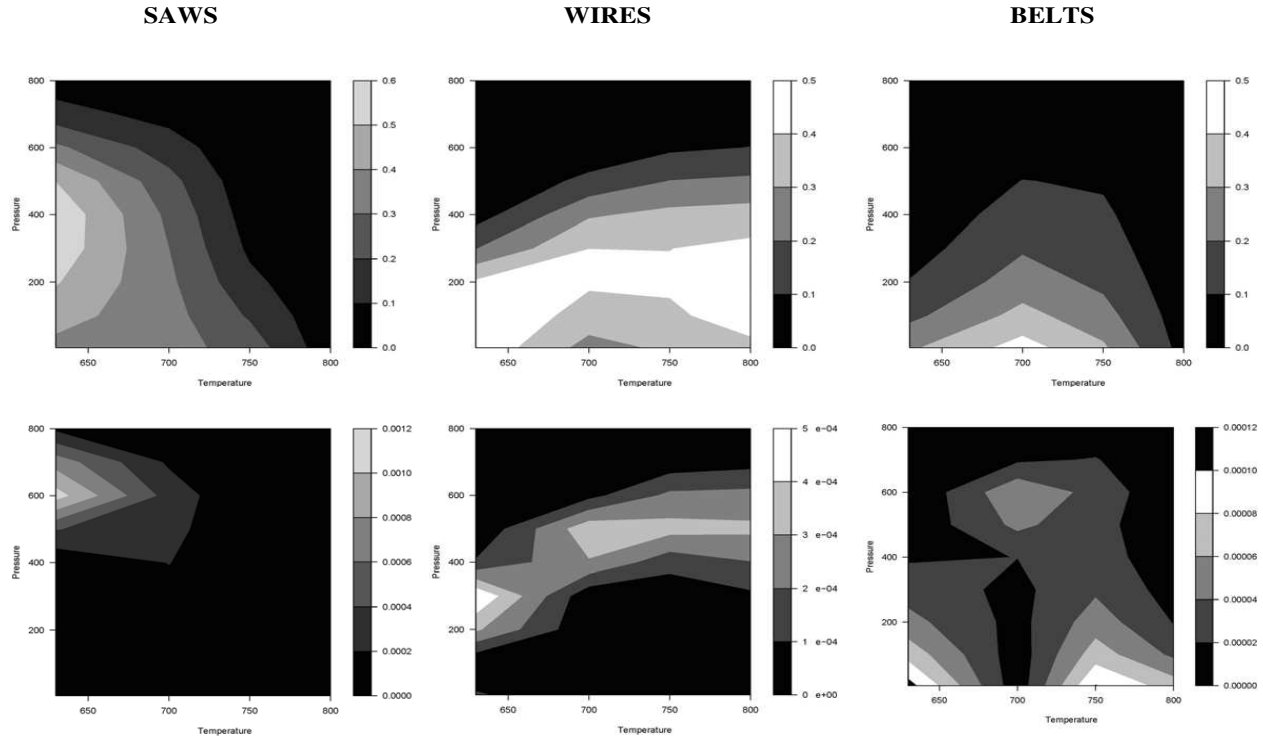


Figure 12: Contour Plots

are shown in Figure 13. The white regions on the top (average) panels and the black regions on the bottom (variance) panels are robust regions that promote high yield with minimal variation.

On the basis of these contour plots and the optimization output summarized in Table 4, the following conclusions can be drawn:

1. For nanosaws, the process is seen to be fairly robust at a pressure below 400 mb, irrespective of source temperature.

2. For nanobelts, temperature affects robustness strongly, and for a pressure of less than 400 mb, the process is very robust only when the temperature is in close proximity of 700^0 C.

3. A temperature of around 630 degrees and pressure of 310 mb simultaneously maximizes the average and minimize the variance of probability of obtaining nanosaws.

4. A temperature of around 700 degrees and pressure of around 120 degrees results in highest average yield for nanowires. Low variance is also observed in this region.

5. Highest yield of nanobelts is achieved at a temperature of 680 degrees and pressure of 4 mb for nanobelts. This is also a low-variance region.

6. There is a large temperature-pressure region (white region in the top-middle panel of Figure 4) that promotes high and consistent yield of nanowires.

7. Highest yields of nanobelts and nanowires are achieved at higher distance (i.e., lower local temperature) as compared to nanosaws.

Except for the robustness-related conclusions, most of the above findings are summarized and discussed by Ma and Wang (2005). They also provide plausible and in-depth physical interpretations of some of the above phenomena.

3.6 Some general statistical issues in nanomaterial synthesis and scope for future research

In this chapter, we report an early application of statistical techniques in nanomaterial research. In terms of reporting results of experiments to synthesize nanostructures, this methodology can be considered a significant advancement over the rudimentary data analysis methods using simple graphs, charts and summary statistics (e.g., Song, Wang, Riedo and Wang 2005; Ma and Wang 2005) that have been reported in the nanomaterial literature so far. Here we discuss features of the data arising from a specific experiment and use a multinomial model to express the probabilities of three different morphologies as functions of the process variables. A new iterative algorithm which is more appropriate than conventional methods for the present problem, is proposed for fitting the multinomial model. Inner noise is incorporated into the fitted models and robust settings of process variables that maximize the expected yield of each type of nanostructure are determined.

Apart from the advantages discussed earlier in this paper and mentioned by Ma and Wang (2005), this study demonstrates how statistical techniques can help in identifying important higher-order effects (like quadratic effects or complex interactions among the process variables) and utilize such knowledge in fine-tuning the optimal synthesis conditions. This work is also an important step towards large-scale controlled synthesis of CdSe nanostructures, since in addition to determining conditions for high yield, it also identifies robust settings of the process variables that are likely to guarantee consistent output.

Although statistical design of experiments (planning, analysis and optimization) have been applied very successfully to various other branches of scientific and engineering research to determine high-yield and reproducible process conditions, its application in nanotechnology has been limited till date. Some unique aspects associated with the synthesis of nanostructures that make the application of the above

techniques in this area challenging are: (i) complete disappearance of nanostructure morphology with slight changes in process conditions; (ii) complex response surface with multiple optima, making exploration of optimal experimental settings very difficult (although in the current case study, a quadratic response surface was found more or less adequate, such is not the case in general); (iii) different types of nanostructures (saws, wires, belts) intermingled; (iv) categorical response variables in most cases; (v) functional inputs (control factors that are functions of time); (vi) a multitude of internal and external noise factors heavily affecting reproducibility of experimental results; and (vii) expensive and time consuming experimentation. In view of the above phenomena, the following are likely to be some of the major statistical challenges in the area of nanostructure synthesis :

1. *Developing a sequential space-filling design for maximization of yield* : Fractional factorial designs and orthogonal arrays are the most commonly used (Wu and Hamada 2000) designs, but are not suitable for nanomaterial synthesis, because the number of runs becomes prohibitively large as the number of levels increases. Moreover, they do not facilitate sequential experimentation, which is necessary to keep the run size to a minimum. Response surface methodology (Myers and Montgomery 2002) may not be useful because the underlying response surface encountered in nano-research can be very complex with multiple local optima, and the binary nature of data adds to the complexity. Space-filling designs like Latin hypercube designs, uniform designs, and scrambled nets are highly suitable for exploring complex response surfaces with minimum number of runs. They are now widely used in computer experimentation (Santner, Williams and Notz 2003). However, they are used in the literature for only one-time experimentation. We need designs that are model independent, quickly "carve out" regions with no observable nanostructure morphology, allow for the exploration of complex response surfaces, and can be used for sequential

experimentation.

2. *Developing experimental strategies where one or more of the control variables is a function of time* : In experiments for nanostructure synthesis, there are factors whose profiles or curves with respect to time are often crucial with respect to the output. For example, although the peak temperature is a critical factor, how this temperature is attained over time is very important. There is an ideal curve that is expected to result in the best performance. Planning and analysis of experiments with such factors (which are called *functional factors*) are not much discussed in the literature and may be an important topic for future research.

3. *Scale up* : One of the important future tasks of the nanomaterial community is to develop industrial-scale manufacture of the nanoparticles and devices that are rapidly being developed. This transition from laboratory-level synthesis to large scale, controlled and designed synthesis of nanostructures poses plenty of challenges. The key issues to be addressed are rate of production, process capability, robustness, yield, efficiency and cost. The following specific tasks may be necessary: (1) Deriving specifications for key quality characteristics of nanostructures based on intended usage. Quality loss functions (Joseph 2004) may be used for this purpose. (2) Expanding the laboratory set-up to simulate additional conditions that are likely to be present in an industrial process. (3) Conducting experiments and identifying robust settings of the process variables that will ensure manufacturing of nanostructures of specified quality with high yield. (4) Statistical analysis of experimental data to compute the capability of the production process with respect to each quality characteristic.

3.7 References

Aitkin, M., Anderson, D.A., Francis, B.J. and Hinde, J.P. (1989), *Statistical Modelling in GLIM*, Clarendon Press, Oxford.

Agresti, A. (2002), *Categorical Data Analysis*, New York: Wiley.

Alivisatos, A.P., Harris, A.L., Levinos, N.J., Steigerwald, M.L. and Brus, L.E. (1988), "Electronic States of Semiconductor Clusters: Homogeneous and Inhomogeneous Broadening of the Optical Spectrum," *Journal of Chemical Physics*, 89, 4001.

Bawendi, M.G., Kortan, A.R., Steigerwald, M.L., and Brus, L.E. (1989), "X-ray Structural Characterization of Larger Cadmium Selenide (CdSe) Semiconductor Clusters," *Journal of Chemical Physics*, 91, 7282-7290.

Cox, D.R. (1970), *Analysis of Binary Data*, London : Chapman & Hall.

Faraway, J.J. (2006), *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Boca Raton : Chapman & Hall/CRC.

Hodes, G., Albu-Yaren, A., Decker, F., Motisuke, P. (1987), "Three-Dimensional Quantum-Size Effect in Chemically Deposited Cadmium Selenide Films," *Physics Review B*, 36, 4215-4221.

Joseph (2004), "Quality Loss Functions for Nonnegative Variables and Their Applications," Journal of Quality Technology 36, 129-138.

Lieber, C.M. (1998), "One-Dimensional Nanostructures: Chemistry, Physics and Applications," *Solid State Communications*, 107, 607-616.

Ma, C., Moore, D.F., Ding, Y., Li, J. and Wang, Z.L. (2004), "Nanobelt and Nanosaw Structures of II-VI Semiconductors," *International Journal of Nanotechnology*, 1, 431-451.

Ma,C., Ding, Y., Moore, D.F., Wang, X. and Wang,Z.L. (2004), "Single-Crystal CdSe Nanosaws," *Journal of American Chemical Society*, 126, 708-709.

Ma, C. and Wang, Z.L. (2005), "Roadmap for Controlled Synthesis of CdSe Nanowires, Nanobelts and Nanosaws," *Advanced Materials*, 17, 1-6.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Myers, H. M. and Montgomery, D.C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, NewYork: Wiley.

Naglekerke, N. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691-692.

Song, J., Wang, X., Riedo, E. and Wang, Z.L. (2005), "Systematic Study on Experimental Conditions for Large-Scale Growth of Aligned ZnO Nanwires on Nitrides," *Journal of Physical Chemistry B*, 109, 9869-9872.

Santer, T. J., Williams, B. J. and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.

Tolbert, S.H. and A.P. Alivisatos, A.P. (1994), "Size Dependence of a First Order Solid-Solid Phase Transition: The Wurtzite to Rock Salt Transformation in CdSe Nanocrystals," *Science*, 265, 373-376.

Tran, P.T., Goldman, E.R., Anderson, G.P., Mauro, J.M. and Mattoussi, H. (2002),"Use of Luminescent CdSe-ZnS Nanocrystal Bioconjugates in Quantum Dot-Based Nanosensors,"*Physics Status Solidi B-Basic Research*, 229, 427-432.

Wu, C.F.J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Venables, W.N. and Ripley, B.D. (2000), *Modern Applied Statistics with S-PLUS*, New York: Springer.

CHAPTER IV

SEQUENTIAL MINIMUM ENERGY DESIGNS FOR SYNTHESIS OF NANOSTRUCTURES

4.1 Introduction

Nanowires of inorganic materials are emerging as remarkably powerful building blocks in nanoscience and nanotechnology with the potential to impact numerous areas ranging from electronics, photonics and optoelectronics to life sciences and healthcare (Hu, Odon and Lieber 1999, Lieber 2003, Li et al. 2006, Wang 2004, Samuelson 2003, Yang 2005, Patolsky and Lieber 2005, Patolsky, Zhang and Lieber 2006a, 2006b). Critical to the advances now being made with nanowires has been a reasonable understanding of the growth mechanism. Nanowires grown on two types of source materials - Cadmium Selenide (CdSe) and Zinc Oxide (ZnO) have been considered extremely important from the point of view of potential applications. Growth of high quality, aligned and uniform ZnO nanowires is essential for the newly discovered nanogenerators (Wang and Song 2006). The practical applications of such nanogenerators are huge (Wang et al. 2007). Wireless devices may allow in-situ, real-time biomedical monitoring and detection, but such devices still requires a power source. Ideally, such devices should be self-powered rather use a battery. The body provides numerous potential power sources - mechanical energy, vibrational energy, chemical energy (glucose), and hydraulic energy, but the challenge is their efficient conversion into electric energy. If accomplished on the nanoscale, such power sources could have greatly reduced the size of the integrated nanosystems for optoelectronics (Duan et al. 2003), biosensors (Zheng et al. 2005), resonators (Bai et al. 2003) and more

(Wang and Song 2006). CdSe is the most extensively studied quantum-dot material and is therefore regarded as the model system for investigating a wide range of nanoscale processes. Among the three types of CdSe nanostructures, nanowires are likely to be the most useful structure for practical device applications.

Experiments to determine the best process conditions that maximize the yield of nanowires are expensive and time consuming. An effort was made to systematically investigate the best process conditions for CdSe nanowires (see Dasgupta et al. 2007) using a 5×9 full-factorial experimental design involving two key process variables temperature and pressure. The optimum conditions were obtained after maximizing the fitted quadratic multinomial logit models with respect to the process variables. It was observed that, although the models for the other two types of nanostructures (nanosaws and nanobelts) fit the data reasonably well, the fit for nanowires was not as satisfactory (with a generalized R^2 of about 50%). Further, it was felt that the experimental effort spent in this project was prohibitive, and much more efficient experimental methods were necessary.

Table 11: Average percentage yield of nanowires

Temperature	Pressure								
	0.00	0.12	0.25	0.37	0.50	0.62	0.75	0.87	1.00
0.00	41.27	42.41	0.00	21.11	11.06	20.49	0.00	0.00	0.00
0.32	29.09	35.19	55.56	20.00	21.82	44.77	0.00	0.00	0.00
0.55	21.96	47.71	34.00	16.94	27.47	22.22	20.00	0.00	0.00
0.77	14.29	11.11	36.06	35.32	49.40	33.33	0.00	0.00	0.00
1.00	8.33	14.29	0.00	28.93	0.00	0.00	3.85	0.00	0.00

A phenomenon that makes the exploration of optima very difficult is the *complete disappearance of morphology* in several regions of the operable design space as well as the *presence of multiple optima*. This is illustrated in Figure 1, which shows a contour plot of average percentage yield (averaged over all substrates) of CdSe nanowires against temperature and pressure (both variables scaled to [0,1]). The figure is generated from the data shown in Table 11 obtained from the experimental

data reported in Dasgupta et al. (2006). Several points emerge from this figure :

1. There is a large no-yield (deep green) region in the entire design space.
2. There is a small region of no-yield (deep green) completely embedded within a region of moderate and high yields.
3. The regions of highest yield (white) are scattered at different places.

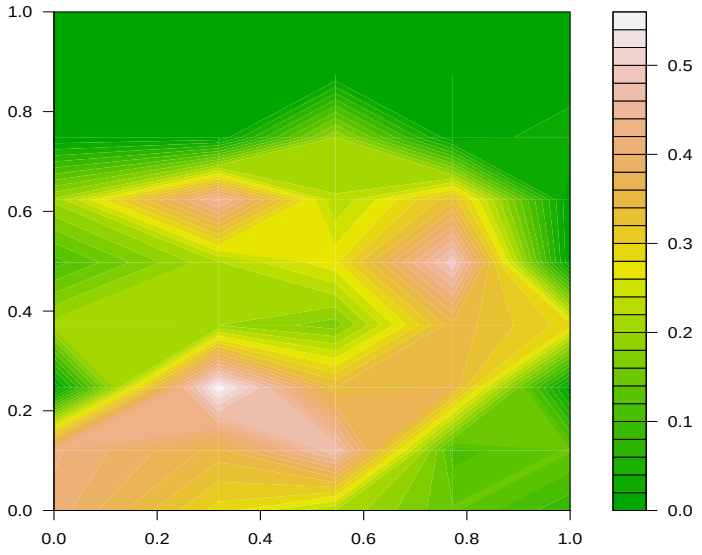


Figure 13: Contour plot for average yield of nanowires

These aspects of the problem make the application of design of experiments to find the maximum yield very challenging. Traditional experimental designs cannot be used in such situations due to the following reasons.

Fractional factorial designs and orthogonal arrays are the most commonly used (Wu and Hamada 2000) designs, but are not adequate for our purpose, because the number of runs becomes prohibitively large as the number of levels increases. Moreover, they do not facilitate sequential experimentation, which is necessary to keep the run size to a minimum. Response surface methodology (Myers and Montgomery 2002) may not be useful for our purpose because the underlying response surface

encountered in nano-research is very complex with multiple local optima. Moreover, the data being mostly categorical, and binary on some occasions, maximum likelihood estimates may not exist during the initial stages of experimentation (Silvapulle 1981), thereby prohibiting a model-driven approach. Space-filling designs like Latin hypercube designs (McKay, Beckman and Conover 1979) or uniform designs (Fang 2002), and scrambled nets (Owen 1995) can be used for exploring complex response surfaces with minimum number of runs. Now widely used in computer experimentation (Santner, Williams and Notz 2003, Fang et al. 2006), they are model independent and do not have problems associated with categorical or binary data. However, they are used in the literature for only one-time experimentation. Can they be used for sequential experimentation?

Therefore, the objective is to obtain designs that are model independent, can quickly "carve out" regions with no observable nanostructure morphology, allow for the exploration of complex response surfaces, and can be used for sequential experimentation. Thus, a major objective in this research is to develop a novel and efficient sequential space-filling design.

Because our aim is to apply these experimental designs for material research, we would like to propose a new procedure that will appeal to the nanoscientists who are mostly trained in physics, chemistry, or engineering. The new method should be easily understood by the nanoscientists and be well-grounded in statistical theory.

In Section 2, we review some available sequential procedures for global optimization of complex functions and discuss why they are not appropriate for our purpose. In Section 3, we introduce the new *minimum energy design principle* and investigate its performance for deterministic functions. A modified algorithm for random functions is proposed and illustrated in Section 4. Some concluding remarks and directions for future research are stated in Section 5.

4.2 Sequential design procedures for global optimization of complex multimodal functions

In global optimization, the use of stochastic processes is called Bayesian global optimization (Betro 1991). Motivated by problems arising from computer experiments, several researchers have tried to use this idea to develop sequential experimental strategies for computer experiments and global optimization of complex functions. Cox and John (1997) introduced a method called the Sequential Design for Optimization (SDO). Their approach uses a random function model, and lower confidence bounds on predicted values are used for sequential selection of evaluation points. Jones et al. (1998) proposed a sequential procedure based on the expected improvement criterion. The criterion balances the need to exploit the approximating surface (by sampling where the prediction is optimal) with the need to improve the approximation (by sampling where the prediction error may be high). Wang (2003) developed an adaptive response surface method that sequentially reduces the design space by defining a threshold value. At each step, a new Latin hypercube design is generated in the reduced design space and second order model is fitted. None of these approaches, however, guarantee an efficient optimization of the nanowire synthesis process owing to the following reasons:

1. As mentioned in the introductory section, there is a large no-yield (or failure) region in the design space. The boundaries of this failure region is unknown. Since the entire design space has to be explored, many zero yield data points have to be expected. Note that 17 out of the 45 design points in Table 11 result in zero yield. It is evident that all the approaches stated above would require a moderate number of experimental runs at the first stage of the procedure to fit a reasonable meta-model. For example, the SDO approach by Cox and John (1997) uses a 16-run design for two dimensional functions to build a meta-model at the initial stage. Figure 14 shows a 20-run Latin hypercube design

superimposed on the contour plot for the CdSe nanowire yield. It is seen that 9 out of the 20 points fall in the no-yield region, which is a huge waste from the point of view of the experimental effort involved (each run involves several steps spread over a number of days). Also, the estimates obtained from such data may not be very stable.

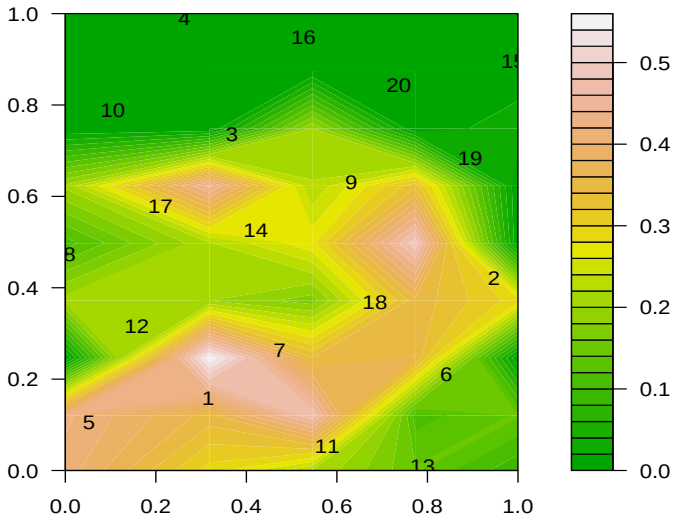


Figure 14: 20-run LHD for synthesis of nanowires

An adaptive sequential optimization procedure (ASOP) to generate a space-filling design that minimizes the number of no-yield points was developed by Henkenjohann et al. (2005). However, their method is based on two assumptions, convex boundaries and a connected non-failure region, which hold good for the motivating engineering problem of sheet metal spinning. None of the above two assumptions can be assumed to hold good for nanowire synthesis.

2. All the methods stated above work well with continuous responses. However, as already mentioned, in our case, the data is mostly categorical. The response yield is a binomial random variable, where the number of trials may range from very high (180 in the current example) to 1 (thereby generating binary data).

Estimation problems associated with such data may severely affect the search for optimum.

3. Unlike the global optimization problems and algorithms in computer experiments, here we are less concerned about the convergence of the algorithm to the true optimum. This is because, it is very difficult to describe the growth mechanism with a precise model, and in reality there will always be several noise variables involved. Therefore our principle objective is to explore quickly as much of the yield region as possible, and "carve out" regions of no morphology.

Almost all the sequential procedures mentioned earlier in this section use the idea of "sampling points from prospective regions that are more likely to improve performance". There are some sequential methods that work the opposite way. They explore the global optimum of black box functions by sequentially eliminating poor combinations of factor levels. Two such methods are SEL (Wu, Mao and Ma 1990) and SELC (Mandal, Wu and Johnson 2006). The latter method uses a genetic algorithm to improve upon the former. However, these methods are not appropriate for our purpose owing to the large number of runs that they usually require.

4.3 Sequential Minimum Energy Designs (SMED)

Each experimental run with a specific combination of process variables corresponds to a point in the design space and is called a design point. Visualize a box containing some charged particles. If they have the same sign for the charge, then they will repel each other and occupy positions inside the box so as to minimize the total potential energy. *Here, the box is the experimental region and the each position taken by the charged particles is a design point.* Consequently, the experimental design consists of all the positions occupied by the particles. Because this design is obtained by

minimizing the potential energy, we call it *minimum energy design*. The box can be in a high dimensional space. Let there be m factors. For simplicity, we scale the experimental range of each factor into $[0, 1]$, so that the box under consideration is $[0, 1]^m$.

The minimum energy designs are clearly a type of space-filling designs. They can be used for sequential experimentation as follows. Suppose we have n charged particles inside the box, which occupy certain positions according to the minimum potential energy criterion. Fix the charged particles at these locations and now introduce a new particle into the box. This new particle will now attain a position to minimize the total potential energy, giving us the $(n+1)^{th}$ design point. This follows from Thomson's theorem in electrodynamics (Zhou 1999, p. 59). We can intelligently choose the charge of each particle depending on the response observed at that design point. This will allow us to find the next design point by efficiently utilizing the already observed data.

Without loss of generality, assume that the particle charge is positive. Let $q(\mathbf{x}_i)$ be the charge of the particle at the ith design point $\mathbf{x}_i$. Then the potential energy between the ith and jth particle is proportional to $q(\mathbf{x}_i)q(\mathbf{x}_j)/d(\mathbf{x}_i, \mathbf{x}_j)$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between the two points. Without loss of generality, we can take the proportionality constant to be 1. Thus, the total potential energy for n charged particles is given by

$$E_n = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)}. \tag{19}$$

The minimum energy design can be obtained by minimizing E_n with respect to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

How to choose the charge for each particle? Since the objective of the experiment is to maximize the yield of nanostructures, we should assign a large charge to a particle where the yield is low and vice versa. By doing this, the new particle to be

introduced into the box will repel more from the low yield region and will occupy a position in a region which is more likely to be of higher yield. This strategy will allow us to quickly find points with maximum yield. Thus, the amount of positive charge at each point should be a decreasing function of the yield at that specific point. There are many choices for a decreasing function; we choose a simple function

$$q(\mathbf{x}_i) = (1 - \alpha p(\mathbf{x}_i))^{\gamma}, \tag{20}$$

where $100p(\mathbf{x}_i)$ is the true percentage of growth at $\mathbf{x}_i$ and α and γ are positive tuning constants. To make the charge positive for all $\mathbf{x}$, we must have

$$\alpha \leq [max_{\mathbf{x}} p(\mathbf{x})]^{-1}. \tag{21}$$

When n design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ have already been selected, the next design point is obtained as:

$$\begin{aligned} \mathbf{x}_{n+1} &= \arg\min_x \left(E_n + \sum_{i=1}^{n} \frac{q(\mathbf{x})q(\mathbf{x}_i)}{d(\mathbf{x}, \mathbf{x}_i)} \right) \\ &= \arg\min_x \sum_{i=1}^{n} \frac{q(\mathbf{x})q(\mathbf{x}_i)}{d(\mathbf{x}, \mathbf{x}_i)}. \end{aligned} \tag{22}$$

Note that in order to perform the above minimization, we need to assign some charge $q(\mathbf{x})$ to each point $\mathbf{x}$ where we have not observed any yield. This can be done by predicting the yield at each of these points based on the observed data and plugging in these estimates in (20). In order to predict these yields, we need an interpolating function $\hat{p}(\mathbf{x})$ which satisfies $\hat{p}(\mathbf{x}_i) = p(\mathbf{x}_i)$ for $i = 1, \ldots, n$. There are several possible choices of such an interpolating function, e.g., kriging (Santner et al. 2003). However, as already mentioned earlier, at the initial stages owing to complexities associated with parameter estimation and computation, kriging may not be a very good choice. Keeping in mind that we want to use a simple function that would not involve parameter estimation problems, we propose to use inverse distance

weighting to perform this prediction. This means, after every single observation, the estimated yield at a point $\mathbf{x}$ would be given by

$$\hat{p}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \big(d(\mathbf{x}, \mathbf{x}_i)\big)^{-k} p(\mathbf{x}_i)}{\sum_{i=1}^{n} \big(d(\mathbf{x}, \mathbf{x}_i)\big)^{-k}}, \tag{23}$$

where k is a constant. We take $k = 2$, which is a common practice. This estimate will be plugged into (20) to update the charge of each point after an iteration, and subsequently the next design point will be chosen from (22). This gives us the desired sequential space-filling design, for which following results can easily be established (see Appendix B):

Proposition 4.1. Assume that the entire design space $[0, 1]^m$ is divided into a finite number of grid points $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$. Let $p(\mathbf{X}_g) = \max\big\{p(\mathbf{X}_1), \ldots, p(\mathbf{X}_N)\big\}$, and

$$\mathbf{x}_n = \min_{\mathbf{x} \in \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}} \sum_{i=1}^{n-1} \frac{q(\mathbf{x}) q(\mathbf{X}_i)}{d(\mathbf{x}, \mathbf{X}_i)}.$$

Let $\mathbf{x}_g = \arg\max_{\mathbf{x}} p(\mathbf{x})$. For $\alpha = 1/p(\mathbf{x}_g)$, $\mathbf{x}_n = \mathbf{X}_g$ for some $n \leq N$.

Proposition 4.2. Consider the definition of $\mathbf{x}_n$ given by (22). Let $\mathbf{x}_g = \arg\max_{\mathbf{x}} p(\mathbf{x})$. For $\alpha = 1/p(\mathbf{x}_g)$, if $\mathbf{x}_n = \mathbf{x}_g$ for some $n = n_0$, then $\mathbf{x}_n = \mathbf{x}_g$ for all $n > n_0$.

Physically, Proposition 4.1 implies that there exists a value of the tuning parameter α for which if a grid search is done with the minimum energy algorithm, then the global optimum on the grid will be reached in a finite number of iterations. Proposition 4.2 implies that for the same value of α, the design will continue to stick to the global optimum once it picks that point in some iteration. However, because $\mathbf{x}_g$ is not known to the experimenter, practically it is not possible to use the optimal choice of α. We will develop some adaptive estimation techniques for its practical implementation. Note that here we assume the function $p(\mathbf{x})$ is known, which is not true for physical experiments. Thus it needs to be estimated from data. The data is random and hence the estimator of $p(\mathbf{x})$ has some uncertainty, which complicates the sequential design problem. These will be addressed in Section 4.

Note that when $\gamma = 0$, the algorithm does not take into consideration the observed yield data; rather chooses points sequentially to minimize $\sum_i \sum_j [d(\mathbf{x}_i, \mathbf{x}_j)]^{-1}$. The resulting design is equivalent, in some sense, to the maximin distance design proposed by Johnson, Moore and Ylvisaker (1990). See also Morris and Mitchell (1995).

A similar idea based on an algorithm introduced in the statistical software JMP (version 6) by Bradley Jones called *minimum potential design* (MPD) was used by Kessels, Jones, Goos and Vandebroek (2006) for generating space filling designs. Such designs are obtained by minimizing the sum total of the elastic potential energy and the electrostatic potential energy of a system. The former is proportional to the squared distance and the latter is proportional to the inverse of the distance. The *sequential minimum energy design* (SMED) proposed here is completely different from MPD owing to two reasons. First, the MPD algorithm assumes equal charge for each point in the design space, which is not true for SMED. Therefore, the selection of points in MPD is independent of the underlying response function, essentially making it similar to other space-filling designs like LHD or uniform designs. Second, the objective function for MPD is the sum of two different types of energy with the implicit assumption that the associated constants are equal. (Note that the spring constant associated with the elastic energy and the Coulomb's constant associated with the electrostatic energy need not be the same.) On the other hand, SMED is based on the electrostatic energy alone, and does not require any such assumption.

4.3.1 Sequential minimum energy designs with deterministic functions and known α

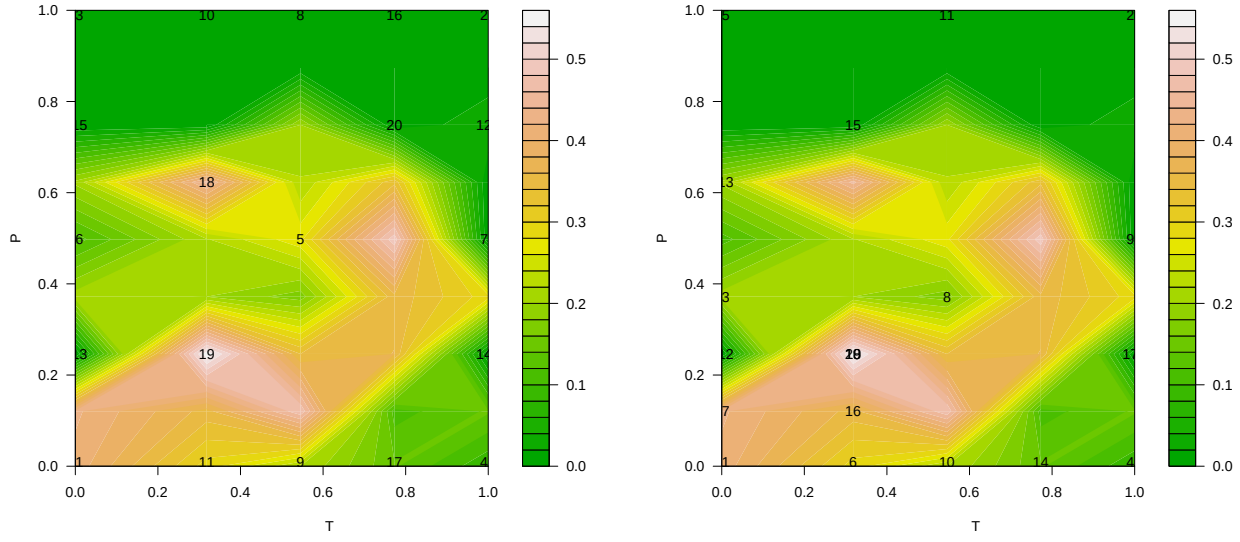
The performance of the algorithm is demonstrated by using it to generate a 20-run design with the objective of maximizing yield of CdSe nanowires. As described in Dasgupta et al. (2006) and shown in Table 11 we have 45 available design points corresponding to 9 levels of pressure and 5 levels of temperature, both variables scaled down to [0,1]. The design point (0.32, 0.25) is globally optimal and results

in a maximum yield of 55.56%. The 20-run design chosen by the algorithm fixing $\alpha = 1/0.55 = 1.8$, $\gamma = 3$, and starting our search from (0,0) is shown in the bottom right panel of Figure 15. The numbers in the figure denote the sequence in which the 20 points were chosen. We find that the algorithm detects the optimal setting at the 10th iteration, and henceforth sticks to that point in the subsequent iterations. Note that although the 7th selected point results in no yield, point 8 immediately makes a far-away move; this is not possible with improvement-based algorithms and bears testimony to this algorithm's self-correcting property. Thus, with proper choice of α and γ, it is possible to reach the optimum setting with less than one-fourth of the total number of runs originally conducted in this experiment.

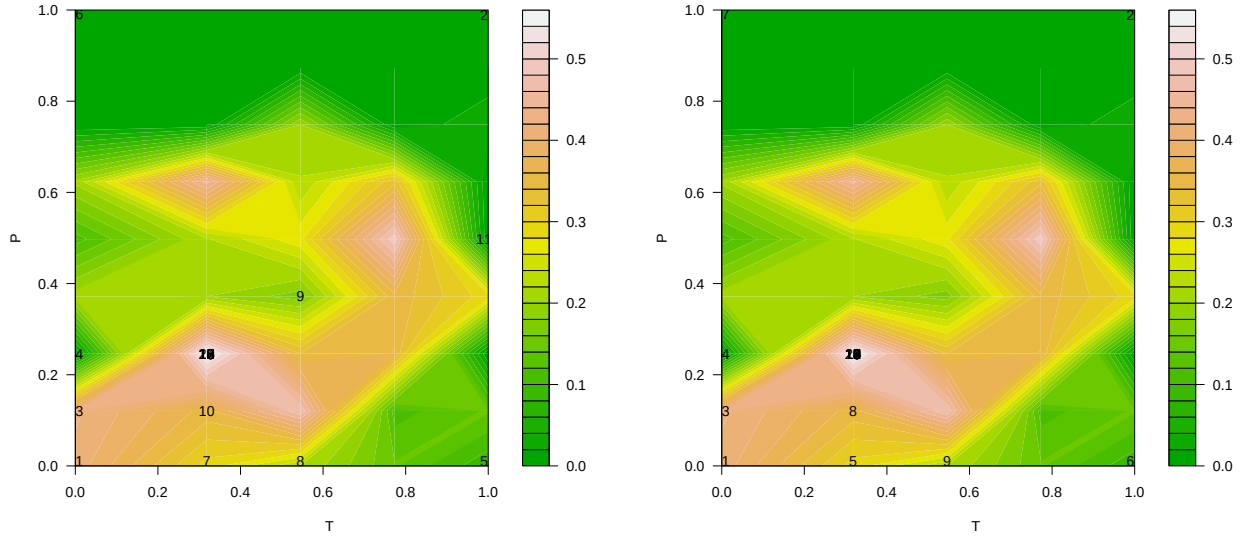
Figure 15 shows four 20-run designs generated with four different values of γ ($\gamma = 0, 1, 2, 3$). In each case, the search starts from (0,0), i.e., the bottom-left corner of the two-dimensional design space. The design picks the global optimum at the 17th iteration when $\gamma = 1$, at the 11th iteration when $\gamma = 2$ and at the 9th iteration when $\gamma = 3$. When $\gamma = 0$, as mentioned earlier, the resulting design is, in some sense equivalent to the maximin-distance design, and is independent of the observed data.

Figure 16 shows four designs generated with four other starting points - one at the center (0.55, 0.50), one at the top right corner (1,1), and two other at two locally optimal points (0.32,0.62) and (0.77,0.50). The performance, in general, is not very sensitive to the choice of the initial point. However, as seen from the bottom right panel, if the initial point is chosen very close to a local optimum, the algorithm takes a longer time to converge to the global optimum.

So far we have assumed that the parameters α and γ are known, which will not be the case in practice. Further, we need to define a universe of design points from which the minimum energy algorithm will sequentially pick up points. We develop a practically implementable algorithm by addressing the above issues in the following sections.

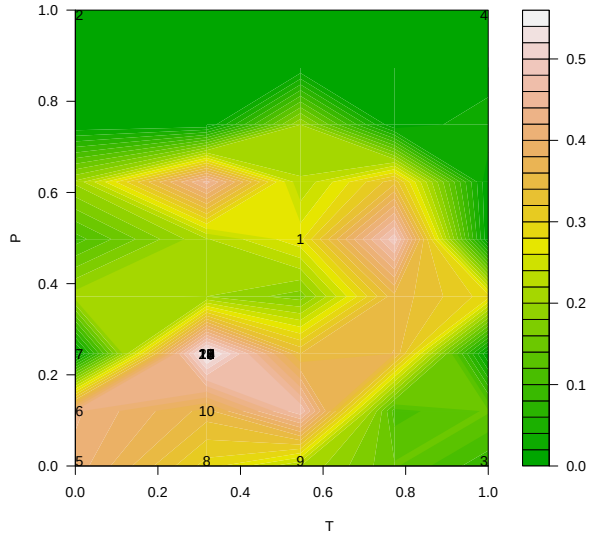


$\gamma = 0$ $\gamma = 1$

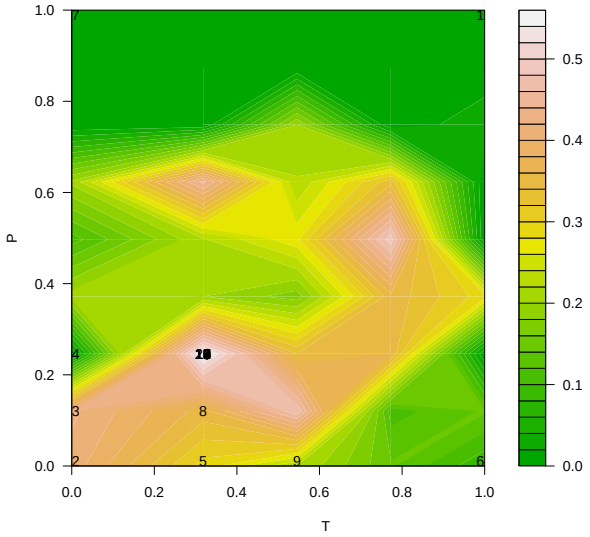


$\gamma = 2$ $\gamma = 3$

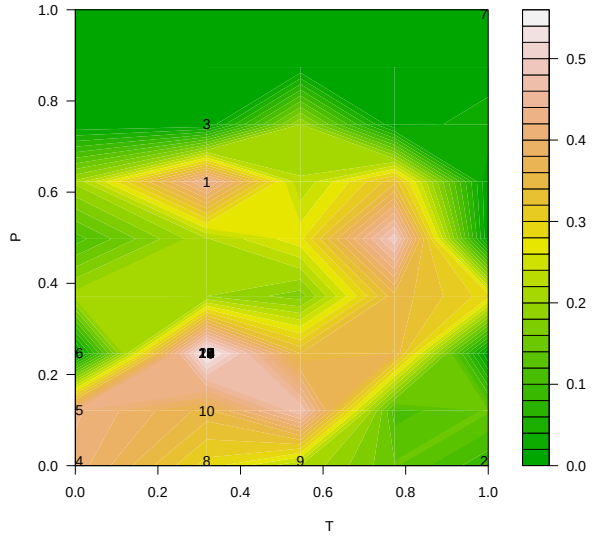
Figure 15: Performance of the design with initial point (0,0) and different γ



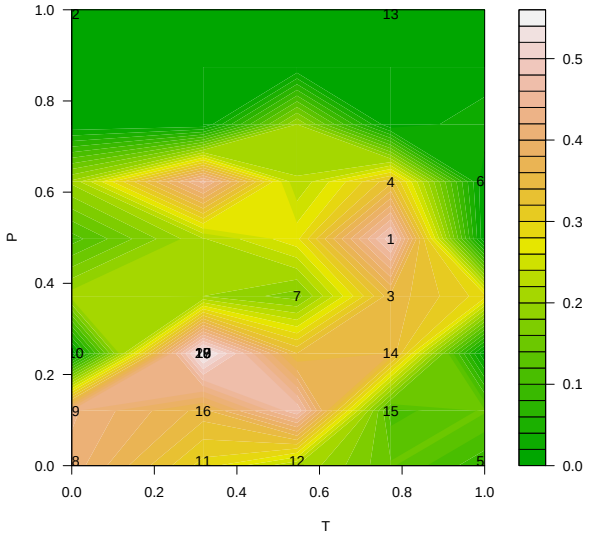
$\mathbf{x}_1 = (0.55, 0.50)$



$\mathbf{x}_1 = (1, 1)$



$\mathbf{x}_1 = (0.32, 0.62)$



$\mathbf{x}_1 = (0.77, 0.50)$

Figure 16: Performance of the design with different initial points and $\gamma = 3$

4.3.2 Estimation of parameters

The choice of the parameters α and γ are important from the point of view of performance of the minimum energy algorithm. So far we have chosen α as the inverse of the maximum yield. Since the maximum yield will not be known in practice, an estimation procedure has to be developed.

Ideally, we would like to have a design which picks different points before reaching the optimum and sticks to the optimum when it reaches there. When $p(\mathbf{x}_g)$ is known, the convergence criterion of the algorithm is defined as follows: if $q(\mathbf{x}) = 0$ for some $\mathbf{x} = \mathbf{x}^*$ then the next point to be picked will be $\mathbf{x}^*$. Equation (20) would then ensure that $\mathbf{x}^* = \mathbf{x}_g$ and Theorem 1 will ensure that all the subsequent points chosen by the design will be $\mathbf{x}_g$.

When we estimate α, our convergence criterion has to be $0 \leq q(\mathbf{x}) < \delta$, where $\delta > 0$ is a pre-defined small number. Let us introduce the following notation:

- α_n : Estimator of α after the nth iteration.
- $p_{(n)} = \max_{1 \leq i \leq n} p(\mathbf{x}_i)$.
- $p_g = p(\mathbf{x}_g)$.
- n_g : The iteration at which the point $\mathbf{x}_g$ is picked for the first time.

To make sure that the design does not (i) converge to some point before reaching $\mathbf{x}_g$, and (ii) does not wander away from $\mathbf{x}_g$ once it reaches there, the following two conditions must be satisfied:

$$q(\mathbf{x}_n) \geq \delta \quad \text{when} \quad n < n_g. \tag{24}$$

$$0 \leq q(\mathbf{x}_n) < \delta \quad \text{when} \quad n = n_g. \tag{25}$$

The above two conditions, along with (20) imply the following:

$$\alpha_n \leq \frac{1-\delta^{1/\gamma}}{p_{(n)}} \quad \text{when} \quad n < n_g. \tag{26}$$

$$\frac{1-\delta^{1/\gamma}}{p_g} < \alpha_n \leq \frac{1}{p_g} \quad \text{when} \quad n = n_g. \tag{27}$$

The problem is that n_g is unknown. If we choose $\alpha_n = (1-\delta^{1/\gamma})/p_{(n)}$, condition (26) will be satisfied for all n, but condition (27) will never be satisfied. This means, the algorithm will never converge, but, at the same time, not get stuck to a local optima.

Keeping in mind that α_n has to be less than $1/p_{(n)}$ (otherwise the design will converge to immediately), and should converge to $1/p_g$ as $n \longrightarrow \infty$, another possible choice of α_n is

$$\alpha_n = \frac{1}{p_{(n)} + a_n\big(1-p_{(n)}\big)}$$

where $\{a_n\}$ is a sequence such that $a_n \longrightarrow 0$. This choice ensures convergence at the nth iteration if

$$a_n < \frac{p_{(n)}}{1-p_{(n)}}.\frac{\delta^{1/\gamma}}{1-\delta^{1/\gamma}}.$$

This may result in an early convergence to a point with yield much lower than the optimum. For example, let $\gamma = 3$ and $\delta = 0.001$. If we choose $a_n = 1/n$ and by the 10th iteration, an yield of 50% is achieved, then the algorithm will converge to that point at the 10th iteration.

To strike a balance between very early convergence to a point other than $\mathbf{x}_g$ and no convergence at all, we introduce another parameter n_c, which denotes the number of runs till which the experimenter wants to pick distinct points from the design space and does not want convergence, i.e., $n_c < \min\{n : q(\mathbf{x}_n) < \delta\}$. Practically, n_c can be taken as 75% to 90% of the maximum affordable number of runs. Now, if we choose α_n as

$$\alpha_n = \frac{1-b_n\delta^{1/\gamma}}{p_{(n)} + a_n\big(1-p_{(n)}\big)}, \tag{28}$$

where $\{a_n\}$ and $\{b_n\}$ are two sequences such that $a_n \longrightarrow 0$ and $b_n = n_c/n$, then it is easily seen that the following two conditions are satisfied:

$$\alpha_n \longrightarrow 1/p_g, \tag{29}$$

$$q(\mathbf{x}_n) > \delta \quad \text{when} \quad n < n_c. \tag{30}$$

We also propose $a_n = 1/n^{\phi}$, where ϕ is a positive constant. Therefore, our estimator of α after the nth iteration becomes

$$\alpha_n = \frac{1 - (n_c/n)\delta^{1/\gamma}}{p_{(n)} + (1/n^{\phi})\big(1 - p_{(n)}\big)}. \tag{31}$$

To make sure that the numerator of α_n defined by (31) is non-negative, we must have $n_c\delta^{1/\gamma} \leq 1$. This condition gives us an upper bound for γ as

$$\gamma \leq \frac{\log \delta}{\log n_c^{-1}}.$$

We shall choose γ slightly less than this upper bound. A reasonable choice could be $\gamma = \log \delta / \log n_c^{-1} - 1/n_c$.

4.3.3 Choice of a universe of design points and the initial point

When no nanostructures are obtained (i.e., $p(\mathbf{x}) = 0$), the minimum energy design will continue to sample points from the corners of the box (similar to a situation when $\gamma = 0$, see Figure 1, top left panel). But when the number of factors m is high, the number of experiments that can be afforded is usually much less than the number of corner points 2^m. Since we intend to perform a grid search, it is important to choose a universe (large set) of design points in the design space which will form the grid from which points will be chosen sequentially.

To address this issue, we propose sampling from a large space-filling design. Suppose we plan to perform an experiment with n runs. Then select an efficient fixed design in N runs, where $N \gg n$. We can use an orthogonal array (Wu and Hamada

2000), Latin hypercube design (McKay et al. 1979), maximin design (Johnson et al. 1990), or an orthogonal maximin Latin hypercube design (Joseph and Hung 2006) for this purpose. Now, the n runs will be chosen from the N runs using the minimum energy design algorithm.

Without any prior information, the initial point in the sequential design may be chosen at random.

4.3.4 Algorithm for deterministic response

The SMED algorithm for deterministic response thus consists of the following steps:

1. Decide upon the number of trials (n) to be conducted.
2. Choose δ, ϕ and n_c.
3. $\gamma \longleftarrow (\log \delta)/(\log n_c^{-1}) - 1/n_c$.
4. $\Omega \longleftarrow \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ (Select a universe of $N(\gg n)$ design points using a space-filling design).
5. $\mathbf{x}_1 \longleftarrow \mathbf{z}_{INITIAL}$ where $INITIAL = \max(1, integer(nU))$ where $U \sim uniform[0, 1]$.
6. Observe $p(\mathbf{x}_1)$.
7. $pmax \longleftarrow p(\mathbf{x}_1)$.
8. $\alpha_1 \longleftarrow 1 - n_c\delta^{1/\gamma}$.
9. For $j = 1, \ldots N$, $q(\mathbf{z}_j) \longleftarrow \big(1 - \alpha_1 p(\mathbf{x}_1)\big)^{\gamma}$.
10. $J \longleftarrow 1$.
11. $\mathcal{D} \longleftarrow \mathbf{x}_1$.
12. WHILE $J < n$

- IF $q(\mathbf{z}) < \delta$ (convergence criterion) for $\mathbf{z}^* \in \Omega$,
  - $\mathbf{x}_{J+1} \longleftarrow \mathbf{z}^*$.
- ELSE
  - $E_{J+1} \longleftarrow \sum_{\mathbf{x}\in\mathcal{D}} \frac{q(\mathbf{x})q(\mathbf{z})}{d(\mathbf{x},\mathbf{z})}$ for each $z \in \Omega$. (Evaluate the total energy; assign a large value M to E_{J+1} whenever $\mathbf{x} = \mathbf{z}$).
  - $\mathbf{x}_{J+1} \longleftarrow \arg\min_{\mathbf{z}\in\Omega} E_{J+1}$ (Choose the $(J+1)$th design point that minimizes the total energy).
  - $\mathcal{D} \longleftarrow \mathcal{D} \bigcup \{\mathbf{x}_{J+1}\}$.
  - $pmax \longleftarrow \max(pmax, p(\mathbf{x}_{J+1}))$.
  - $\alpha_{J+1} \longleftarrow \frac{1-(n_c/J)\delta^{1/\gamma}}{pmax+1/(J+1)^{\phi}(1-pmax)}$ (update α).
  - For all $\mathbf{x} \in \mathcal{D}$, $q(\mathbf{x}) \longleftarrow \big(1 - \alpha_{J+1}p(\mathbf{x})\big)^{\gamma}$ (update charge of selected design points).
  - For all $\mathbf{z} \in \Omega \bigcap \mathcal{D}^c$ (points not yet selected),
    - $\hat{p}(\mathbf{z}) = \frac{\sum_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x})/d(\mathbf{x},\mathbf{z})}{\sum_{\mathbf{x}\in\mathcal{D}} 1/d(\mathbf{x},\mathbf{z})}$ (estimate yield by inverse distance weighting).
    - $q(\mathbf{z}) \longleftarrow \big(1 - \alpha_{J+1}\hat{p}(\mathbf{z})\big)^{\gamma}$ (update charge).
- $J \longleftarrow J + 1$.

13. The final design $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$.

4.3.5 Performance of the algorithm

We use the following measures to evaluate the performance of the algorithm:

(i) Average yield (AY) is defined as the arithmetic average of the yields observed at the n selected design points and is given by $AY = \sum_{i=1}^{n} p(\mathbf{x}_i)/n$.

(ii) Maximum yield ($Y_{\max}$) denotes the maximum yield observed within n runs, i.e. $Y_{\max} = \max_{1\le i\le n} p(\mathbf{x}_i)$

(iii) No yield (NY) denotes the number of no-yield points picked up by the design, i.e., $NY = \sum_{i=1}^{n} \mathcal{I}\big\{p(\mathbf{x}_i) = 0\big\}$, where $\mathcal{I}$ is the indicator function.

(iv) Convergence to the global optimum, defined as $CNGO = \mathcal{I}\{\mathbf{x}_n = \mathbf{x}_{n-1} = \mathbf{x}_g\}$.

(v) Convergence to a local optimum, defined as $CNLO = \mathcal{I}\{\mathbf{x}_n = \mathbf{x}_{n-1} \neq \mathbf{x}_g\}$.

We want $AY, Y_{\max}$ and $CNGO$ to be large and NY and $CNLO$ to be small. Among the five performance measures, (ii) and (iii) will be considered most important from the point of view of the objective of experimentation. As already mentioned earlier, convergence is not the primary concern here; rather the experimenter would like to generate reasonable amount of non-zero yield data and locate high-yield zones as quickly as possible. AY is not as important as $Y_{\max}$ or NY, because even if the algorithm converges fast to a local optima very early, AY will be very high; which is not the true reflection of the performance of the algorithm.

Performance evaluation 1: The nanowire synthesis experiment

We examine the performance of the algorithm by choosing 20-run sequential minimum energy designs from a universe of $N = 45$ design points in the CdSe experiment (Table 11). Recall that the point (0.32,0.25) with an yield of 0.5556 is considered as the global optimum A 3^3 full factorial experiment is performed with the three tuning parameters n_c, δ and ϕ at three levels each. Each run is replicated 100 times to generate 100 different designs with different (random) choice of the starting point. Table 12 summarizes the experimental results.

The following observations can be made from Table 12:

(i) The overall performance of the algorithm is quite satisfactory. Irrespective of the choice of the three parameters $n_c, \delta, \quad \phi$, the algorithm picks the global optimum almost everytime.

Table 12: Performance of the algorithm for deterministic version of the nanowire yield function

n_c	δ	γ	ϕ	AY		$Y_{\max}$		NY		$\sum CNGO$	$\sum CNLO$
				mean	s.d.	$\sum \mathcal{I}\{Y_{\max} = \mathbf{x}_g\}$	$\min(Y_{\max})$	mean	s.d.		
12	0.0010	2.7	1	0.23	0.02	100	0.56	4.9	0.41	0	0
12	0.0010	2.7	2	0.34	0.01	93	0.49	3.2	0.48	93	7
12	0.0010	2.7	3	0.34	0.02	82	0.42	3.3	0.53	82	18
12	0.0005	3.0	1	0.22	0.01	100	0.56	4.9	0.29	0	0
12	0.0005	3.0	2	0.34	0.01	84	0.48	3.0	0.51	84	16
12	0.0005	3.0	3	0.34	0.01	83	0.42	3.1	0.57	83	17
12	0.0001	3.6	1	0.23	0.01	100	0.56	4.9	0.24	0	0
12	0.0001	3.6	2	0.30	0.01	100	0.56	4.0	0.47	100	0
12	0.0001	3.6	3	0.29	0.01	98	0.49	4.1	0.49	98	2
15	0.0010	2.5	1	0.23	0.01	100	0.56	4.9	0.29	0	0
15	0.0010	2.5	2	0.30	0.01	100	0.56	4.0	0.47	100	0
15	0.0010	2.5	3	0.29	0.01	98	0.49	4.1	0.49	98	2
15	0.0005	2.7	1	0.23	0.01	100	0.56	4.9	0.29	0	0
15	0.0005	2.7	2	0.29	0.01	100	0.56	4.0	0.37	100	0
15	0.0005	2.7	3	0.29	0.01	98	0.48	4.0	0.49	98	2
15	0.0001	3.3	1	0.23	0.01	100	0.56	4.8	0.38	0	0
15	0.0001	3.3	2	0.30	0.01	93	0.49	3.5	0.54	93	7
15	0.0001	3.3	3	0.30	0.01	93	0.49	3.5	0.50	93	7
17	0.0010	2.4	1	0.22	0.01	100	0.56	5.2	0.69	0	0
17	0.0010	2.4	2	0.27	0.01	100	0.56	4.5	0.67	100	0
17	0.0010	2.4	3	0.26	0.01	100	0.56	4.6	0.51	100	0
17	0.0005	2.6	1	0.23	0.01	100	0.56	5.0	0.39	0	0
17	0.0005	2.6	2	0.26	0.01	93	0.49	4.2	0.60	93	7
17	0.0005	2.6	3	0.26	0.01	100	0.56	4.6	0.51	100	0
17	0.0001	3.2	1	0.23	0.01	100	0.56	4.9	0.43	0	0
17	0.0001	3.2	2	0.26	0.01	93	0.49	4.2	0.60	93	7
17	0.0001	3.2	3	0.26	0.01	96	0.49	4.3	0.54	96	4

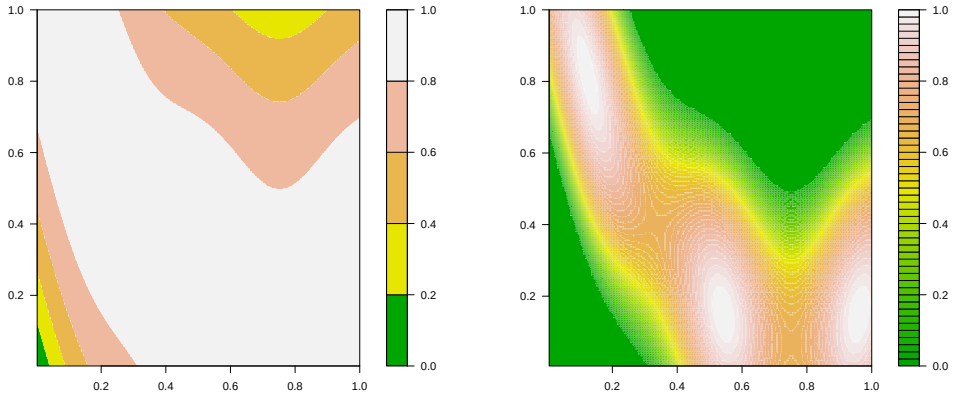
(ii) The algorithm does not converge within 20 iterations if $\phi = 1$, but picks the optimum 100% of the times. The convergence improves as ϕ increases, but, as expected, when $\phi = 3$, we have a few instances of convergence to some local optimum.

(iii) Even when the algorithm did not converge to the global optimum, it mostly converged to a local optimum with the second highest yield.

(iv) The average number of no-yield points is 5 when $\phi = 1$ and 4 when $\phi = 2$ or 3.

Performance evaluation 2: A standard global optimization test function

We now study the performance of the algorithm using a standard test function used in global optimization literature. A popular test function used to evaluate the performance of any global optimization algorithm is the Branin function (Branin, 1972). It has three global minima. However, test functions for our algorithm must



Contour plot of f Contour plot of p

Figure 17: Simulated yield functions using the Branin function

satisfy two properties - first, it has to be multimodal, and second, there has to be a large non-convex and disconnected no-yield region. We therefore transform the Branin function to obtain another function that satisfies the above two criteria.

Let $B(x, y)$ denote the negative of the Branin function such that

$$B(x,y) = -\left((y-5x^2/4\pi^2+5x/\pi-6)^2+10(1-1/8\pi)\cos x+10\right), -5 \leq x \leq 10, 0 \leq y \leq 15.$$

The above function attains maximum (-0.3979) at three distinct points $(-\pi, 12.25)$, $(\pi, 2.25)$ and (9.4248,2.25). The minimum is around 305. We transform B to a function f with a domain $[0, 1] \times [0, 1]$ and range [0,1] by applying the following transformation:

$$f(u,v) = \max\left(0, \frac{f(15u-5, 15v)+305}{-0.3979+305}\right), 0 \leq u, v \leq 1.$$

However, about 66% of the values of the function f lie above 0.80 (see the left panel of Figure 17). To create a function that is more similar to the yield function of nanostructures, we define $p(u, v) = \max(\frac{f(u,v)-0.8}{0.2}, 0) \leq u, v \leq 1$. The contour plot of p is shown in the right panel of Figure 17. The function p attains its maximum value 1 at three points (0.1239,0.8167), (0.5428,0.15) and (0.9617,0.15).

The algorithm given in Section 3.4 is used to generate a 25-run SMED for the modified Branin yield function p defined above. The 25 design points are sampled from a 150-run LHD. Among the 150 points chosen by the LHD, we consider the point (0.1208,0.8322) for which $p(u, v) = 0.9996$ as the global optimum. Four different parameter combinations of n_c and ϕ are used, keeping δ fixed at 0.001 and the results are summarized and compared with the performance of three standard space-filling designs in Table 13. Clearly, the results are very encouraging. It is seen that the global optimum was picked on each of the four occasions. Most interestingly, prior to convergence, all the three distinct regions of the highest yield were probed, as seen from Figure 18. The number of no-yield points was also negligible.

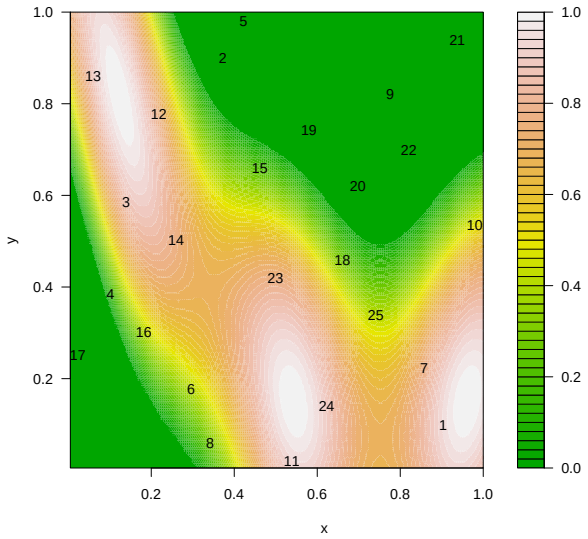
Table 13: Performance of SMED and other space filling designs with the modified Branin function

Design	Performance characteristics			
	AY	$Y_{\max}$	NY	$CNGO$
Uniform	40.8%	94.33%	8/25	N.A.
Minimum potential	39.9%	93.94%	8/25	N.A.
Maximin LHD	38.9%	88.13%	9/25	N.A.
SMED(2.3,1,0.001,20)	74.9%	99.96%	2/25	YES
SMED(2.2,1,0.001,22)	67.3%	99.96%	3/25	YES
SMED(2.3,2,0.001,20)	72.5%	99.96%	2/25	YES
SMED(2.2,2,0.001,22)	68.9%	99.96%	3/25	YES

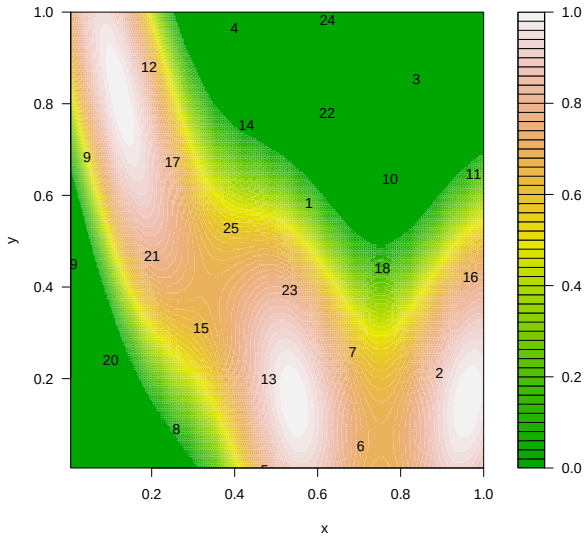
N.A. : Not Applicable.

4.4 Random functions

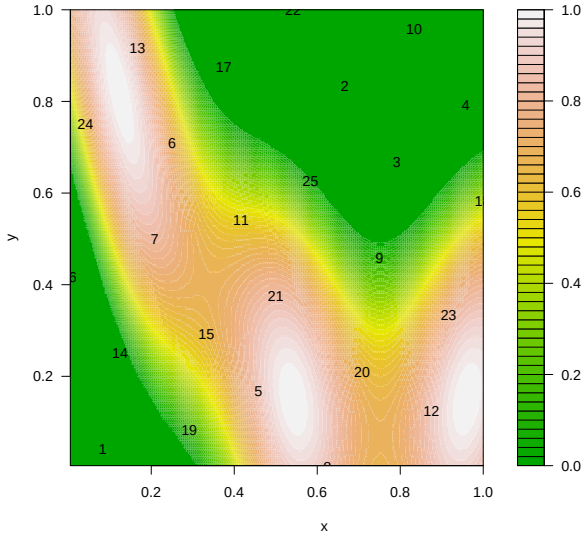
So far the SMED algorithm assumes that the yield function $p(\mathbf{x})$ is known. However, in practice it has to be estimated from the data. Suppose, r nanostructure samples are collected from each experimental run (in the CdSe nanostructures experiment r was 180.) Then the observed yield at the ith point y_i is a binomial random variable with parameters r and $p(\mathbf{x_i})$. Thus, we actually observe $\hat{p}(\mathbf{x}_i) = y_i/r$. In the CdSe experiment, $r = 180$. Figure 19 shows two contour plots of observed yield for $r = 5$



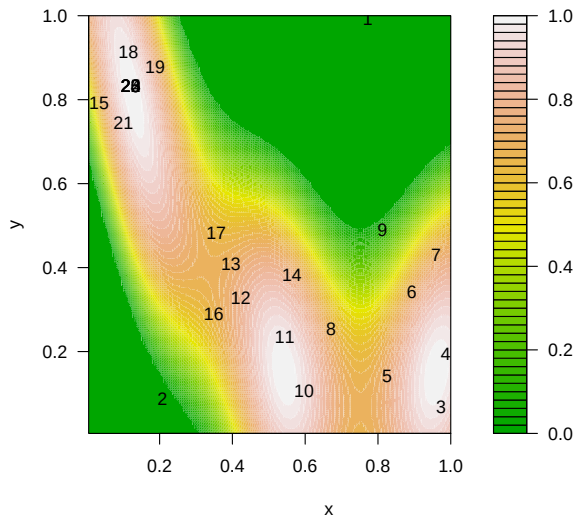
Uniform design



Minimum potential design



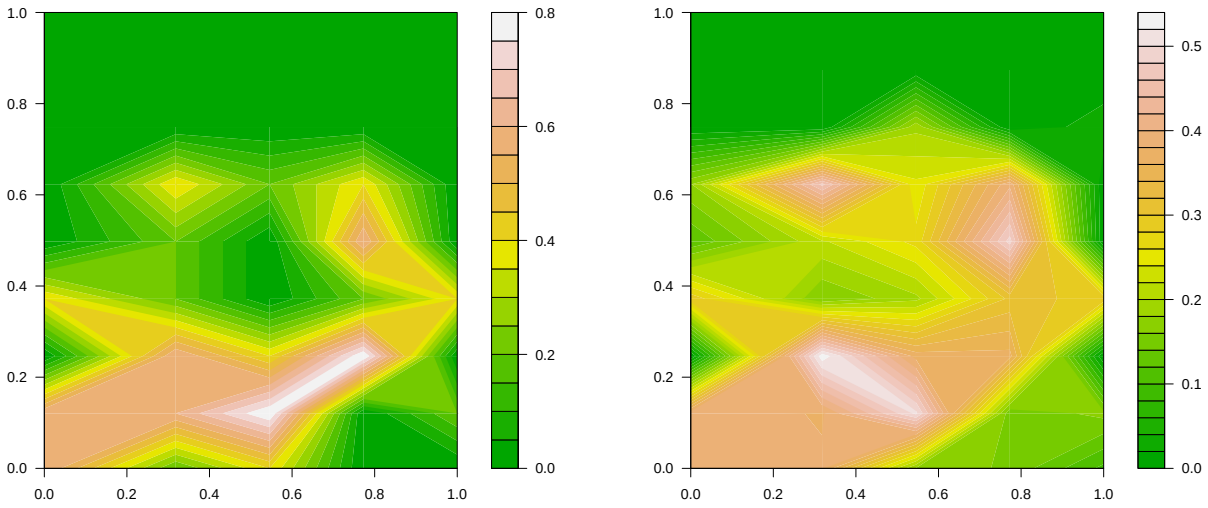
Maximim LHD



SMED(2.3,1,0.001,20)

Figure 18: Comparison with standard designs

and $r = 180$ respectively, assuming that the true yields at the 45 points are those given in Table 11. These plots are generated with the estimated yields $\hat{p}(\mathbf{x})$ at the 45 points shown in Table 11, where $\hat{p}(\mathbf{x}_i) = y_i/r,\ \ i = 1, \ldots, 45$ and y_i is a randomly generated observation from a binomial $\big(r, p(\mathbf{x}_i)\big)$.



$r = 5$ $r = 180$

Figure 19: Contour plots with random yields

The random nature of the yield function complicates the matter. Despite having reached the global optimum, we may not observe maximum yield at that point because of the random error. This may result in assigning a high positive charge to that point and thus the algorithm will never pick that point again in near future, which is undesirable.

To understand how the algorithm described in Section 3.4 works for a random yield function, we simulate 20-run designs with parameters $\phi = 2$, $\delta = 0.001$, $n_c = 15$ and $\gamma = 2.5$ (the best combination seen from Table 12) using the CdSe yield function for various values of r ranging from 5 to 180. As explained earlier, the underlying function, defined at 45 points, is given by $\hat{p}(\mathbf{x}_i) = y_i/r,\ \ i = 1, \ldots, 45$, where y_i is a randomly generated observation from a binomial $\big(r, p(\mathbf{x}_i)\big)$. The values of $p(\mathbf{x}_i)$ are

Table 14: Performance of the algorithm for random functions

r	AY		$Y_{\max}$		NY		$\sum CNGO$	$\sum CNLO$
(Sample size)	mean	s.d.	$\sum \mathcal{I}\{Y_{\max} = \mathbf{x}_g\}$	$\min(Y_{\max})$	mean	s.d.		
5	0.23	0.030	56	0.42	4.39	1.15	17	83
25	0.27	0.025	75	0.42	4.27	0.81	51	49
50	0.28	0.018	85	0.45	4.08	0.66	64	36
75	0.29	0.017	86	0.48	4.02	0.67	68	32
100	0.29	0.015	91	0.48	4.13	0.73	71	29
125	0.29	0.016	92	0.48	4.02	0.71	82	18
150	0.29	0.016	96	0.48	4.07	0.66	87	13
180	0.29	0.016	96	0.49	4.01	0.66	87	13
∞	0.30	0.015	100	0.56	4.06	0.47	100	0

obtained from Table 11.

The results, based on 100 simulations for each value of r, are sumarized in Table 14. As expected, for high values of r, the performance is not very different from the deterministic case. As r decreases, the performance worsens; for extremely small values of r (e.g., $r = 5$), the performance of the algorithm is not quite satisfactory although not drastically poor; it is seen to pick the global optimum 56% of the times, and the maximum yield identified is not less than 0.42.

4.4.1 An improved algorithm using Bayesian estimation

Although the existing algorithm is seen to work fairly well for random functions, for small values of r, the percentage of times the global optimum is picked needs to be improved (for $r = 5$, this was seen to be 56%). To improve the performance of the algorithm, we propose recursive Bayesian estimation of yield at each selected design point. Let y_n denote the yield observed at the nth selected point $\mathbf{x}_n$ and let $\hat{p}_F(\mathbf{x}_n) = y_n/r$ denote the frequentist estimate of $p(\mathbf{x}_n)$. Let us assume that $y_n|p(\mathbf{x}_n) \sim binomial(r, p(\mathbf{x}_n))$ and $p(\mathbf{x}_n)$ has a prior distribution $Beta(\alpha(\mathbf{x}_n), \beta(\mathbf{x}_n))$. Then, the Baye's estimate of $p(\mathbf{x}_n)$, which is its posterior expectation, is given by

$$\begin{aligned} \hat{p}_B(\mathbf{x}_n) &= \frac{y(\mathbf{x}_n) + \alpha(\mathbf{x}_n)}{r + \alpha(\mathbf{x}_n) + \beta(\mathbf{x}_n)} \\ &= \frac{r\hat{p}_F(\mathbf{x}_n) + n_0 p_0(\mathbf{x}_n)}{r + n_0}, \end{aligned} \tag{32}$$

where $p_0(\mathbf{x}_n) = \alpha(\mathbf{x}_n)/(\alpha(\mathbf{x}_n) + \beta(\mathbf{x}_n))$ is the prior mean of $p(\mathbf{x}_n)$ and $n_0 = \alpha(\mathbf{x}_n) + \beta(\mathbf{x}_n)$. We assume that the hyperparameter n_0 is the same for all points in the design space.

Now, $p_0(\mathbf{x}_n)$ can be estimated from the Bayesian estimates of yields obtained at the already sampled points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}\}$ by using inverse distance weighting. This estimator is given by

$$\begin{aligned} \hat{p}_D(\mathbf{x}_n) &= \hat{p}_D(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) \\ &= \frac{\sum_{j=1}^{n-1} \hat{p}_B(\mathbf{x}_j)\big(d(\mathbf{x}_n, \mathbf{x}_j)\big)^{-2}}{\sum_{j=1}^{n-1} \big(d(\mathbf{x}_n, \mathbf{x}_j)\big)^{-2}}. \end{aligned} \tag{33}$$

Thus from (32), we can write

$$\hat{p}_B(\mathbf{x}_n) \widehat{=} \frac{r\hat{p}_F(\mathbf{x}_n) + n_0\hat{p}_D(\mathbf{x}_n)}{r + n_0}. \tag{34}$$

Note that the Bayesian estimate of $p(\mathbf{x}_n)$ defined by (34) is a convex combination of $\hat{p}_F(\mathbf{x}_n)$, the frequentist estimate, and $\hat{p}_D(\mathbf{x}_n)$, the estimate of the prior mean based on the previous observations. The weights associated with the former is $r/(n_0 + r)$ and that with the latter is $1 - r/(n_0 + r) = n_0/(n_0 + r)$. Assuming that all the previous observations are on $\mathbf{x}_n$, a reasonable choice for n_0 is $(n - 1)$. Finally, we obtain the Bayesian estimate of $p(\mathbf{x}_n)$ as

$$\hat{p}_B(\mathbf{x}_n) = \frac{r}{r + n - 1}\hat{p}_F(\mathbf{x}_n) + \frac{n - 1}{r + n - 1}\hat{p}_D(\mathbf{x}_n). \tag{35}$$

When $n = 1$, the Bayesian estimate equals the frequentist estimate. As $n \longrightarrow \infty$, the weight associated with the frequentist estimate tends to zero. In particular, if $n = r + 1$, both components in the RHS of (35) receive equal weights of 0.5.

The Bayesian estimation procedure described above is recursive. That is, after selecting the nth design point, updating the set of design points to $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and obtaining the Bayesian estimate of $p(\mathbf{x}_n)$ using (35), the estimates of $p_0(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ will be updated as follows:

$$\hat{p}_D^{(n)}(\mathbf{x}_k) = \frac{\sum_{j:\ \mathbf{x}_j \in \mathcal{D}, j \neq k} \hat{p}_B^{(n)}(\mathbf{x}_j)\big(d(\mathbf{x}_k, \mathbf{x}_j)\big)^{-2}}{\sum_{j:\ \mathbf{x}_j \in \mathcal{D}, j \neq k} \big(d(\mathbf{x}_k, \mathbf{x}_j)\big)^{-2}} \quad \text{for all} \quad \mathbf{x}_k \in \Omega. \tag{36}$$

Here the upper suffix denotes the iteration number. Note that the above estimator has a slightly different from the usual inverse-distance weighting estimator, and is not an interpolator.

After selection of the $(n+1)th$ point, we shall have $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n+1}\}$. The Bayesian estimator of $p(\mathbf{x}_k)$ for all $\mathbf{x}_k \in \mathcal{D}$ will be given by

$$\hat{p}_B^{(n+1)}(\mathbf{x}_k) = \frac{r}{r+n}\hat{p}_F(\mathbf{x}_k) + \frac{n}{r+n}\hat{p}_D^{(n)}(\mathbf{x}_n). \tag{37}$$

In the new algorithm for random functions, step 12 of the algorithm described in Section 3.4 is changed as follows:

Step 12. WHILE $J < n$

- IF $q(\mathbf{z}) < \delta$ (convergence criterion) for $\mathbf{z}^* \in \Omega$
  - $\mathbf{x}_{J+1} \longleftarrow \mathbf{z}^*$,
- ELSE
  - $E_{J+1} \longleftarrow \sum_{\mathbf{x}\in\mathcal{D}} \frac{q(\mathbf{x})q(\mathbf{z})}{d(\mathbf{x},\mathbf{z})}$ for each $z \in \Omega$. (Evaluate the total energy; assign a large value M to E_{J+1} whenever $\mathbf{x} = \mathbf{z}$).
  - $\mathbf{x}_{J+1} \longleftarrow \arg\min_{\mathbf{z}\in\Omega} E_{J+1}$ (Choose the $(J+1)$th design point that minimizes the total energy).
  - $\mathcal{D} \longleftarrow \mathcal{D} \bigcup \{\mathbf{x}_{J+1}\}$.
  - For all $\mathbf{x} \in \mathcal{D}, \hat{p}_B^{(J+1)}(\mathbf{x}) \longleftarrow \frac{J\hat{p}_D^{(J)}(\mathbf{x}) + y(\mathbf{x})}{J+r}$.
  - $pmax \longleftarrow \max_{1\le i\le(J+1)} \left(p_B^{(J+1)}(\mathbf{x}_i)\right)$.
  - $\alpha_{J+1} \longleftarrow \frac{1-(n_c/J)\delta^{1/\gamma}}{pmax + 1/(J+1)^{\phi}(1-pmax)}$ (update α).
  - For all $\mathbf{z} \in \Omega \bigcap \mathcal{D}^c$, $\hat{p}_D^{(J+1)}(\mathbf{z}) \longleftarrow \frac{\sum_{\mathbf{x}\in\mathcal{D}} p_B^{(J+1)}(\mathbf{x})/d^2(\mathbf{x},\mathbf{z})}{\sum_{\mathbf{x}\in\mathcal{D}} 1/d^2(\mathbf{x},\mathbf{z})}$ (estimate p_0 for points not yet selected).
  - For all $\mathbf{z} \in \Omega \bigcap \mathcal{D}^c$, $q^{(J+1)}(\mathbf{z}) \longleftarrow \left(1 - \alpha_{J+1} p_D^{(J+1)}(\mathbf{x})\right)^{\gamma}$ update charge of points not yet selected).

– For all $\mathbf{x} \in \mathcal{D}$, $q^{(J+1)}(\mathbf{x}) \longleftarrow \left(1 - \alpha_{J+1} p_B^{(J+1)}(\mathbf{x})\right)^{\gamma}$ (update charge of selected design points).

– For all $\mathbf{x}_k \in \mathcal{D}$, $\hat{p}_D^{(J+1)}(\mathbf{x}_k) = \dfrac{\sum_{j:\ \mathbf{x}_j \in \mathcal{D}, j \neq k} \hat{p}_B^{(J+1)}(\mathbf{x}_j) \left(d(\mathbf{x}_k, \mathbf{x}_j)\right)^{-2}}{\sum_{j:\ \mathbf{x}_j \in \mathcal{D}, j \neq k} \left(d(\mathbf{x}_k, \mathbf{x}_j)\right)^{-2}}$ for all $\mathbf{x}_k \in \Omega$ (Update estimate of p_0 for selected points).

- $J \longleftarrow J + 1$.

4.4.2 Performance evaluation of the new algorithm

The modified algorithm was run for various values of r with the same combination of parameters ($\gamma = 2.5, \quad \phi = 2, \quad \delta = 0.001, \quad n_c = 15$) with the same test function used to generate Table 14. As before, one hundred 20-run designs were generated for value of r. Table 15 summarizes the results.

Table 15: Performance of the new algorithm for random functions

r (Sample size)	Algorithm	AY		$Y_{\max}$		NY		$\sum CNGO$	$\sum CNLO$
		mean	s.d.	$\sum \mathcal{I}\{Y_{\max} = \mathbf{x}_g\}$	$\min(Y_{\max})$	mean	s.d.		
5	Original	0.23	0.030	56	0.42	4.39	1.15	17	83
5	Modified	0.19	0.024	75	0.42	5.20	1.18	0	0
25	Original	0.27	0.025	75	0.42	4.27	0.81	51	49
25	Modified	0.22	0.020	88	0.42	4.91	0.84	0	0
100	Original	0.29	0.015	91	0.48	4.13	0.73	71	29
100	Modified	0.23	0.015	95	0.48	4.52	0.72	0	0
180	Original	0.29	0.016	96	0.49	4.01	0.66	87	13
180	Modified	0.24	0.016	98	0.49	4.48	0.67	0	0
∞	-	0.30	0.015	100	0.56	4.06	0.47	100	0

The observations can be summarized as follows:

1. The modified algorithm never converged within 20-iterations to a local or global optimum. As a consequence of non-convergence, the average yield was also much less for the new algorithm. This is expected, though, because, with the introduction of new influential (high or low yield) points, the yield (and consequently the charge) of points selected earlier may change quite drastically. Let $q^{(n)}(\mathbf{x}_n)$ denote the charge of the nth selected design point after the nth iteration, and suppose $q^{(n)}(\mathbf{x}_n) = \min_{1 \leq i \leq n} q^{(n)}(\mathbf{x}_n)$. After the $(n+1)$th iteration, it is possible that $q^{(n+1)}(\mathbf{x}_n) > q^{(n+1)}(\mathbf{x}_i)$ for some $i < n$, based on either a high-yield point observed near $\mathbf{x}_i$ or a very low-yield point observed near $\mathbf{x}_n$.

2. The average number of no-yield points for the modified algorithm is slightly more than that for the original algorithm. This is also expected, since in the modified algorithm, after sampling the first point with non-zero yield, the Bayesian estimate of p at all sampled points will be strictly positive. If $p(\mathbf{x}_i) = 0$, we must have $\hat{p}_F(\mathbf{x}_i) = 0$ but $\hat{p}_B(\mathbf{x}_i) > 0$ after a few iterations. Whereas the modified algorithm protects points at which the true yield is positive but the observed yield is zero, it does so at the cost of assigning positive yields to points at which the true (and observed) yield is zero. However, the modified algorithm does not increase the number of no-yield points substantially.

3. In spite of the above two points, the modified algorithm has a clear advantage over the original algorithm in terms of the number of times the global optimum is identified, especially for low values of r. For example, for $r = 5$, the number of times the global optimum is picked increases to 75% from 56% by using the new algorithm, which is considered a significant improvement.

4.5 Summary and conclusions

In this chapter, we have proposed a novel sequential space filling design called SMED for exploring best process conditions for synthesis of nanowires. The SMED is a novel approach to generate designs that are model independent, can quickly "carve out" regions with no observable nanostructure morphology, allow for the exploration of complex response surfaces, and can be used for sequential experimentation. Owing to its origination from laws of electrostatics, it should be appealing and comprehensible to a broad spectrum of scientific researchers. The basic idea has been developed into a practically implementable algorithm for deterministic functions, and guidelines for choosing the parameters of the design have been proposed. Performance of the algorithm has been studied using experimental data on nanowire synthesis as well as the modified Branin function. A modification of the algorithm based on Bayesian

estimation has been proposed for random functions.

The broad areas of future research should be the following:

1. The current algorithm is based on grid search, and therefore may not be applicable to functions with three or more input variables. We need to incorporate a suitable continuous optimization algorithm that will minimize the complex objective function given by (22) at each iteration of the algorithm and test its performance for functions of higher dimensions.

2. Although the proposed modification to handle random functions yields fairly satisfactory results, we would like to enhance its performance further by improving the recursive estimation method. There are many possibilities, ranging from using a k-nearest neighbor inverse-distance weighting to a more sophisticated kriging model.

3. One of the important future objectives of research related to synthesis of nanowires is to optimize continuous quality characteristics like diameter, length and density. Extension of the minimum energy algorithm to situations where the response is continuous will be extremely helpful in designing such experiments economically. The minimum energy algorithm can help in search for optimum of bounded functions. For continuous response, we need a generalization that will also explore the optima for unbounded functions.

4.6 References

Bai, X. D., Gao, P. X., Wang, Z. L. and Wang, E. G. (2003), "Dual-mode mechanical resonance of individual ZnO nanobelts," *Appl. Phys. Letts.* 82, p. 4806-4808.

Betro, B. (1991), "Bayesian Methods in Global Optimization," *Journal of Global Optimization*, 1(1), 1-14.

Branin, F. H. (1972), "Widely Convergent Method for Finding Multiple Solutions of Simultaneous Nonlinear Equations," *IBM J. Res. Develop*, 16, 504-522.

Cox, D. D. and John, S. (1997), "SDO : A Statistical Method for Global Optimization," *Multidisciplinary Design Optimization : State of the Art*, Alexandrov N, Hussaini MY (eds.), SIAM: Philadelphia, 315-329.

Dasgupta, T., Ma, C., Joseph, R., Wang, Z. L. and Wu, C. F. J. (2007), "Statistical Modeling and Analysis for Robust Synthesis of Nanostructures," *Journal of the American Statistical Association*, to appear.

Duan, X. F., Huang, Y., Agarwal, R. and Lieber, C. M. (2003), "Single-nanowire electrically driven lasers," *Nature* 421, p. 241.

Fang, K.T., Li, R., and Sudijanto, A. (2006), *Design and Modeling for Computer Experiments*, Boca Raton, FL: Chapman & Hall.

Fang, K.T., "Theory, method and applications of the uniform design," *International J. Reliability, Quality, and Safety Engineering*, 9(4), 305-315.

Hu, J.T., Odon, T. W. and Lieber, C. M. (1999), "Chemistry and physics in one dimension: Synthesis and properties of nanowires and nanotubes," *Acc. Chem. Res.* 32, p.435.

Henkenjohann, N., Gobel, R., Kleiner, M. and Kunert, J. (2005), "An Adaptive Sequential Procedure for Efficient Optimization of the Sheet Metal Spinning Process," *Quality and Reliability Engineering International*, 21, 439-455.

Johnson, M. E., Moore, L. M. and Ylvisaker, D. (1990), "Minimax and Maximin Distance Designs," *Journal of Statistical Planning and Inference*, 26, 131-148.

Jones, D. R., Schonlau, M. and Welch, W. J. (1998), "Efficient Global Optimization of Expensive Black Box Functions," *Journal of Global Optimization*, 13, 455-492.

Joseph, V. R. and Hung, Y. (2007), "Orthogonal-Maximin Latin Hypercube Designs," *Statistica Sinica*, to appear.

Kessels, R., Jones, B., Goos, P. and Vanderbrook, M. (2006), "An Efficient Algorithm for Constructing Bayesian Optimal Choice Designs," Research Report KBI 0616, Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven.

Li, Y., Qian, F., Xiang, J. and Lieber, C.M. (2006), "Nanowire electronic and optoelectronic devices," *Materials Today* 9(10), p.18.

Lieber, C. M. (2003), "Nanoscale science and technology: Building a big future from small things," *MRS Bull.* 28, p.486.

Mandal, A., Wu, C. F. J., and Johnson, K. (2006),"SELC : Sequential Elimination of Level Combinations by means of Modified Genetic Algorithms," *Technometrics*, 48, 273-283.

Morris, M.D., and Mitchell, T.J. (1995),"Exploratory Designs for Computer Experiments," *Journal of Statistical Planning and Inference*, 43, 381-402.

Myers, H. M. and Montgomery, D.C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, NewYork: Wiley.

Owen, A. B. (1995), "Randomly permuted (t, m, s)-nets and (t, s)-sequences," in Harald Niederreiter and Peter Jau-Shyong Shiue, ed., *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of Lecture Notes in Statistics, 299315, Springer-Verlag.

Patolsky, F. and Lieber, C. M. (2005), "Nanowire nanosensors," *Materials Today* 8(4), p.20.

Patolsky, F., Zheng, G. and Lieber, C. M. (2006), "Nanowire-Based Biosensors," *Anal. Chem.* 78, p.4261.

Patolsky, F., Zheng, G. and Lieber, C.M. (2006), "Nanowire sensors for medicine and the life sciences," *Nanomedicine* 1, p.51.

Samuelson, L. (2003), "Self-forming nanoscale devices," *Materials Today*, 6(10), p.22.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.

Silvapulle, M. J. (2007),"On the Existence of Maximum Likelihood Estimators for the Binomial Response Models," *Journal of the Royal Statistical Society. Series B*, 43(3), 310-313.

Wang, G. G. (2003), "Adaptive Response Surface Method using Inherited Latin Hypercube Design Points," *ASME Journal of Mechanical Design*, 125, 210-220.

Wang, X.D., Song, J. H., Liu, J., and Wang, Z. L. (2007)," Direct-Current Nanogenerator Driven by Ultrasonic Waves," *Science*, 316, 102-105.

Wang, Z. L. (2004), "Nanostructures of Zinc Oxide," *Materials Today*, 7(6), p.26.

Wang, Z. L. and Song, J. H. (2006), "Piezoelectric Nanogenerators Based on Zinc Oxide Nanowire Arrays," *Science* 312, p. 242-246.

Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Wu, C. F. J., Mao, S.S., and Ma, F. S. (1990), "SEL : A Search Method based on Orthogonal Arrays," In S. Ghosh (ed.), *Statistical Design and Analysis of Industrial Experiments*, Marcel Dekker Inc., New York, 279-310.

Yang, P. (2005), "Chemistry and physics of semiconductor nanowires," *MRS Bull.* 30, p.85.

Zheng, G. F., Patolsky, F., Cui, Y., Wang, W. U. and Lieber, C. M. (2005), "Multiplexed electrical detection of cancer markers with nanowire sensor arrays," *Nature Biotechnology* 23, p. 1294.

Zhou, S., (1999), *Electrodynamics of Solids and Microwave Superconductivity*, New York: Wiley.

APPENDIX A

PROOF OF THEOREM 3.1

Proof. **Part (a)** :

For simplicity, consider a single predictor variable and assume that $\eta_{ij} = \beta_j x_i$, where β_j is a scalar ($i = 1, 2, \ldots, N, j = 1, 2, 3$). Let $Q(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{N} \left(\sum_{j=1}^{3} y_{ij}\eta_{ij} - n_i \log\left(1 + \sum_{j=1}^{3} \exp\left(\eta_{ij}\right)\right)\right)$. Recall that $\beta_j^{(k)}$ denotes the estimate of β_j obtained after the k^{th} iteration. Then, it suffices to show that

(i) $Q(\beta_1, \beta_2, \beta_3)$ is a concave function of β_j, $j = 1, 2, 3$, and

(ii) $Q(\beta_1^{(k+1)}, \beta_2^{(k)}, \beta_3^{(k)}) \geq Q(\beta_1^{(k)}, \beta_2^{(k)}, \beta_3^{(k)})$.

It is easy to see that for $l = 1, 2, 3$,

$$\frac{\partial^2 Q}{\partial \beta_l^2} = -\sum_{i=1}^{N} \frac{n_i x_i^2 e^{\beta_l x_i}(1 + \sum_{j \neq l} e^{\beta_j x_i})}{(1 + \sum_{j=1}^{3} e^{\beta_j x_i})^2} \leq 0,$$

which proves the concavity of Q.

To prove (ii), we note that for given $\beta_2^{(k)}, \beta_3^{(k)}$ the solution for β_1 to the equation

$$\sum_{i=1}^{N} \left(y_{i1} - n_i \frac{e^{\beta_1 x_i}}{1 + e^{\beta_1 x_i} + \sum_{j=2}^{3} e^{\beta_j^{(k)} x_i}} \right) x_i = 0$$

maximizes $Q(\beta_1, \beta_2^{(k)}, \beta_3^{(k)})$.

From the first equation of (9) and steps 1-3 of the algorithm, we have

$$\sum_{i=1}^{N} \left(y_{i1} - n_i \frac{e^{\beta_1^{(k+1)} x_i}}{1 + e^{\beta_1^{(k+1)} x_i} + \sum_{j=2}^{3} e^{\beta_j^{(k)} x_i}} \right) x_i = 0,$$

which means $\beta_1^{(k+1)} = \arg\max Q(\beta_1, \beta_2^{(k)}, \beta_3^{(k)})$. Therefore (ii) holds.

Part (b) :

Again, for simplicity, consider a single predictor variable and assume that $\eta_{ij} = \beta_j x_i$ where β_j is a scalar ($i = 1, 2, \ldots, N, j = 1, 2, 3$). Let $\beta_j^{(k)}$ denote the estimate of β_j obtained after steps 1-3 of the k^{th} iteration and β_j^* denote the final estimate of β_j obtained by the proposed algorithm.

The estimated asymptotic variance of $\beta_1^{(k)}$, denoted by $s^2(\beta_1^{(k)})$, is given by the negative expectation of $\frac{\partial^2 \log L_{b1}}{\partial \beta_1^2}|_{\beta_1^{(k)}, \beta_2^{(k-1)}, \beta_3^{(k-1)}}$, where $\log L_{b1}$ denotes the binomial log-likelihood function of $y_{i1} (i = 1, \ldots N)$ that corresponds to the first of the three equations in (9) and is given by

$$\log L_{b1} = \sum_{i=1}^{N} \log \binom{n}{y_{i1}} + \sum_{i=1}^{N} y_{i1}(\eta_{i1} + \gamma_{i1}) - n_i \sum_{i=1}^{N} \log \Big(1 + \exp(\eta_{i1} + \gamma_{i1})\Big).$$

Now, $s^2(\beta_1^*)$, the estimated asymptotic variance of β_1^*, is given by the negative expectation of $\frac{\partial^2 \log L}{\partial \beta_1^2}|_{\beta_j = \beta_j^*, j=1,2,3}$, where $\log L$ is the multinomial likelihood given by (6).

It can easily be seen that

$$\frac{\partial^2 \log L_{b1}}{\partial \beta_1^2} = -\sum_{i=1}^{N} n_i x_i^2 \frac{1 + \exp(\eta_{i2}) + \exp(\eta_{i3})}{\big(1 + \exp(\eta_{i1}) + \exp(\eta_{i2}) + \exp(\eta_{i3})\big)^2} = \frac{\partial^2 \log L}{\partial \beta_1^2}.$$

By convergence of $\beta_j^{(k)}$ to β_j^* for $j = 1, 2, 3$, it follows that $s^2(\beta_1^{(k)}) \longrightarrow s^2(\beta_1^*)$.

Similarly, each component in the covariance matrix $\mathbf{\Sigma}_{\beta^{(k)}}$ can be proven to converge to each component of $\mathbf{\Sigma}_{\beta^*}$.

□

APPENDIX B

PROOF OF PROPOSITIONS 4.1 AND 4.2

Proof. Proposition 4.1 follows from the fact that, by construction, the sequence $\mathbf{x}_n$ has to consist of distinct points, and therefore has be finite.

To prove Proposition 4.2, note that if $\mathbf{x}_{n_0} = \mathbf{x}_g$ and $\alpha = 1/p(\mathbf{x}_g)$, then by (20), $q(\mathbf{x}_{n_0}) = 0$. Define

$$E_n(\mathbf{x}) = \sum_{i=1}^{n} \frac{q(\mathbf{x}_i)q(\mathbf{x})}{d(\mathbf{x}, \mathbf{x}_i)}.$$

Then, $E_n(\mathbf{x}_{n_0}) = 0$, which implies

$$\mathbf{x}_{n_0} = \arg \min_{\mathbf{x} \in [0,1]^m} E_n(\mathbf{x}).$$

By (22), $\mathbf{x}_{n_0+1} = \mathbf{x}_{n_0} = \mathbf{x}_g$. Following a similar argument, $\mathbf{x}_n = \mathbf{x}_g$ for all $n > n_0$.

□

VITA

Tirthankar Dasgupta was born in India. He received a Bachelor's degree in Statistics from University of Calcutta and two Master's degrees - M. Stat (Applied Statistics) and M. Tech (Quality, Reliability and Operations Research) from Indian Statistical Institute, Calcutta, India. He spent eight years as a faculty of the Statistical Quality Control and Operations Research Division of the Indian Statistical Institute, imparting advisory services and training to a wide spectrum of industries, prior to joining the School of Industrial and Systems Engineering at Georgia Institute of Technology as a doctoral student in 2003. In January 2008, he will join the Department of Statistics, Harvard University, as an Assistant Professor.