MULTI-STAGE EXPERIMENTAL PLANNING AND ANALYSIS FOR FORWARD-INVERSE REGRESSION APPLIED TO GENETIC NETWORK MODELING

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy
in the Graduate School of The Ohio State University

By

Cenny Taslim

****

The Ohio State University

2008

Dissertation Committee:  Approved By

Professor Theodore T. Allen, Adviser

Professor Hakan Ferhatosmanoglu  Industrial and Systems Engineering

Professor Clark Mount-Campbell  Graduate Program
ABSTRACT

This dissertation proposes methods for steady state linear system identification for both forward cases in which prediction of outputs for new inputs are desired and also inverse prediction of which inputs fostered measured outputs are needed. Special attention is given to genetic network modeling applications. Inverse prediction matters here because then one can predict the effective genetic perturbation associated with a new target drug compound or therapy. The primary application addressed in this dissertation is motivated by our on-going contributions related to Down syndrome which affects approximately 1 out of every 800 children.

First, single shot experimentation and analysis to develop network models is considered. The discussion focuses on linear models because of the relevance of equilibrium conditions and the typical scarcity of perturbation data. Yet, deviations from linear systems modeling assumptions are also considered. For system identification, we propose forward network identification regression (FNIR) and experimental planning involving simultaneously perturbing more than a single gene concentration using D-optimal designs. The proposed methods are compared with alternatives using
simulation and data sets motivated by the SOS pathway for Escherichia coli bacteria. Findings include that the optimal experimental planning can improve the sensitivity, specificity, and efficiency of the process of deriving genetic networks. In addition, topics for further research are suggested including the need to develop more numerically stable analysis methods, improved diagnostic procedures, sequential design and analysis procedures.

Next, multi-stage design and analysis procedures are proposed for experimentation in which both forward and inverse predictions are relevant. Methods are proposed to derive desirable experimental plans for the next batch of tests based on both space filling and D-optimality. The space filling designs are intended to support both linear and nonlinear modeling while D-optimality methods are relatively model-dependent. Rigorous results related to linear optimality criteria are presented in relation to multi-criteria formulations of the forward-inverse problem. Computational results are presented based on the SOS pathway and inspired by an on-going study of the genetic network associated with Down syndrome. In the studied cases, the biologists added a multiple choice constraint to the formulation for their simplicity.
Dedicated to my Mom and Dad.

In Loving Memory of the late Mr. Mulyana.
ACKNOWLEDGMENTS

I would like to thank my adviser, Dr. Theodore Allen. He is a harsh task master but sometimes he is less mean than other times. Dr. Mario Lauria, my external adviser, has been a constant source of information and inspiration. I have sincerely enjoyed all forms of his support. I would also like to thank my committee members, Drs. Clark Mount-Campbell and Hakan Ferhatosmanoglu. Dr. Mount-Campbell suggested that constrained regression be applied to address numerical issues which we have included as a topic for future study. Also, I thank Dr. Mount-Campbell for his help in developing the numerical results for the multi-stage problems. I would also like to thank Dr. Sen, whose insistence on rigorous optimality results inspired the development of new methods with desirable properties.

I am very grateful to Sandya Liyanarachchi and Dr. Ramana Davuluri. Ms. Liyanarachchi shared her expertise in the analysis of Microarray data. Dr. Ramana Davuluri gave me the opportunity to volunteer in his Bioinformatics and Computational Biology group. I would also like to thank Dr. William Notz for insightful and helpful discussion relating to statistics and academia. I would also like to express my gratitude
to Dr. Edwina Carreon for her guidance in writing dissertation and to my supervisor Cedric Sze for all of his support.

I am thankful for the assistance, friendship and camaraderie provided by my student colleagues. In particular, Dr. Nilgun Ferhatosmanoglu motivated my adviser to pursue biological applications. Drs. Ning Zheng and Mikhail Bernshteyn provided substantial help in developing numerical results throughout the dissertation.

Most importantly, I would like to thank my family and friends. My parents Taslim and Lenny Thamrin, my sisters Selliana, Pilliana and Olivia, and my brother Setiono, who always believe I can accomplish anything I set my mind to. My close friends, whose friendship I always treasure. Their perseverance and support has also been critical for maintaining my sanity. I thank them sincerely.

Finally, I would like to sincerely thank Rachmat Mulyana for his continuous support and endless encouragement. Without his support this dissertation would never be completed.
VITA

1977 .......................................................... Born – Jakarta, Indonesia

1999 .......................................................... BS Industrial and Systems Engineering
                             The Ohio State University
                             Columbus, OH

2008 .......................................................... MS Industrial and Systems Engineering
                             The Ohio State University
                             Columbus, OH

PUBLICATION

Maximize the Probability of Identifying The Important Factors,” The International
Journal of Industrial & Systems Engineering.

FIELDS OF STUDY

Major Field: Industrial and Systems Engineering

Minor Fields: Computer Science, Statistics, Bioinformatics
# TABLE OF CONTENTS

Abstract ........................................................................................................................................... ii  
Acknowledgments .............................................................................................................................. v  
Vita ..................................................................................................................................................... vii  
List of Tables ........................................................................................................................................ x  
List of Figures ........................................................................................................................................ xi  
Chapter 1  Introduction ..................................................................................................................... 1  
Chapter 2  First Stage Design and Analysis for Forward-Inverse system identification .... 6  
  Abstract ........................................................................................................................................... 6  
  2.1 Introduction ............................................................................................................................... 7  
  2.2 Genetic Networks ..................................................................................................................... 9  
  2.3 Existing Methods ..................................................................................................................... 13  
  2.4 Proposed Methods ................................................................................................................... 15  
  2.5 Criteria and Simulation ........................................................................................................... 18  
  2.6 Numerical Examples and Simulation Results ......................................................................... 23  
  2.7 Discussion and Future Work .................................................................................................... 42  
Acknowledgments ............................................................................................................................. 44  
Appendices ........................................................................................................................................ 45  
Chapter 3  Second Stage Design and Analysis for Forward-Inverse Analysis .................... 47  
  Abstract ........................................................................................................................................... 47  
  3.1 Introduction ............................................................................................................................... 48  
  3.2 Forward-Inverse Empirical Modeling .................................................................................... 50  
  3.3 Rigorous Results for D-Efficiency and Starved Data Cases ............................................... 54  
  3.4 Numerical Examples ............................................................................................................... 59
3.5 Computational Results .................................................................................. 63
3.6 Conclusions .................................................................................................. 69
Chapter 4 Conclusions and Future Research ................................................. 71
Bibliography ........................................................................................................ 76
LIST OF TABLES

Table 2.1 Amount of perturbation applied to each gene.......................................................... 16
Table 2.2 Example D-optimal designs for (a) six factors and (b) nine factors. .................. 17
Table 2.3 The nine gene SOS pathway: (a) genetic network and (b) the estimated A..... 23
Table 2.4 (a) “True” resilient A matrix, (b) Qualitative “true” A matrix, (c) X matrix run 1,
(d) $A_{est}$ run 1 with SEE = 0.66, (e) X matrix run 2, (f) $A_{est}$ run 2 with SEE = 350.5. .......... 25
Table 2.5 Test results based on 10,000 runs and random resilient matrices (RRM)....... 31
Table 2.6 Simulation results using the “recovered” E. coli A matrix and $n = m = 9$ ........ 40
Table 3.1 Three sets of test function coefficients for numerical examples ..................... 60
Table 3.2 The SOS network sets of function coefficients for numerical examples ......... 61
Table 3.3 The forward-inverse D-optimal design efficiencies for the test functions ...... 66
Table 3.4 The forward-inverse Maximin design efficiencies for the test functions ........ 69
LIST OF FIGURES

Figure 2.1  (a) A matrix and (b) The associated network.................................................. 12
Figure 2.2  Half normal plot for effects on coverage probability (w). ................................. 30
Figure 2.3  Interaction plot for coverage probability (w). .................................................... 32
Figure 2.4  Half-normal plot for effects on false positive probability................................. 33
Figure 2.5  AB interaction plot for false positive probability............................................. 34
Figure 2.6  BD interaction plot for false positive probability............................................. 35
Figure 2.7  Half-normal plot for effects on the natural logarithm of the ESEE................. 36
Figure 2.8  Interaction plot for the natural logarithm of ESEE. ........................................ 37
Figure 2.9  Half-normal plot for effects on the natural logarithm of the EASER............. 38
Figure 2.10 Interaction plot for the natural logarithm of EASER...................................... 39
Figure 2.11 Interaction plot for SOS pathway and the natural logarithm of ESEE .......... 41
Figure 3.1  Input and output experimental region for the first set of test function (F1) 62
Figure 3.2  Input and output experimental region for the 2nd order test function (F3) 62
Figure 3.3  AMPL code to implement Drezner and Erkut (1995) Maximin method........ 67
CHAPTER 1

INTRODUCTION

Design of experiments has been used in a wide range of fields from marketing and finance to computer science and manufacturing to help managers reduce cost and/or maximize profit. This dissertation will focus on planning optimal design of experiments motivated by the challenge of inferring genetic networks from data in the area of bioinformatics. The overall goal is to help scientists determine empirically the network of rules that govern specific biological organisms. These rules should be estimated in as cost effective and accurate a manner as possible. Hypothetical benefits include the more rapid development of drugs and gene therapies to help all types of organisms cope with their environments better including mitigating the effects of genetic disorders such as Down syndrome.

Many articles proposing design of experiments methods have been written in recent years relating to genetics. Primarily, these articles have concentrated on the
accuracy of the models derived from microarray experiments with the samples to be measured assumed to be given (e.g., see Kerr and Churchill, 2001; Churchill, 2002; Glonek et al., 2004; Khanin and Wit, 2004; Yang, 2003, and Ferhatomanoglu, Allen, Catalyurek, and Seillier-Moiseiwitsch, under review). For example, one might have two samples corresponding to individuals manifesting a disorder and two without that disorder. Then, the goal is to deploy resources to achieve the most accurate measurements possible.

In this dissertation, experiments are planned specifically to create samples that correspond to selected genetic perturbations. These perturbations are chosen to model the intricate web of interactions of different genes with as much accuracy as possible. In general, one is working on a subset of many thousands of genes which is selected by clustering because they have relatively strong promoter and inhibitor relationships. Therefore, the experiments considered here can be considered as a next step after initial clustering based on microarray data. Possible roles for the preliminary data used in clustering are briefly considered in Chapter 3.

Three special characteristics of genetic networks motivate the method development here. First, perhaps the most common measurements of biological systems are made in steady state. This means that linear systems models are likely to provide reasonable approximations to predict gene concentrations. This dissertation focuses on genetic networks identification at or near equilibrium points. Even though
cells are always growing and may never be in steady state (strictly speaking), they can be in a dynamic steady state. Cho, Gasparich, Sledgeski, Ezzell and Vermeulen (1984) define a state of dynamic equilibrium in E. coli as a period where the average rate of change in the cell remains constant. Also, biological subsystems might be in steady state to a good approximation and these steady state points may vary from subsystem to subsystem. In Gardner et al. (2003), steady state points are apparently achieved approximately 5.5 hours after the application of the perturbations (longer times for mammalian systems are generally expected).

For these steady state systems, the linear systems models used to approximate the genetic networks are equivalent to ordinary linear regression models with no constant terms. While this might seem like a minor detail, we will show that related numerical difficulties can pose significant challenges.

Second, in research related to genetic network modeling both forward and inverse predictions have importance. This follows because it is relatively easy to input known perturbations using so-called plasmids. Yet, it is less easy to predict the effective genetic perturbations associated with a compound of interest which could be a drug. Therefore, biologists are interested in applying many sets of known perturbations, measuring concentrations, fitting models, and then using these models to make inverse predictions about target compounds based on newly measured concentrations.
Third, genetic experiments are generally expensive. As a result, typically there only enough runs are available to barely fit a first order system model. Even while researchers may be skeptical of first order models of chemical behavior, it is not clear that models involving more parameters are viable.

These considerations motivate the five major problem statements to be addressed in this dissertation:

1. Relevant to first stage design and analysis, what modern experimental planning and analysis methods offer advantages for system identification with no constant term and with both forward and inverse identification goals? Presumably, these are the methods that are most relevant for “green field” network modelers who are starting to model a network for which no historical data is available.

2. Related to problem statement 1, what test beds and metrics can be developed to evaluate progress in related method development? As we will describe in Chapter 2, some test functions already exist but they do not incorporate all relevant information including the presence of equilibrium and the fact that biological systems are generally stable.

3. Relevant to second stage or sequential design and analysis, what methods offer advantages for system identification? This question should be answered with and without constant terms, with both forward and inverse prediction accuracy related goals, and in starved and non-starved data situations. Presumably, these
methods would be most relevant for modelers either in the middle of a large genetic network modeling project or for cases in which historical datasets are available and augmentation is possible.

4. Can rigorous results be generated to facilitate first and second stage experimental planning involving both forward and inverse related objectives? These rigorous methods should simplify array generation and ensure that any heuristics applied can generate optimal or near optimal solution quality.

5. Can the relevance of developed methods be established in the context of an on-going biological research program? The methods proposed here should be viable to guide a major on-going activity.

Chapter 2 focuses on the first two of these problem statements. Chapter 3 addresses the remaining three problems. Chapter 4 summarizes the conclusions.
ABSTRACT

The challenge of empirically deriving genetic networks that predict which genes are targeted by drug compounds is described. The discussion focuses on the inverse problem of predicting inputs from measured outputs in the context of linear systems in steady state. For system identification, we propose forward network identification regression (FNIR) and experimental planning involving simultaneously perturbing more than a single gene concentration using D-optimal designs. The proposed methods are compared with alternatives using simulation and data sets motivated by the SOS pathway for Escherichia coli bacteria. Findings include that the optimal experimental planning can likely improve the sensitivity, specificity, and efficiency of the process of
deriving genetic networks. Finally, topics for further research are suggested including the need to develop more numerically stable analysis methods, improved diagnostic procedures, and sequential design and analysis procedures.

1. INTRODUCTION

Many authors have considered the problem of empirical model building related to genetic systems (e.g., see Ambesi and di Bernardo, 2006, Bansal, Della Gatta, and di Bernardo, 2006, Basso, Margolin, Stolovitzky, Klein, Dalla-Favera, and Califano, 2005). Possible benefits of related research include aiding in the custom tailoring of drug treatments to specific individuals and predicting the side effects of new drugs. A subset of related research focuses on: (1) linear systems models and (2) the common case in which only steady state conditions are measured. In particular, Gardner, di Bernardo, Lorenz, and Collins (2003) propose methods based on one-factor-at-a-time (OFAT) experimentation and inverse subset regression. Presumably because of their focus on the biological engineering application, those authors made little effort to compare their approach with alternatives or to consider more sophisticated experimental plans than OFAT. In general, the topic of experimental design for genetic network identification has received relatively little attention even though the associated experiments are costly and many perturbations could be relevant (Rosa, Leon, and Rosa, 2006). One exception is Tegnér, Yeung, Hasty, and Collins (2003) who showed that efficiency can be gained by
randomly perturbing multiple genes simultaneously in a fully sequential experimentation and analysis procedure.

The primary purposes of this paper are to describe the experimental design problem of how to perturb cells to derive genetic networks and to suggest opportunities for future research. Also, we apply statistical simulation to compare the methods proposed by Gardner, Bernardo, Lorenz, and Collins (2003) with simple alternatives that we propose. Furthermore, the relevance of system “resilience” is considered. The concept of system resilience is currently of great interest to many communities (Hollnagel, Woods, and Leveson, 2006 and Fiksel 2003), and methods to accurately measure that resilience can help in building more profitable and long lasting enterprises. In the genetic network context, we describe the possible application of the estimated resilience as a modeling diagnostic. In addition, we propose a method for generating simulated system matrices with the resilience appropriate for modeling actual biological systems.

In Section 2, genetic networks are defined and the associated challenge for empirical model building is described. Section 3 details the experimental planning and modeling methods proposed by Gardner, Bernardo, Lorenz, and Collins (2003). Their experimental planning is based on one type of one-factor-at-a-time (OFAT) experimentation, and their analysis is a type of inverse regression. In Section 4, alternative planning procedures based on D-optimal designs and ordinary least squares
regression are proposed. Section 5 describes the simulation methodology used to compare the alternative procedures. Section 6 describes a simple numerical example and the results from two types of simulation experiments. The simulation results are based on the nine-gene sub-network example of the SOS pathway from Escherichia coli bacteria. Section 7 concludes with a discussion and opportunities for future research.

2. GENETIC NETWORKS

Deoxyribonucleic acid (DNA) effectively dictates a set of rules such that if certain entities are present other entities should be increased or reduced in concentration (Gardner et al., 2003). The entities include genes, proteins, and metabolites. Here, $x$ is an $m$-dimensional vector of these assorted concentrations which are the response in experimentation. For simplicity, we refer to all relevant entities as “genes” even though the mathematics applies to proteins and metabolites. The assumptions that define “genetic networks” include: (i) none of the $n$ perturbations considered $(u_1, \ldots, u_n)$ drives the network out of the basin of attraction of the stable steady-state point and (ii) the basic linear systems model (Shi, 2006, Chen, Lin, and Shamash, 2004, Ljung, 1999, and Paulo and Arbib, 1974) approximates behavior:

$$\frac{dx}{dt} = Ax + u + \varepsilon$$

\textbf{Equation 2.1}
where \( \mathbf{A} \) is the \( m \times m \) “system matrix”, \( \mathbf{u} \) is an \( m \)-dimensional vector of additive disturbance, \( t \) is time, and \( \varepsilon \) represent a combination of system fluctuations and measurement errors. In the biological context, \( \mathbf{A} \) is called the “genetic network” model. In general, this equation can describe an intrinsically linear system or a linearization of a nonlinear system near an equilibrium point (Neubert and Caswella, 1996). Also, considering that \( x_2 \) in \( \mathbf{x} \) might be the time derivative of \( x_1 \), Equation 2.1 can describe second and higher order linear differential equations.

The approximation of a genetic network using a linear systems model is largely for the sake of practicality because of the limited amount of available data. In most genetic experimentations, the number of runs is only marginally sufficient even if only linear models are considered. Also, a study done by Bansal, Belcastro, Ambesi-Impiombato and di Bernardo (2007) comparing different methods using in-silico data as well as real genetic data indicates that linear or steady-state assumptions are valid and can be used to correctly infer regulatory interactions between genes at or near a steady state point. Furthermore, the results of DREAM2 competition show that the linearity assumption does not hinder the algorithm ability to make reasonable predictions for non-linear networks (Lauria, Iorio, and di Bernardo, submitted for publication).

In steady state, concentrations are constant so that \( \mathbf{A} \mathbf{x} + \mathbf{u} + \varepsilon = \mathbf{0} \). Therefore, predictions using the system matrix, \( \mathbf{A} \), for perturbation, \( \mathbf{u}_p \), are given by the steady state solution to a stochastic differential equation:
\[ x_p = -A^{-1}u_p + \varepsilon \]

which again assumes that steady state can be reached under the associated perturbation and the linear approximation applies. As noted in Gardner et al. (2003), it is generally of greater interest to predict which unknown perturbation, \( u_p \), caused the measured, \( x_p \), using:

\[ u_p = -Ax_p + \varepsilon \]

This follows because researchers may want to know which genes are “active” when a compound is added to a cell. This inverse estimation aspect distinguishes genetic network modeling from many elementary linear systems identification problems.

Another complication of the biological context is that measurements of gene concentrations are often performed using so-called “microarray” experiments in which two samples are compared on the same slide because the efficient testing methods used provide more accurate comparison values than absolute measurements. As a result, researchers generally use scaled concentrations and perturbations to define the genetic network:

\[ x_i = \frac{[gene_i]_{\text{perturbed}}}{[gene_i]_{\text{unperturbed}}} - 1 \quad \text{and} \quad u_i = \frac{[gene_i]_{\text{added}}}{[gene_i]_{\text{unperturbed}}} \quad i = 1, \ldots, m \]

where the “[ ]” symbols refer to the concentrations of the genes in the brackets.
With the scaling in Equation 2.4, Gardner et al. (2003) use “negative regulation” to refer to values in $A$ less than $-1$, “positive regulation” to refer to values greater than $-1$, and 0 values as no regulation. Here, we expand the definition of zero regulation to include values within $\Delta$ of zero, with $\Delta = 0.1$ as the default value. This conforms to the intuition that small connection strengths can be negligible. Gardner et al. (2003) does not need this expansion since their analysis method generates coefficients that are exactly zero. For example, Figure 2.1 shows a system matrix $m = 5$ genes and the associated network. Circles denote negative regulation, and triangles denote positive regulation. Circles and triangles which do not have lines denote self inhibitor and self promoter respectively. Each row in the system matrix $A$ shows the influence of the gene associated with each column to the gene associated with each row after the application of the external perturbations. Thus in this example, recA promotes itself, lexA, recF and dinI.

$$
A = \begin{bmatrix}
-0.60 & -0.18 & -0.01 & 0 & 0.10 \\
0.39 & -1.67 & -0.01 & 0 & 0.09 \\
0.04 & -0.19 & -1.28 & 0 & 0.05 \\
-0.18 & 0.24 & -0.02 & -1 & -0.05 \\
0.28 & 0 & 0 & 0 & -2.09
\end{bmatrix}
$$

Figure 2.1 (a) A matrix and (b) The associated network
The key questions considered in this article are: (i) which set of perturbations, \(u_1, \ldots, u_n\) or "experimental design" should be applied? And (ii) which analysis method should be used to analyze the measured steady-state concentration responses, \(x_1, \ldots, x_n\)?

The goals are estimation of \(A\) for accurate prediction based on Equation 2.2 and Equation 2.3. Note that \(x_i\) is a multi-response vector associated with the input \(u_i\) for run \(i\) because all of the \(m\) gene concentrations are responses. Also, it is tempting to use the symbol \(y\) instead of \(x\) because these concentrations are measured outputs. However, \(y\) is typically reserved in the systems theory literature (Ljung, 1999) for outputs when \(x\) cannot be directly measured and the equation \(y = Bx\) is added to Equation 2.1.

3. EXISTING METHODS

For their experimental planning, Gardner et al. (2003) used an approach that we call the “one-factor-at-a-time” (OFAT) constant perturbation method such that in each experiment only one type of gene was added at a constant rate, i.e., over-expressed at a constant rate (through insertion of so-called “plasmids” which form sites for continual gene formation). If \(u_1, \ldots, u_n\) are the perturbation inputs associated with the \(n\) runs, the OFAT experimental plan (\(U\)) is:
OFAT

\[ U = \begin{bmatrix} u_1' & \vdots & u_n' \end{bmatrix} = \begin{cases} u_{i,j} = H_i & i \mod m = j \\ u_{i,j} = 0 & i \mod m \neq j \end{cases} \quad \forall i = 1, \ldots, n \quad j = 1, \ldots, m \]

where \( H_1, \ldots, H_n \) are the ranges selected for specific problems. The OFAT experimental plan is a diagonal matrix with \( H_i \) along its main diagonal. Gardner et al. (2003) focused on the \( n = m \) saturated case. Tegnér, Yeung, Hasty, and Collins (2003) proposed a fully sequential procedure based on an initial single gene perturbation. The study of sequential design and analysis methods is suggested for future work.

Gardner et al. (2003) proposed the network identification regression (NIR) algorithm to estimate \( A \) from the measured responses \( x_1, \ldots, x_n \). The NIR method is detailed in Appendix A, which is more detailed than in Gardner et al. (2003) and represents the computer code those authors used in their examples.

The central concept associated with the NIR method is the inverse regression of measured outputs \( (X) \) on inputs \( (U) \). For example, inverse regression is:

\[ A_{\text{est}} = -[(X'X)^{-1}X'U]' \quad \text{where} \quad U = (u_1 \mid \ldots \mid u_n)' \quad \text{and} \quad X = (x_1 \mid \ldots \mid x_n)', \]

which assumes that \( X'X \) is full rank. The NIR method involves fitting models similar to Equation 2.6 except subset regression is applied. \( \tilde{X}_i \) is defined as the design matrix associated with only the genes included in subset \( i \). In the Gardner et al. (2003) method the set of models under consideration, \( S \), contains only models having exactly \( m^* \) genes.
included, where \( m^* \) is selected by testing the dynamic stability of the models. For simplicity and also following Gardner et al. (2003) choice of \( m^* = 5 \) for a nine gene network as a dynamically stable model, in our examples we fixed \( m^* = 3 \) for six gene problems and \( m^* = 5 \) for nine gene problems.

The subset approach offers the potential benefit that the \( \tilde{X}'X \), matrices are more likely to be well conditioned than the \( XX' \) associated with all genes in Equation 2.6. Also, it is widely believed that the genetic networks are sparse, so forcing several terms to equal zero has intuitive appeal. Yet, cases can easily be produced such that all the \( \tilde{X}'X \) are ill conditioned and the NIR algorithm returns a zero matrix for \( A \). In general, the conditioning of \( XX' \) or “design matrices” is a central concern in regression because poorly conditioned matrices can cause random errors to have a large effect on estimation results (e.g., see Myers and Montgomery, 2002, p. 393).

4. PROPOSED METHODS

In this section, alternative methods are proposed for both experimental planning and analysis. The primary motivation of these methods is to improve estimation accuracy by increasing the conditioning properties of matrices being inverted. Also, the proposed alternatives are intended to be representative of the simplest methods from the experimental design and system identification literatures (e.g., see Ljung, 1999).
For experimentation, we propose using D-optimal designs assuming the fitted model is first order with no constant term. Myers and Montgomery (2002) describe the construction of D-optimal designs which can be accomplished using widely available software. Examples of putatively D-optimal designs generated using genetic algorithm (used in the numerical studies) are in Table 1. Each gene is perturbed with a prescribed concentration selected to ensure the system returns to its original steady-state point. In each experiment, a gene is either perturbed with the prescribed concentration or not perturbed at all. The amount of perturbation applied to each gene is formulated in the following table:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.65</td>
<td>1.17</td>
<td>13.41</td>
<td>1.67</td>
<td>4.54</td>
<td>2.36</td>
<td>4.71</td>
<td>12.87</td>
<td>4.11</td>
</tr>
</tbody>
</table>

Table 2.1 Amount of perturbation applied to each gene

Note that because the constant term is omitted, generating D-optimal designs using (−1,1) ranges and then shifting and scaling does not generally produce D-optimal or even competitive designs, so design generation must be accomplished in engineering units. This fact is proven by a simple theorem in Chapter 3.
Table 2.2 Example D-optimal designs for (a) six factors and (b) nine factors.

Gardner et al. (2003) indicated that experimentation with \( n < m \) is of future interest because of the large size of practical networks and the costs involved. Therefore, we propose the application of the supersaturated designs, e.g., adapted from Allen and Bernshteyn (2003) methods for future study. In chapter 3, we present multi-stage design of experiments based on Maximin and D-optimal approach for the starved case.

Next, for analysis we propose the “forward network identification regression (FNIR) algorithm” based on ordinary least squares regression. The FNIR algorithm is called forward because it involves regressing the controlled inputs \((u_1, \ldots, u_n)\) on the outputs \((x_1, \ldots, x_n)\). It is simply estimation using Equation 2.7:

\[
A^{-1}_{est} = -[(U'U)^{-1}U'X]' 
\]
Since the FNIR algorithm fits the same functional form to all responses, \( y_1, \ldots, y_n \), Equation 2.7 generates the best linear unbiased estimation (BLUE) regardless of the correlation between random errors (Khuri and Cornell, 1996, p. 254). A major limitation of FNIR in the context of saturated experimentation with \( n = m \) is the fact that there are no degrees of freedom for model diagnostics. Therefore, we describe the application of estimated resilience as model diagnostics. The next section focuses on method evaluation and includes a “resilience” diagnostic that can be used with FNIR based analysis.

5. CRITERIA AND SIMULATION

In this section, we review the criteria and simulation methods from Gardner, Bernardo, Lorenz, and Collins (2003) and propose additional approaches.

Criteria

The “coverage percentage” referred to by Gardner, Bernardo, Lorenz, and Collins (2003) can be expressed as a probability (similar to the power in Allen and Bernshteyn, 2003):

\[
\text{Coverage probability} = w = E \left\{ \frac{H}{T} \right\}
\]

Equation 2.8
where $E$ is the expected value operator, $T$ is the true number of positive or negative regulation connections in the network, $H$ is the number of correctly identified, nonzero connections or “hits” by the algorithm. Therefore, coverage probability quantifies the “sensitivity” of the method. Also, the “false positive percentage” from Gardner, Bernardo, Lorenz, and Collins (2003) expressed as a probability is:

$$\text{False positive probability} = E\left\{ \frac{M}{H + M} \right\}$$

where $M$ is total number of connections which are identified incorrectly or “missed” by the algorithm, i.e., connections declared to be nonzero when the true associated connections are either zero or different in sign than those identified. Therefore, false positive probability quantifies the “specificity” of the method. For example, if the true network is shown in Figure 2.1(b) and the algorithm identified recA as having only positive self regulation and recF as having negative regulation on recA, then $T = 11$, $H = 1$, and $M = 11$.

To these criteria, we add the “expected squared estimation errors” (ESEE):

$$\text{ESEE} = m^{-2} E\left\{ \sum_{j=1}^{m} \sum_{i=1}^{m} (A_{est,i,j} - A_{i,j})^2 \right\}$$

Thus, the ESEE summarizes the likely quantitative errors in the estimations, $A_{est}$. 
Also, the most commonly used measure describing the “resilience” of a system is defined using the eigenvalues of $A$ (Neubert and Caswell, 1996). Denote the eigenvalues of $A$ as $\lambda_1(A), \ldots, \lambda_m(A)$ ordered largest to smallest by their real part. The resilience is:

$$\text{Equation 2.11}$$

$$\text{Resilience} = -\text{Real}[\lambda_1(A)]$$

Where $\text{Real}[z]$ is a function that returns the real part of the complex number $z$.

Another criterion is “expected absolute error in resilience” (EASER):

$$\text{Equation 2.12}$$

$$\text{EASER} = \mathbb{E}\{ |\text{Real}[\lambda_1(A)] - \text{Real}[\lambda_1(A_{\text{est}})] | \}$$

which is relevant partly because positive resilience is typical of actual biological systems so estimated resilience can be used as a model diagnostic. If the estimated resilience is negative and/or very large, the analyst should suspect that the empirical model cannot be trusted.

**Simulation**

Gardner, Bernardo, Lorenz, and Collins (2003) also proposed assumptions for their method simulations. In their assumptions, they randomly generated a network and then arbitrarily made up coefficients in $A$ consistent with that network. This approach is problematic in that it is likely to generate an unrealistic network with negative
resilience. Therefore, their generated networks will likely be unable to endure some of the perturbations in a set of experiments, i.e., the system will never reach equilibrium and Equation 2.2 and Equation 2.3 are not relevant.

For our simulations, we apply two approaches for generating true A matrices. First, we propose “random resilient matrices” (RRM) based on random diagonal matrices, D, and matrices of normally distributed random variables, Q, and independent, identically distributed (IID) random variables, Zj for j = 1,...,m and αjk for k, j = 1,...,m.

RRM

Repeat { λj ~ |Zj|  Zj ~ N(mean = 0, standard deviation = σA) for j = 1,...,m
αjk ~ IID N(mean = 0, standard deviation = σA) for k, j = 1,...,m

D = \begin{pmatrix} λ_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & λ_m \end{pmatrix}  Q = \begin{pmatrix} α_{11} & \cdots & α_{1m} \\ \vdots & \ddots & \vdots \\ α_{m1} & \cdots & α_{mm} \end{pmatrix}  D^* = \begin{pmatrix} λ_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & λ_m^{-1} \end{pmatrix}

A = Q^{-1}DQ and A^{-1} = Q^{-1}D^*Q}

Until { condition number(A) < K }

where the condition number is the ratio of the largest to the smallest eigenvalue, and we consider two cut-off values K = 10 and K = 100 and σA = 1. While the RRM simulation method is guaranteed to generate positively resilient systems matrices, it has two types of limitations. First, the generated A might not have a realistic distribution of positive,
negative, and zero regulations. Second, A could be associated with non-physical, negative values for concentrations simulated using \( x = A^{-1}u \). In the biological case, this might not be a problem if \( x \) refers to differential between an ambient concentration and the actual concentration according to the scaling in Equation 2.2. Yet, investigating methods that can efficiently generate positive solutions and realistic distribution of regulations are proposed as a subject for future research. For these, “sign solvable” systems (Brualdi and Shader, 1995) might be helpful.

Second, we simulated “true” A matrices based on the SOS network shown in Table 2.3 (a) and (b) which includes the “recovered” or estimated A matrix as derived by Gardner, Bernardo, Lorenz, and Collins (2003). Once A is generated using either approach, simulation is accomplished using the steady state formula:

\[
X = A^{-1}U + \varepsilon
\]

where the \( \varepsilon_{ij} \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \) have potentially unequal standard deviations for different genes, \( \sigma_{ij} \) for \( j = 1, \ldots, m \).
We considered two assumptions about the random errors: \( \sigma_{ij} = 0.1 \) and \( \sigma_{ij} = 1.0 \) for all \( j = 1, \ldots, m \). The complications of unequal standard errors are recommended for future study. The standard deviations of the random errors in Gardner et al. (2003) showed no strong relationship with the ranges or the values in the estimated \( A \) matrix.

### 6. NUMERICAL EXAMPLES AND SIMULATION RESULTS

#### Numerical Example

We begin with a simple example involving six genes to illustrate the proposed simulation and estimation procedures. Table 2.4(a) shows an \( A \) matrix generated using the random resilient matrices (RRM) method. Table 2.4(b) shows the qualitative system connections associated with the generated \( A \) matrix. The D-optimal experimental design used is in Table 2.2. Table 2.4(c) shows one set of simulated concentrations (run 1)
derived from adding random errors with standard deviations $\sigma_{ij} = 1.0$ for all $j = 1,\ldots,6$. Table 2.4(d) shows the estimated or “recovered” $A$ matrix derived using straightforward least squares estimation in Equation 2.7. The empirical sum squared estimation errors (SEE) for the derived $A_{est}$ are 0.66. The remaining parts (e) and (f) are based on a second set of random errors (run 2) resulting in a much higher SEE equal to 350.5.

This example illustrates that the prediction accuracy of the proposed methods can be highly sensitive to random errors. Specifically, the true overall number of non-zero connections is $T = 27$. In the first run, the number of correctly identified connections is $H = 17$, and the number of misses is $M = 18$ (Note: entry(4,1) is considered a miss). For the second run, the numbers are $H = 15$ and $M = 20$. In practice, such errors and variation could derive from problems in measuring gene concentrations and contaminants in the experimental system.
Random Resilient Matrices Simulation Results

Next, a full factorial Monte Carlo simulation experiment was performed using random resilient matrices (RRM) with results in Table 2.5. This experiment involved all combinations of experimental planning method, analysis methods, factor numbers ($m$), and resilient matrices. The results are presented in Table 2.4 below:

Table 2.4: Simulation results for resilience.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>Gene</td>
</tr>
<tr>
<td>Gene</td>
<td>Gene</td>
</tr>
<tr>
<td>Gene</td>
<td>Gene</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
<th>Gene 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.61</td>
<td>1.48</td>
<td>2.38</td>
<td>0.18</td>
<td>-3.38</td>
<td>6.38</td>
</tr>
<tr>
<td>2</td>
<td>7.92</td>
<td>3.57</td>
<td>14.55</td>
<td>3.66</td>
<td>-3.08</td>
<td>5.18</td>
</tr>
<tr>
<td>3</td>
<td>8.96</td>
<td>7.68</td>
<td>15.37</td>
<td>2.42</td>
<td>-1.56</td>
<td>9.79</td>
</tr>
<tr>
<td>4</td>
<td>10.38</td>
<td>6.05</td>
<td>15.69</td>
<td>-0.37</td>
<td>-1.63</td>
<td>9.32</td>
</tr>
<tr>
<td>5</td>
<td>2.27</td>
<td>1.23</td>
<td>-0.21</td>
<td>3.80</td>
<td>1.27</td>
<td>2.73</td>
</tr>
<tr>
<td>6</td>
<td>-0.12</td>
<td>0.31</td>
<td>1.12</td>
<td>0.31</td>
<td>7.80</td>
<td>-1.54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
<th>Gene 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.10</td>
<td>3.66</td>
<td>14.39</td>
<td>1.73</td>
<td>-5.96</td>
</tr>
<tr>
<td>2</td>
<td>11.07</td>
<td>6.98</td>
<td>16.25</td>
<td>3.22</td>
<td>-1.72</td>
</tr>
<tr>
<td>3</td>
<td>10.48</td>
<td>8.16</td>
<td>15.29</td>
<td>0.04</td>
<td>-1.61</td>
</tr>
<tr>
<td>4</td>
<td>3.46</td>
<td>-0.17</td>
<td>3.55</td>
<td>1.95</td>
<td>-2.91</td>
</tr>
<tr>
<td>5</td>
<td>-0.52</td>
<td>2.88</td>
<td>1.87</td>
<td>2.54</td>
<td>5.49</td>
</tr>
<tr>
<td>6</td>
<td>-26.36</td>
<td>-9.70</td>
<td>20.42</td>
<td>-4.26</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
<th>Gene 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-36.50</td>
<td>-14.54</td>
<td>28.69</td>
<td>-6.10</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>13.55</td>
<td>4.11</td>
<td>-10.06</td>
<td>1.38</td>
<td>0.51</td>
</tr>
<tr>
<td>3</td>
<td>60.42</td>
<td>23.48</td>
<td>-48.33</td>
<td>10.55</td>
<td>-0.74</td>
</tr>
<tr>
<td>4</td>
<td>-28.09</td>
<td>-9.93</td>
<td>21.42</td>
<td>-4.75</td>
<td>-0.14</td>
</tr>
<tr>
<td>5</td>
<td>-20.71</td>
<td>-6.91</td>
<td>15.32</td>
<td>-2.89</td>
<td>-1.18</td>
</tr>
<tr>
<td>6</td>
<td>-26.36</td>
<td>-9.70</td>
<td>20.42</td>
<td>-4.26</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

Table 2.4 (a) “True” resilient A matrix, (b) Qualitative “true” A matrix, (c) X matrix run 1, (d) $A_{est}$ run 1 with SEE = 0.66, (e) X matrix run 2, (f) $A_{est}$ run 2 with SEE = 350.5.
at two levels, random error standard deviations \((\sigma_{ij})\) at two levels, and two choices of condition number cut-off \((K)\). In all cases, the number of runs, \(n\), equals the number of genes, \(m\), to compare with the OFAT method in Gardner, Bernardo, Lorenz, and Collins (2003), i.e., \(n = m\). In each simulation, 10,000 replications are applied resulting in approximately three significant digits for all numbers except for the ESEE and EASER values, which only have a single digit because of the high sample variances. In all cases MATLAB was used with its default double precision. Also, MATLAB performs numerical inversions using Gaussian elimination with partial pivoting in general and Cholesky factorization for the positive definite, symmetric matrices.

The Gardner, Bernardo, Lorenz, and Collins (2003) method returns zero matrices in some cases because no models fitted pass their condition number test (see appendix). The second-to-last column of Table 2.5 shows counts of the number of times the methods return zero matrices. The run times on a 3.2 GHz Pentium IV with 1 GB RAM computer are provided in the last column of the table. Clearly, the relatively simple forward network identification regression (FNIR) method offers a potentially important advantage in computational efficiency compared with the subset regression based network identification regression (NIR) method. For example, in some cases FNIR required less than 1.4% of the time of NIR.

Figure 2.2 through Figure 2.10 show analysis results of the full factorial experiment in Table 2.5. The half normal plots show the factors and their interactions
proven to have significant effects using Lenth’s method with experiment-wise error rate equal to 0.05. For example, Figure 2.2 and Figure 2.3 together show that the simple forward network identification regression (FNIR) method proposed in section 4 achieves significantly higher coverage probability than the Gardner, Bernardo, Lorenz, and Collins (2003) network identification regression (NIR) method (factor B). Also, predictably the performance of both methods significantly degrades as the experimental error standard deviations, $\sigma_{ij}$ for $j = 1, \ldots, m$ (factor D), increase.

Note that the NIR method is applied using $m^* = 3$ for six gene problems and $m^* = 5$ for nine gene problems. Also, choosing higher values of $m^*$ does not, automatically, improve the NIR method coverage because the chance of returning a zero matrix increases. The other statistically significant effects, while possibly important, are relatively small. The results show that the condition number cut-off on the system matrix (factor E) has minimal, if any, effect on the coverage probability associated with the FNIR method, within the range $K = 10$ to $K = 100$. This holds despite the fact that the proposed FNIR method involves the inversion of the $A^{-1}_{est}$ matrix.

Figure 2.4 shows the significant effects of the experimental factors on the false positive probability. The results show the effect of the experimental planning approach which interacts with the analysis method as illustrated in Figure 2.5. The proposed forward network identification regression (FNIR) method has its performance significantly enhanced through the application of D-optimal experimental plans. This
follows presumably because of the positive effects on the properties of the $U'U$ matrix, which is inverted in that method. It is not surprising, perhaps, that the NIR method performance is minimally effected by improved properties of $U'U$ since this matrix plays no direct role in that method.

The largest effects on false positive performance were the analysis method choice (factor B), the magnitude of the standard error (factor D), and their interaction (BD). Figure 2.6 shows that the FNIR method achieves lower false positive probabilities when the experimental errors are smaller in magnitude (factor D). However, as experimental errors increase, the FNIR false positive probability increases to the extent that the NIR method compares favorably. This occurs (in part) because of the many zero matrices returned by the NIR method for difficult cases and because the subset regression approach (by design) limits the possible number of false positives. Yet, the differences might not be practically important.

As noted previously, even with 10,000 simulation runs, the expected squared estimation errors (ESEE) of the estimated system matrix, $A_{est}$, and the expected squared errors of the resilience estimates (EASER) cannot be evaluated with greater than a single significant digit. Also, the Monte Carlo estimation errors are generally highly correlated with the mean values. To stabilize the variance and reduce the effects of outliers, we used the natural logarithm transformation of the estimated ESEE and EASER values in our analyses.
The results in Figure 2.7 and Figure 2.8 together establish the significant and substantial improvement in the estimation accuracy associated with the network identification regression (NIR) method from Gardner, Bernardo, Lorenz, and Collins (2003) compared with the simple forward network identification regression (FNIR) method proposed here. Yet, much of the relative advantage of the NIR method can be attributed to the zero matrices returned by the NIR method in difficult cases. In real situations, if the NIR method returned all zeros, the analysts would either try a different analysis method or request additional data. Similarly, using the FNIR method, in a small number of cases, the estimated system matrix \((A)\) entries are very large indicating exceedingly strong connections. In real applications, these non-physical values would likely send up “red flags”, and the experimenter would try a different analysis method and/or request additional data. This suggests a need for additional strategies to augment existing data.

Figure 2.9 and Figure 2.10 together indicated that the estimated resilience values are often highly inaccurate. As the standard deviations of the random errors \((\sigma_{e_j})\) increase the estimated values typically differ by more than 5 units from the true values. Using the forward network identification regression (FNIR) method, the estimated resilience might be the only available diagnostic. Very large or negative values of the estimate resilience could signal that the estimated system matrix is unreliable and alternative analyses and/or additional data is needed.
To understand the high errors associated with the FNIR method, consider the conditioning of the $U'U$ design matrix. The choice of D-optimal design generally improves the coverage, logarithm ESEE, and logarithm EASER. However, even using the D-optimal design, the diverse experimental ranges cause the condition numbers of $U'U$ to equal 993 for 6 runs and 1093 for 9 runs. Therefore, the absence of a constant term in the model causes even an optimally efficient design to have poor conditioning properties. Using an identity matrix in place of $U$ results in ESEE and EASER values less than 20 for all cases considered.

![Half normal plot for effects on coverage probability ($w$)](image)

**Figure 2.2** Half normal plot for effects on coverage probability ($w$)
<table>
<thead>
<tr>
<th>DOE Type</th>
<th>Analysis Method</th>
<th>$n = m$</th>
<th>$\sigma_{ij}$</th>
<th>$K$</th>
<th>Coverage Prob. ($w$)</th>
<th>False positive</th>
<th>$A_{est}$ Error (ESEE)</th>
<th>Resilience Error (EASER)</th>
<th>#Zero Matrices</th>
<th>Time (Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>6</td>
<td>1.0</td>
<td>100</td>
<td>0.667</td>
<td>0.354</td>
<td>5552.21</td>
<td>4.291</td>
<td>0</td>
<td>76.1</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>6</td>
<td>1.0</td>
<td>100</td>
<td>0.342</td>
<td>0.281</td>
<td>0.50</td>
<td>0.408</td>
<td>122</td>
<td>241.2</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>6</td>
<td>1.0</td>
<td>100</td>
<td>0.678</td>
<td>0.347</td>
<td>1584.22</td>
<td>4.320</td>
<td>0</td>
<td>76.4</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>6</td>
<td>1.0</td>
<td>100</td>
<td>0.335</td>
<td>0.287</td>
<td>0.49</td>
<td>0.359</td>
<td>292</td>
<td>231.7</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>9</td>
<td>1.0</td>
<td>100</td>
<td>0.680</td>
<td>0.353</td>
<td>2051.43</td>
<td>5.108</td>
<td>0</td>
<td>57.2</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>9</td>
<td>1.0</td>
<td>100</td>
<td>0.384</td>
<td>0.269</td>
<td>0.35</td>
<td>0.435</td>
<td>136</td>
<td>1479.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>9</td>
<td>1.0</td>
<td>100</td>
<td>0.701</td>
<td>0.343</td>
<td>203761.66</td>
<td>19.639</td>
<td>0</td>
<td>58.0</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>9</td>
<td>1.0</td>
<td>100</td>
<td>0.359</td>
<td>0.297</td>
<td>0.32</td>
<td>0.312</td>
<td>696</td>
<td>1307.6</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>6</td>
<td>0.1</td>
<td>100</td>
<td>0.890</td>
<td>0.116</td>
<td>35.98</td>
<td>0.292</td>
<td>0</td>
<td>83.1</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>6</td>
<td>0.1</td>
<td>100</td>
<td>0.371</td>
<td>0.226</td>
<td>0.37</td>
<td>0.182</td>
<td>304</td>
<td>246.2</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>6</td>
<td>0.1</td>
<td>100</td>
<td>0.907</td>
<td>0.097</td>
<td>10.91</td>
<td>0.093</td>
<td>0</td>
<td>83.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>6</td>
<td>0.1</td>
<td>100</td>
<td>0.356</td>
<td>0.245</td>
<td>0.38</td>
<td>0.206</td>
<td>443</td>
<td>233.3</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>9</td>
<td>0.1</td>
<td>100</td>
<td>0.887</td>
<td>0.124</td>
<td>123.58</td>
<td>0.298</td>
<td>0</td>
<td>87.7</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>9</td>
<td>0.1</td>
<td>100</td>
<td>0.402</td>
<td>0.218</td>
<td>0.24</td>
<td>0.179</td>
<td>354</td>
<td>1412.7</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>9</td>
<td>0.1</td>
<td>100</td>
<td>0.925</td>
<td>0.080</td>
<td>2.51</td>
<td>0.065</td>
<td>0</td>
<td>89.2</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>9</td>
<td>0.1</td>
<td>100</td>
<td>0.367</td>
<td>0.271</td>
<td>0.26</td>
<td>0.193</td>
<td>917</td>
<td>1307.9</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>6</td>
<td>1.0</td>
<td>10</td>
<td>0.687</td>
<td>0.413</td>
<td>2138.36</td>
<td>5.164</td>
<td>0</td>
<td>81.8</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>6</td>
<td>1.0</td>
<td>10</td>
<td>0.413</td>
<td>0.311</td>
<td>0.25</td>
<td>0.620</td>
<td>0</td>
<td>256.6</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>6</td>
<td>1.0</td>
<td>10</td>
<td>0.699</td>
<td>0.400</td>
<td>3567.57</td>
<td>3.757</td>
<td>0</td>
<td>82.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>6</td>
<td>1.0</td>
<td>10</td>
<td>0.425</td>
<td>0.289</td>
<td>0.23</td>
<td>0.534</td>
<td>0</td>
<td>245.9</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>9</td>
<td>1.0</td>
<td>10</td>
<td>0.691</td>
<td>0.445</td>
<td>20502.50</td>
<td>9.084</td>
<td>0</td>
<td>13.0</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>9</td>
<td>1.0</td>
<td>10</td>
<td>0.462</td>
<td>0.343</td>
<td>0.19</td>
<td>0.707</td>
<td>0</td>
<td>1034.0</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>9</td>
<td>1.0</td>
<td>10</td>
<td>0.717</td>
<td>0.416</td>
<td>8339.31</td>
<td>9.835</td>
<td>0</td>
<td>14.1</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>9</td>
<td>1.0</td>
<td>10</td>
<td>0.489</td>
<td>0.299</td>
<td>0.13</td>
<td>0.476</td>
<td>0</td>
<td>1003.6</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>6</td>
<td>0.1</td>
<td>10</td>
<td>0.899</td>
<td>0.119</td>
<td>0.10</td>
<td>0.038</td>
<td>0</td>
<td>81.8</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>6</td>
<td>0.1</td>
<td>10</td>
<td>0.515</td>
<td>0.138</td>
<td>0.07</td>
<td>0.163</td>
<td>0</td>
<td>257.4</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>6</td>
<td>0.1</td>
<td>10</td>
<td>0.913</td>
<td>0.100</td>
<td>0.02</td>
<td>0.021</td>
<td>0</td>
<td>82.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>6</td>
<td>0.1</td>
<td>10</td>
<td>0.501</td>
<td>0.147</td>
<td>0.08</td>
<td>0.210</td>
<td>0</td>
<td>246.3</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>9</td>
<td>0.1</td>
<td>10</td>
<td>0.896</td>
<td>0.128</td>
<td>0.14</td>
<td>0.028</td>
<td>0</td>
<td>13.6</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>9</td>
<td>0.1</td>
<td>10</td>
<td>0.589</td>
<td>0.130</td>
<td>0.04</td>
<td>0.153</td>
<td>0</td>
<td>1022.8</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>9</td>
<td>0.1</td>
<td>10</td>
<td>0.929</td>
<td>0.083</td>
<td>0.01</td>
<td>0.014</td>
<td>0</td>
<td>13.9</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>9</td>
<td>0.1</td>
<td>10</td>
<td>0.583</td>
<td>0.126</td>
<td>0.04</td>
<td>0.172</td>
<td>0</td>
<td>992.2</td>
</tr>
</tbody>
</table>

Table 2.5 Test results based on 10,000 runs and random resilient matrices (RRM).
Figure 2.3 Interaction plot for coverage probability ($w$).
Figure 2.4 Half-normal plot for effects on false positive probability.
Figure 2.5 AB interaction plot for false positive probability
Figure 2.6 BD interaction plot for false positive probability

<table>
<thead>
<tr>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

False Positive Probability

D:0.1  D:1.0

Figure 2.6 BD interaction plot for false positive probability
Figure 2.7 Half-normal plot for effects on the natural logarithm of the ESEE
Figure 2.8 Interaction plot for the natural logarithm of ESEE.
Figure 2.9 Half-normal plot for effects on the natural logarithm of the EASER
**SOS Pathway Simulation Results**

For the SOS pathway simulations, only nine runs are studied because Gardner, Bernardo, Lorenz, and Collins (2003) argue that these nine genes form a natural set. The condition number for the system A matrix in Table 2.3(b) is 9.25. In all cases, the ranges for the experimental plans as in Table 2.1 were used. To facilitate the application of hypothesis testing three factors (i.e. DOE type, analysis method and $\sigma_{e}$), two replicates each representing 10,000 runs are provided in Table 2.6. In general, the results were similar to those for random resilient matrices (RRM), and we omit them for space...
reasons. Specifically, for coverage probability, the D-optimal design and FNIR offered benefits compared with OFAT and NIR. For false positives, D-optimal designs and NIR were beneficial but the effects were small and any interaction could not be detected between these factors.

<table>
<thead>
<tr>
<th>DOE Type</th>
<th>Analysis Method</th>
<th>$\sigma_{ij}$</th>
<th>Coverage Prob. (w)</th>
<th>False positive</th>
<th>$A_{res}$ Error (ESEE)</th>
<th>Resilience Error (EASER)</th>
<th>#Zero Matrices</th>
<th>Time (Sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>1</td>
<td>0.6312</td>
<td>0.7744</td>
<td>9727.4</td>
<td>12.8325</td>
<td>0</td>
<td>12.9</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>1</td>
<td>0.4692</td>
<td>0.7074</td>
<td>0.549</td>
<td>1.2625</td>
<td>0</td>
<td>1032.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>1</td>
<td>0.6866</td>
<td>0.7460</td>
<td>2918.7</td>
<td>11.3578</td>
<td>0</td>
<td>13.9</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>1</td>
<td>0.5395</td>
<td>0.6552</td>
<td>0.442</td>
<td>0.8932</td>
<td>0</td>
<td>958.5</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>0.1</td>
<td>0.8990</td>
<td>0.3665</td>
<td>0.025</td>
<td>0.0018</td>
<td>0</td>
<td>69.7</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>0.1</td>
<td>0.8240</td>
<td>0.2909</td>
<td>0.019</td>
<td>0.0017</td>
<td>0</td>
<td>1538.9</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>0.1</td>
<td>0.9281</td>
<td>0.2588</td>
<td>0.009</td>
<td>0.0011</td>
<td>0</td>
<td>35.4</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>0.1</td>
<td>0.8864</td>
<td>0.1790</td>
<td>0.007</td>
<td>0.0010</td>
<td>0</td>
<td>1352.8</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>1</td>
<td>0.6307</td>
<td>0.7744</td>
<td>3238.7</td>
<td>10.5275</td>
<td>0</td>
<td>79.4</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>1</td>
<td>0.4692</td>
<td>0.7072</td>
<td>0.551</td>
<td>1.2675</td>
<td>0</td>
<td>1687.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>1</td>
<td>0.6852</td>
<td>0.7464</td>
<td>26995409.8</td>
<td>10.1937</td>
<td>0</td>
<td>80.4</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>1</td>
<td>0.5390</td>
<td>0.6554</td>
<td>0.440</td>
<td>0.8918</td>
<td>0</td>
<td>1566.5</td>
</tr>
<tr>
<td>OFAT</td>
<td>FNIR</td>
<td>0.1</td>
<td>0.8986</td>
<td>0.3664</td>
<td>0.025</td>
<td>0.0019</td>
<td>0</td>
<td>78.7</td>
</tr>
<tr>
<td>OFAT</td>
<td>NIR</td>
<td>0.1</td>
<td>0.8242</td>
<td>0.2909</td>
<td>0.019</td>
<td>0.0017</td>
<td>0</td>
<td>1516.2</td>
</tr>
<tr>
<td>D-opt</td>
<td>FNIR</td>
<td>0.1</td>
<td>0.9280</td>
<td>0.2583</td>
<td>0.009</td>
<td>0.0011</td>
<td>0</td>
<td>60.3</td>
</tr>
<tr>
<td>D-opt</td>
<td>NIR</td>
<td>0.1</td>
<td>0.8860</td>
<td>0.1805</td>
<td>0.007</td>
<td>0.0010</td>
<td>0</td>
<td>1378.7</td>
</tr>
</tbody>
</table>

Table 2.6 Simulation results using the “recovered” E. coli A matrix and $n = m = 9$

Figure 2.11 shows the effects of the choice of analysis method on the natural logarithm of the ESEE. The only difference from the RRM experiment is that the forward network regression identification (FNIR) was competitive with the network identification regression (NIR) method for the case in which the experimental noise standard deviations ($\sigma_{ij}$) were small. We feel that the similarity of the results from the random
resilient matrices (RRM) simulations and the SOS pathway simulation provides some level of confirmation that both simulation approaches are relevant for evaluating alternative experimentation and analysis methods.

Figure 2.11 Interaction plot for SOS pathway and the natural logarithm of ESEE
7. **DISCUSSION AND FUTURE WORK**

The study of gene networks constitutes an application area of experimental design and analysis of great current interest to researchers in many fields. This article has proposed the application for network identification of only a few of the most well-known methods from the experimental design literature, i.e., D-optimal designs and least squares estimation in the forward network identification regression (FNIR) method. Our conclusion is that these straightforward alternatives are competitive with the more complicated methods from Gardner, Bernardo, Lorenz, and Collins (2003). Using the measures of coverage and false positives proposed by Gardner, Bernardo, Lorenz, and Collins (2003), FNIR was either dominant or competitive with their network identification regression (NIR) related to the 40 simulation conditions considered in our two computational experiments. Also, D-optimal designs generally improved the performance compared with the one-factor-at-a-time (OFAT) approach in Gardner, Bernardo, Lorenz, and Collins (2003).

Yet, many challenges remain in that all of the methods considered might seem viable only for preliminary investigations. This follows because of the potentially problematic coverage probabilities, false positive probabilities, and other measures of errors observed here and in Gardner, Bernardo, Lorenz, and Collins (2003). Also, our research has identified the sources of some of the errors including, most notably, numerical issues associated with the design matrix $U'U$. These errors were caused by
the lack of a constant term in the fitted model and diversity of the ranges used in experimentation. Further research on planning experiments including augmentation experiments and analysis methods to address these issues is potentially of great interest to biologists.

Specifically, little attention has been given to the subject of model misspecification. Genes not being considered could change and affect the results. In addition, systems might not reach steady state and non-linear phenomena not described by Equation 2.1 might be important. Both experimental design and analysis methods mitigating the effects of misspecification can be developed. For example, lack of fit analysis procedures that address the sparsity and multi-response nature of microarray data might aid in decisions about whether additional experimentation is needed. Also, Bayesian design and analysis procedures that include both primary and potential terms to mitigate the abovementioned possible biasing effects can be investigated analogous to those in DuMouchel and Jones (1994).

For analysis methods, replacing estimation using all genes in Equation 2.7 with stepwise regression could constitute a natural first step. This would permit comparisons using the less arbitrary definitions of zero regulation from Gardner, Bernardo, Lorenz, and Collins (2003) for the saturated $n = m$ case, because hypothesis testing would be possible. Furthermore, standardized regression might help eliminate numerical errors
while requiring constrained optimization to address the absence of a constant in Equation 2.1.

All of the simulation criteria in Section 5 could be used to generate supersaturated and other optimal experimental plans. To generate criteria that address model misspecification, assumption schemes need to be developed that include the possibility of bias analogous to those used in Allen and Berntshteyn (2003). Tegnér, Yeung, Hasty, and Collins (2003) explore one example of a nonlinear network to study their linear network based analysis. Their scheme could be generalized to address multiple sources of misspecification. Also, it may be of interest to develop criteria related to the accuracy of predictions from equations Equation 2.2 and Equation 2.3.

Given the inverse analysis aspect of identifying which genes are targeted by specific compounds, sequential design and analysis procedures seem to offer the greatest promise for efficiency and accuracy. Tegnér, Yeung, Hasty, and Collins (2003) proposed one such fully sequential scheme. The development and evaluation of sequential methods that address model misspecification has apparently received little or no attention in the existing literature.

ACKNOWLEDGMENTS

We thank Allen Miller for contributing significantly to the development of the random resilient matrices (RRM) simulation method. We thank James Collins and
Timothy Gardner for sharing their MATLAB® code with us. Diego Di Bernardo shared a relevant preprint with us. Ning Zheng and Joseph Fiksel provided many helpful discussions and Mikhail Bernshteyn generated the designs.

APPENDICES

In this appendix, the network identification regression (NIR) algorithm from Gardner, Bernardo, Lorenz, and Collins (2003) is reviewed and codified. The algorithm derives the recovered or estimated \( m \times m \) system matrix, \( \text{A}_{\text{est}} \), where \( m \) is the number of genes. The response data for the \( m \) gene concentrations from the \( n \) run experiment is given by the \( m \times n \) matrix, \( \mathbf{X} \). Different subsets, \( \tilde{\mathbf{X}}_i \), containing \( m^* \) columns of \( \mathbf{X} \) are used as specified by elements of the set of models considered \( S \).

NIR algorithm

Initialize \( \{ j = 1, \text{A}_{\text{est}} = 0 \} \)

If \( \{ \text{gene} j \text{ is not perturbed in any of the runs} \} \)

Then \( U_{kj} = \tilde{X}_{kj} \quad \forall \ k = 1,\ldots,n \) <Note: Values over-write original zeros in \( U \).>

Repeat \( \{ \text{Q} = \text{WW}' \text{ where W are optional weight parameters} \}

If \( \exists i \subset S \text{ such that condition number}(\tilde{X}_i'\text{Q}\tilde{X}_i) \leq 1000 \} \)

<At least one subset can be fitted.>
Then \( q = \text{Argmin} \left\{ u_j' \left[ I - \tilde{X}_j'Q\tilde{X}_j + rI \right]^{-1} \tilde{X}_j' \right\} u_j \) 

Subject to \( \{ \text{condition number} \left( \tilde{X}_j'Q\tilde{X}_j \right) \leq 1000 \) and \( \tilde{X}_j = \begin{pmatrix} \tilde{x}_{j,1} \\ \vdots \\ \tilde{x}_{j,m} \end{pmatrix} \) 

and \( r \) is optional ridge regression parameter. 

\[ a_{est,j} = -\left( \tilde{X}_j'Q\tilde{X}_j + rI \right)^{-1} \tilde{X}_j'Qu_j \] 

} <Otherwise the row of zeros is retained.> 

If \{gene \( j \) is not perturbed in any of the runs} 

Then \( \{ a_{est,j,l} = (a_{est,j,l} \div a_{est,l,j}) \) for \( l = 1,\ldots,m \} \)

\[ j = j + 1 \}

While \( \{ j \leq m \} \)

\[ A_{est} = \begin{pmatrix} a_{est,1}' \\ \vdots \\ a_{est,m}' \end{pmatrix} \]

Stop.
CHAPTER 3

SECOND STAGE DESIGN AND ANALYSIS FOR FORWARD-INVERSE ANALYSIS

ABSTRACT

This chapter proposes methods for forward and inverse system modeling using Bayesian and least squares regression. These methods are based on both space-filling design criteria for multiple response problems and linear optimality criteria focusing on D-optimality. Modeling with and without the constant term is considered motivated by the case study application of genetic network modeling. In this application, the regression does not include a constant term because of the assumptions of steady state systems theory. For real world application, we propose extended one-factor-at-a-time (EOFAT) experimentation followed by augmentation of next stage design which offers biologists simplicity. Results are illustrated with numerical examples, a test problem
from the literature, and a case study motivated by an on-going real world biological research related to genetic network modeling of the genetic effects associated with Down syndrome.

1. INTRODUCTION

It can be of interest to develop “inverse” models to predict unknown inputs ($x$) when the outputs have been measured ($y$). For example, Barton (2005 and 2006) and Barton, Meckesheimer, and Simpson (2001) argued that such models are helpful to support inverse design engineering. Inverse design engineering involves selecting inputs so that system performance best meets target values. The above mentioned articles detail the special role that inverse models can play in the context of quality function deployment (QFD).

Here, we focus on the motivation noted in Chapter 2 relating to biological systems and mode of action of compounds as described in Gardner et al. (2003). In these cases, it is desirable to predict the combination of genetic perturbations ($x$) that would give rise to the same measured gene concentrations ($y$) as when the target compound is inserted. Because developing empirical genetic models invariably involves inputting known perturbations and measuring gene concentrations, prediction of the genes “targeted” by compounds is inherently an inverse regression problem. Note that, in this chapter, we depart from the systems notation in the preceding chapter and adopt
regression notation similar to that used in Barton (2006). Therefore, here a design matrix involving inputs for \( n \) runs and \( k \) terms is \( X \). The corresponding matrix of outputs for \( n \) runs and \( r \) responses is \( Y \).

The methods proposed here are extensions of the methods proposed in previous research (in Barton, 2006 and Barton, Meckesheimer, and Simpson, 2001). The primary extension is to address the so-called “data starved” case in which fewer runs than there are terms in the fitted model are performed in the start-up phase. This is motivated by our on-going relationship with biologists relating to Down syndrome. Also, all related experimental planning problems are formulated more concisely here and rigorous results are produced that simplify their solution.

In Section 2, the overall forward-inverse sequential experimental planning problem is formulated. The methods proposed in the literature are described together with a more detailed description of the state-of-the-art heuristics that can be applied. Next, in Section 3 rigorous results are presented suggesting that multi-criteria D-optimality formulations can be solved at least approximately as single criteria formulation. Also, guidelines for collecting the first stage data are supported. Section 4 presents numerical test problems and the relevance of the derived theorems. Section 5 presents the findings from the numerical studies. In section 6, conclusions are summarized including recommendations for biologists.
2. FORWARD-INVERSE EMPIRICAL MODELING

In this section, the general forward-inverse empirical modeling (FIEM) problem is formulated and an associated FIEM modeling method is presented with explicit details omitted in Barton, Meckesheimer, and Simpson (2001). Also, the method presented here represents an extension to cases in which very little data is available at a given stage of experimentation.

Let the \( n \times k \) design matrix \( X \) denote a full design matrix associated with \( n \) runs and \( k \) model terms. This design matrix is “full” because it can be partitioned into rows corresponding to available historical data (in the \( n_1 \times k \) matrix, \( X_{old} \)) as well as planned experimental runs in the \( (n_2 \times k) \) matrix, \( X_{new} \). In this notation, it follows that \( n_1 + n_2 = n \). Similarly, define the \( n \times r \) matrix \( Y \) as the full response matrix containing the available historical response data (in the \( n_1 \times r \) matrix, \( Y_{old} \)) corresponding to the inputs \( (X_{old}) \). This matrix is augmented with predictions for the mean responses for runs that have not yet completed (in the \( n_2 \times r \) matrix, \( Y_{predicted} \)). Therefore, we have:

\[
X = \begin{bmatrix} X_{old} \\ X_{new} \end{bmatrix}, \quad Y = \begin{bmatrix} Y_{old} \\ Y_{predicted}(X_{new}) \end{bmatrix}.
\]

Further, let \( \varphi_1(X) \) refer to the forward design criterion or “information function” relevant to predicting new outputs as a function of new inputs. Similarly, \( \varphi_2(Y) \) is the information function relevant to predicting the inputs that resulted in measured outputs. This and previous research focuses on cases in which both forward and inverse
functions have the same form. Also, they correspond either to the D-optimality criterion, i.e., \( \varphi_1(Z) = \varphi_2(Z) \propto |Z'Z| \), or the maximum minimum distance criterion based on the L2 norm, i.e., \( \varphi_1(Z) = \varphi_2(Z) = \min_{i \neq j} \{(z_i - z_j)'(z_i - z_j)\} \).

Using this notation, the general forward-inverse empirical modeling (FIEM) formulation can be written as a multi-criteria problem as:

\[
\max_{X_{\text{new}}} \{\varphi_1(X), \varphi_2(Y)\} . 
\]

This formulation was originally explored (apparently) by Barton, Meckesheimer, and Simpson (2001). In their article, those authors addressed the multi-criteria aspect using the scaling and details we described below. With the scaling applied, their primary formulation involved weighting the forward objective by the parameter \( \lambda \) and the inverse objective by \( 1 - \lambda \) leading to:

\[
\max_{X_{\text{new}}} \{\lambda \varphi_1(X) + (1 - \lambda) \varphi_2(Y)\} . 
\]

For example, with an equal weighting of forward and inverse objectives, i.e., \( \lambda = 0.5 \) where \( \lambda \) is the weight giving to the two problems in the associated single criterion formulation.

In this dissertation, we will explore predictions based on the Bayesian regression formulation:

\[
Y_{\text{predicted}}(Z) = Z[(X'_{\text{old}}X_{\text{old}} + K)^{-1}(X'_{\text{old}}Y_{\text{old}} + KB)] = ZH
\]

51
where $\mathbf{B}$ is the prior mean for the coefficients which is generally taken to be the zero matrix, i.e., $\mathbf{B} = \mathbf{0}$. This follows in part because one is often working with centered and scaled units in which one has no prior reason to expect that specific coefficients are positive or negative. In addition, $\mathbf{H} = \mathbf{A}_{est}^{-1}$ in the notation from Chapter 2 for cases in which the number of responses $r$ equals the number of terms, $k$. Note that some results will generalize to problems in which other approaches were applied to estimate $\mathbf{H}$. Also, the above equation reduces to ordinary least squares regression, under the assumptions $\mathbf{K} = \mathbf{B} = \mathbf{0}$.

The multi-stage forward-inverse empirical modeling (MFIEM) method is:

**Step 1.** (Initialization) Plan the startup experimental design matrix, $n_1 \times k$ matrix, $\mathbf{X}_{old}$, satisfying:

$$\mathbf{X}_{old} = \underset{X_{i,j} \in [-1,1]}{\text{argmax}} \ q_1(\mathbf{X})$$

**Step 2.** (Initialization) Collect the data for all $r$ responses in the $n_1 \times r$ matrix, $\mathbf{Y}_1$, then shift and scale using the following convention with $\mathbf{Z} = \mathbf{Y}_1$ and $\mathbf{Y}_{old} = \mathbf{Y}_{coded}$:

$$\mathbf{Y}_{coded} = (\mathbf{Z} - \mathbf{M})\mathbf{S}^{-1}$$

where $\mathbf{M} = \frac{1}{2} \begin{bmatrix} \max_i Z_{i,1} + \min_i Z_{i,1} & \cdots & \max_i Z_{i,r} + \min_i Z_{i,r} \\ \vdots & \ddots & \vdots \\ \max_i Z_{i,1} + \min_i Z_{i,1} & \cdots & \max_i Z_{i,r} + \min_i Z_{i,r} \end{bmatrix}$

and $\mathbf{S} = \frac{1}{2} \begin{bmatrix} \max_i Z_{i,1} - \min_i Z_{i,1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max_i Z_{i,r} - \min_i Z_{i,r} \end{bmatrix}$.
Step 3. (Loop) Apply Bayesian regression using $Y_{old}$ and equation 3.3 to estimate $H$.

Step 4. (Loop) Plan the next stage, augmented experimental design matrix, $n_i \times k$ matrix, $X_i$, by solving the multi-criteria formulation in equation 3.2 with:

$$X_{old} = \begin{bmatrix} X_1 \\ \vdots \\ X_{i-1} \end{bmatrix} \text{ and } X_{new} = X_i.$$  \hspace{1cm} \text{Equation 3.6}

Step 5. (Loop) Collect the data for all $r$ responses in the $n_i \times r$ matrix, $Y_i$, and set:

$$Z = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \end{bmatrix}.$$  \hspace{1cm} \text{Equation 3.7}

Again, shift and scale using the convention in Equation 3.5 and $Y_{old} = Y_{coded}$.

Step 6. If $\phi_1(X) < \Delta_1$ and $\phi_2(Y_{old}) < \Delta_2$ stop. Otherwise, $i = i + 1$ and go to Step 3.

Barton, Meckesheimer, and Simpson (2001) introduced this method only considering the non-Bayesian, $K = 0$ and $B = 0$, estimation approach. They focused attention on two stage problems with sufficient initial data such that ordinary least squares can be applied, i.e., $n_1 \geq k$. The addition of Bayesian regression opens the possibility that, at any stage, $n_1 + n_2 + ... + n_i < k$ can be considered. We argue that this is important for the genetic network identification problem because of the expensive and time consuming nature of genetic perturbations. Also, in microarray experiments responses are measured in any run correspond to all possible genetic inputs regardless of which inputs (if any) are perturbed.
Consider also an approximate version of the FIEM formulation that is instructive because it clarifies issues relating to linear optimality criteria.

\[
\max_{x_{\text{new}}} \{ \varphi_1(X), \varphi_2(\bar{Y}) \}
\]

Equation 3.8

\[
\bar{Y} = \begin{bmatrix} Y_{\text{predicted}}(x_{\text{old}}) \\ Y_{\text{predicted}}(x_{\text{new}}) \end{bmatrix} = X[(X'_{\text{old}}X_{\text{old}} + K)^{-1}X'_{\text{old}}Y_{\text{old}}] = XH
\]

The above approximation is relevant particularly when the number of second stage runs is relatively large and/or the sizes of the residuals from the first stage modeling are relatively small.

3. RIGOROUS RESULTS FOR D-EFFICIENCY AND STARVED DATA CASES

In this section, rigorous results establish assumptions under which the multi-criteria FIEM formulation in Equation 3.2 can be solved as a single criterion formulation. Also, they establish conditions on the first stage experimental design (i.e., that all factors should be varied) such that the second stage solutions are potentially desirable. These latter results provide evidence that the commonly observed one-factor-at-a-time (OFAT) approach should be abandoned in data starved situations. This is important because we have observed that OFAT is commonly applied in the context of genetic network modeling (e.g., see Gardner et al., 2003, and Bansal, Gatta, and di Bernardo, 2006).
The first two theorems relate to the popular linear optimality D-optimality or D-efficiency criterion. As noted previously, they establish general conditions under which the multi-criteria FIEM formulation in Equation 3.2 can be solved as a single criterion formulation. In some cases of interest, the conditions are only approximately met. The theorems also provide a way to partially evaluate the accuracy of the approximations.

The first theorem gives results completely independent of the \( K \) and \( B \) values used in the Bayesian regression.

**Theorem 1.** (Zero Residuals Case) Assume \( \varphi_1(Z) = \varphi_2(Z) \propto |Z'Z| \) and \( \bar{Y} = XH \) for any nonsingular \( k \times k \) matrix \( H \), then if \( X_{\text{new}} \) is an optimal solution to:

\[
\max_{X_{\text{new}}} \{ \varphi_1(X) \}
\]

Then, it is also an optimal solution to:

\[
\max_{X_{\text{new}}} \{ \varphi_2(\bar{Y}) \}
\]

**Proof.** \( \varphi_2(\bar{Y}) \propto |\bar{Y}'\bar{Y}| = |H'X'XH| = |H'H||X'X| \). Therefore, \( \varphi_2(\bar{Y}) \propto |X'X| \propto \varphi_1(X) \). Consider two solutions, \( X_1 \) and \( X_2 \) and their associated \( \bar{Y}_1 \) and \( \bar{Y}_2 \). Assume without loss of generality \( \varphi_1(X_1) \leq \varphi_1(X_2) \). Then, \( \varphi_2(\bar{Y}_1) \propto \varphi_1(X_1) \) and \( \varphi_2(\bar{Y}_2) \propto \varphi_1(X_2) \) and, therefore, \( \varphi_2(\bar{Y}_1) \leq \varphi_2(\bar{Y}_2) \). Since this holds for any \( X_1 \) and \( X_2 \), it holds for an optimum \( X_1 \) and its associated \( \bar{Y}_1 \) must also be an optimum solution.

The popular D-optimality or D-efficiency criterion correspond to the assumption \( \varphi_1(Z) \propto |Z'Z| \). Consider cases in which D-optimality is considered relevant for both
forward and inverse regression. Further, assume that the residuals associated with $Y_{old}$ can be neglected. Then, Theorem 1 establishes that generating the D-optimal design for the forward problem is sufficient. Any optimal solution to the forward problem is a dominant solution to the multi-criteria forward-inverse problem (for any value of $\lambda$). This assumes that matrix $H$ is $k \times k$ and nonsingular. This assumption can be checked on a case by case matrix and applies to all of the numerical results relevant to genetic network modeling considered in this dissertation.

Also, note that the above only applies to D-optimality. Optimality for all other linear optimality criteria does depend on the specific scaling matrix $H$ as noted by Pukelsheim on p. 137. The second theorem focuses on the specific case of ordinary least squares regression, i.e., the prior covariance matrix satisfies $K = 0$.

**Theorem 2.** (Nonzero residuals case) Assume $\varphi_1(Z) = \varphi_2(Z) \propto |Z'Z|$, $K = 0$, and $\hat{e} = Y_{old} - X_{old}H$ for any nonsingular $k \times k$ matrix $H$, and $H'X'XH$ is full rank. Then, with the above definitions:

$$\varphi_2(Y) \propto |H'H||X'X| |I + [\hat{e}'0'] (H'X'XH)^{-1} [\hat{e}'0']|.$$  \hspace{1cm} \text{Equation 3.9}

**Proof.**

$$\hat{e} = Y_{old} - X_{old}(X_{old}'X_{old} + K)^{-1}X_{old}'Y_{old} = (I - X_{old}(X_{old}'X_{old} + K)^{-1}X_{old}')Y_{old}$$

$$Y = \begin{bmatrix} Y_{old} \\ Y_{predicted}(X_{new}) \end{bmatrix} = \bar{Y} + [\hat{e}] = XH + [\hat{e}]$$
\[ |Y'Y| = |H'X'XH + 2H'X' \begin{bmatrix} \hat{e} & 0' \end{bmatrix} \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix}| \]

Note that if \( K = 0 \) then \( \{(X'_{\text{old}}X_{\text{old}} + K)^{-1}\} = (X'_{\text{old}}X_{\text{old}} + K)^{-1} \), \( X_{\text{old}}(X'_{\text{old}}X_{\text{old}} + K)^{-1} \) is idempotent and

\[ H'X' \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix} = Y'_{\text{old}}X_{\text{old}}(X'_{\text{old}}X_{\text{old}} + K)^{-1}X'_{\text{old}} \left( 1 - X_{\text{old}}(X'_{\text{old}}X_{\text{old}} + K)^{-1}X'_{\text{old}} \right)Y_{\text{old}} = 0 \]

Therefore, for the \( K = 0 \) case, we have:

\[ |Y'Y| = |H'X'XH + [\hat{e}' \ 0']| \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix} = |H'X'XH| \left| 1 + \begin{bmatrix} \hat{e} \\ 0 \end{bmatrix} (H'X'XH)^{-1}[\hat{e}' \ 0'] \right| \]

where \( |A + x'y| = |A|(1 + yA^{-1}x') \) has been used. Substituting \( |H'X'XH| = |H'H||X'X| \) the result is proven.

Theorem 2 shows that the D-optimal forward-inverse multi-criteria problem can only approximately be solved using the forward solution. It also gives conditions under which this approximation holds, i.e., when the last term in the parenthesis in Equation 3.9 is approximately equal to 1 for all relevant designs.

Intuitively, holding any factor constant in all previous tests should result in degraded ability to make decisions about the merits of varying that factor in the future. The final rigorous result formalizes this intuition. It establishes that Bayesian regression and historical data provide little insight into the effects of factors held constant.

**Theorem 3.** (Irrelevance of constant factors) Assume \( K \) is diagonal, \( B = 0 \), and \( X_{\text{old},ij} = 0 \) for term \( i \) and all runs \( j = 1,...,n_1 \) then \( \varphi_2(Y) \) has no dependence on \( X_{\text{new},ij} \) for \( j = 1,...,n_2 \).
Proof. Without loss of generality, reform $X_{old}$ so that the zero columns are to the right, i.e., $X_{old} = [X_{nonzero} ~ 0]$. Then, the Bayesian formula for estimation becomes:

$$H = \left([X'_{nonzero}X_{nonzero} ~ 0] + K\right)^{-1} \left[X'_{nonzero} \ 0'\right] Y_{old}$$

$$= \left[(X'_{nonzero}X_{nonzero} + K_1)^{-1} \ 0 \ \ 0' \ \ K_2^{-1}\right] \left[X'_{nonzero} \ 0'\right] Y_{old}$$

$$= \left[(X'_{nonzero}X_{nonzero} + K_1)^{-1}X'_{nonzero} \ 0'\right] Y_{old} = [H_1 \ 0]$$

Therefore, consider two $X_{new}$ matrices which differ only for the variables associated with zero columns in $X_{old}$ : $[Z X_1]$ and $[Z X_2]$. The differences between the corresponding $Y$ are:

$$\begin{bmatrix} Y_{old} \\ [Z X_1][H_1] \end{bmatrix} - \begin{bmatrix} Y_{old} \\ [Z X_2][H_1] \end{bmatrix} = 0.$$ 

Since the $Y$s are the same, the $\phi_2(Y)$ values are the same also and the theorem is proven.

The importance of Theorem 3 is that failing to vary inputs in previous experiments is sufficient such that one has little or no guidance about the benefits of varying them in the future. In the context of genetics, this might appear counterintuitive because one might hope that the past recorded responses associated with the inputs could provide guidance relating to the value of varying the inputs in the future. The theorem goes part of the way to establishing that, in the context of linear models, such
hopes are in vain. While the linear model estimation structure has been applied, the result is general for all possible criteria $\varphi_2(Y)$.

Further, it should be noted that all three theorems hold even for second and higher order models. The relationship of all three results and numerical cases of interest is described in the context of numerical results in the next section.

4. NUMERICAL EXAMPLES

This section describes numerical examples used to illustrate the application of the FIEM method from Section 3. The results also facilitate discussion of the theorems from Section 2. In addition, the section describes state-of-the-art algorithms for solving the related optimization problems. For the D-optimality related objectives, solution methods were based on the Johnson and Nachtsheim (1983) algorithm implemented using SAS®. For the space filling objectives, methods were based on the formulation in Stinstra et al. (2003) and methods from Drezner (1995) implemented using AMPL.

First, results are presented based on the assumption that both forward and inverse criteria correspond to D-optimality or percentage D-efficiency. Using the Mitchel (1974) standard assumption, this implies:

$$\varphi_1(Z) = \varphi_2(Z) = \frac{100}{n} |Z'Z|^{1/k}$$

Equation 3.10
Results are presented for three sets of test functions based on linear models which are polynomials.

Table 3.1 shows the coefficients for three sets of test functions. The first two sets involve only first order models. The third involves second order models.

Table 3.2 repeats the coefficients for the SOS network described in Chapter 2 and Gardner et al. (2003). This shows how, in genetic network examples, each input factor is a gene which corresponds to a response. Changing the input factors involves changing how much of the related gene is added to the system (at a constant rate using plasmids). The responses record the changes in the concentrations for those same genes (in the associated mRNA).

<table>
<thead>
<tr>
<th>Term</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5 5</td>
<td>5 5</td>
<td>5 21</td>
</tr>
<tr>
<td>x1</td>
<td>25 1</td>
<td>-2 1</td>
<td>3 -9</td>
</tr>
<tr>
<td>x2</td>
<td>5 3</td>
<td>5 3</td>
<td>9 -2</td>
</tr>
<tr>
<td>x1^2</td>
<td>0 0</td>
<td>0 0</td>
<td>1 5</td>
</tr>
<tr>
<td>x2^2</td>
<td>0 0</td>
<td>0 0</td>
<td>5 2</td>
</tr>
<tr>
<td>x1x2</td>
<td>0 0</td>
<td>0 0</td>
<td>4 -3</td>
</tr>
</tbody>
</table>

Table 3.1 Three sets of test function coefficients for numerical examples
<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( y_4 )</th>
<th>( y_5 )</th>
<th>( y_6 )</th>
<th>( y_7 )</th>
<th>( y_8 )</th>
<th>( y_9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.60</td>
<td>-0.38</td>
<td>0.00</td>
<td>0.14</td>
<td>-0.22</td>
<td>-0.06</td>
<td>0.17</td>
<td>-0.31</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>0.18</td>
<td>-0.57</td>
<td>0.09</td>
<td>-0.14</td>
<td>0.04</td>
<td>0.22</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>-0.79</td>
<td>0.02</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.86</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>-0.07</td>
<td>-0.04</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.49</td>
<td>-0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td>-0.02</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.07</td>
<td>0.90</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.66</td>
<td>0.00</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.03</td>
<td>-2.07</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.22</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2 The SOS network sets of function coefficients for numerical examples

Figure 3.1 shows the experimental region used in our examples for experimental planning of input test runs. The points must be taken from inside the square. It also shows the corresponding output region associated with the first set of test functions and assuming responses are scaled so that they range between -1 and 1 in all dimensions. The shape is a parallelogram because all of the models in the text set are first order. By selecting sites in the x-space, measurements are made in the y-space. If numerical errors can be neglected, one can imagine testing the corner points in the x-space and measuring responses at each vertex of the parallelogram. Figure 3.2 shows the two regions for the third test set involving second order functions. In this case, the corners of the x-space region are not associated with all of the vertices in the y-space. Therefore, spreading runs in the x-space is not necessarily the optimal strategy for spreading them in the y-space.
As an additional consideration, we considered three types of startup experimental plans. For the generic experimentation examples, we considered only putative D-optimal startup designs generated using the exchange algorithm heuristic of Johnson and Nachtsheim (1983) using 200 starting designs. For the case of the SOS test network functions, we considered one-factor-at-a-time (OFAT) experimentation as
described in Chapter 2. We also considered a simple variant of OFAT in which the genes that would otherwise be held constant are varied simultaneously with arbitrarily selected other genes. Therefore, the extended OFAT (EOFAT) approach involves planning the first $m$ perturbations for the first $n_1$ runs by perturbing only the first $n_1$ genes. Then, the perturbations for the remaining $m - n_1$ genes are perturbed in the first $m - n_1$ runs. Therefore, the $i^{th}$ gene is perturbed simultaneously with the $(n_1 + i)$ gene.

5. COMPUTATIONAL RESULTS

Table 3.3 shows the results for applying the FIEM method in Section 3 to generate the second stage experimental plans. The test functions were used to generate first stage results assuming two sets of pseudo random errors. The random errors were assumed to be independent, identically distributed (IID) according to a normal distribution with mean zero and variance 0.4. In our implementation, we simply planned the first and second stage experimental design ignoring the inverse formulation using the SAS® optex function based on the exchange algorithm heuristic of Johnson and Nachtsheim (1983) using 200 starting designs. Next, in selected cases, we attempted to improve these designs by focusing solely on the inverse problem using the conjugate gradient downhill search for the second stage. We were not able to find any improved solutions. In other words, the residuals from the first stage were apparently negligible in
the context of Theorem 2. More systematic checking based on the genetic algorithm built into the SAS® software is suggested for future study.

Therefore, our first finding is that the computational results support the implication of Theorem 2. To a good approximation it appears that, in the cases we considered, the forward-inverse D-optimality problem can be solved as the forward only problem. A second finding is also supported by the results. While the experimental plan at any stage is apparently unaffected by the test set, the stopping criteria are affected. For example, stopping after the second stage in the context of test set 2 is more reasonable than for the other cases. This follows because the inverse D-efficiency of 25.9 in this case might be considered adequate. Since this value is a known prior to second stage testing, it could function as a sample size evaluation technique. For example, in the context of test sets 1 and 3, the experimenter might choose to perform \( n_2 > 5 \) runs while setting on \( n_2 = 5 \) runs in the context of test set 2.

Our third finding is that it is the EOFAT procedure offers potentially important benefits compared with the OFAT procedure. Note that EOFAT like OFAT is also reasonably simple and meets practical constraints that few genes are perturbed at a time. We suggest further study of the EOFAT procedure is warranted. The importance of the EOFAT in the context of other criteria is also supported by the results of Theorem 3. By perturbing all the genes in the first set of runs, the chance to improve virtually any forward-inverse design criterion is likely enhanced.
Figure 3.3 shows the AMPL code used to generate space filling designs. The formulation of Stinstra et al. (2003) was used. The formula was adapted to solve the multi-objective and multi-response optimization problems. This approach effectively replicates the heuristic proposed by Drezner and Erkut (1995) which is apparently regarded as state-of-the-art.

Table 3.4 shows the numerical results based on the assumption that both forward and inverse criteria correspond to maximum-minimum (maximin or Mm) L2 distance criteria. For certain combinations of the test function and the numbers of runs, the forward-inverse formulation details greatly affect performance. For example, for the \( n_1 = n_2 = 4 \) case and the first set of test functions (F1), the affect of changing the formulation from \( \lambda = 1 \ (x) \) to \( \lambda = 0.5 \ (x + y) \) to \( \lambda = 0 \ (y) \) increased the scaled maximin y distance from 0.27 to 0.49. This occurs simultaneous with a large drop in the maximin x distance from 1.07 to 0.41.
| Startup    | Test | $m$ | $r$ | $n_1$ | $n_2$ | Form.    | $100 \frac{1}{n} |X'X|^{1/p}$ | $100 \frac{1}{n} |Y'Y|^{1/p}$ |
|------------|------|-----|-----|-------|-------|----------|-----------------|-----------------|
| D-optimal  | F1   | 2   | 2   | 10    | 5     | $x$      | 99.5            | 12.9            |
| D-optimal  | F1   | 2   | 2   | 10    | 5     | $x+y$    | 99.5            | 12.9            |
| D-optimal  | F2   | 2   | 2   | 10    | 5     | $y$      | 99.5            | 25.9            |
| D-optimal  | F2   | 2   | 2   | 10    | 5     | $x+y$    | 99.5            | 25.9            |
| D-optimal  | F3   | 2   | 2   | 10    | 5     | $x$      | 99.5            | 19.8            |
| D-optimal  | F3   | 2   | 2   | 10    | 5     | $x+y$    | 99.5            | 19.8            |
| D-optimal  | F1   | 2   | 2   | 4     | 4     | $x$      | 100.0           | 31.9            |
| D-optimal  | F1   | 2   | 2   | 4     | 4     | $x+y$    | 100.0           | 31.9            |
| D-optimal  | F2   | 2   | 2   | 4     | 4     | $x$      | 100.0           | 63.9            |
| D-optimal  | F2   | 2   | 2   | 4     | 4     | $x+y$    | 100.0           | 63.9            |
| D-optimal  | F2   | 2   | 2   | 4     | 4     | $y$      | 100.0           | 63.9            |
| OFAT       | SOS  | 9   | 9   | 5     | 4     | $x$      | 73.3            | 49.5            |
| OFAT       | SOS  | 9   | 9   | 5     | 4     | $x+y$    | 73.3            | 49.5            |
| OFAT       | SOS  | 9   | 9   | 5     | 4     | $y$      | 73.3            | 49.5            |
| EOFAT      | SOS  | 9   | 9   | 5     | 4     | $x$      | 93.2            | 59.3            |
| EOFAT      | SOS  | 9   | 9   | 5     | 4     | $x+y$    | 93.2            | 59.3            |
| EOFAT      | SOS  | 9   | 9   | 5     | 4     | $y$      | 93.2            | 59.3            |

**Table 3.3** The forward-inverse D-optimal design efficiencies for the test functions
# maximin design augmenting previous runs
set TERMS;
set TOTRUNS;
set RESP_TYPE;
set OLDRUNS;
set NEWRUNS;
set PAIRS := {i in TOTRUNS, j in TOTRUNS : i < j};
var x {i in TOTRUNS, v in TERMS} := Uniform(-1,1);
var y {i in TOTRUNS, m in RESP_TYPE} := Uniform(-1,1);
var dmin_x;
var dmin_y;
param alpha;
param beta {TERMS, RESP_TYPE};
param oldx {OLDRUNS, TERMS};
param oldy {OLDRUNS, RESP_TYPE};
maximize min_dist: alpha*dmin_x + (1-alpha)*dmin_y;
subject to
distance_x {(i, j) in PAIRS}: sum {v in TERMS} (x[i,v] - x[j,v])^2 >= dmin_x;
range_x {i in TOTRUNS, v in TERMS}: -1 <= x[i,v] <= 1;
# forcing the first column of x to be 1
constant_x {i in TOTRUNS}: x[i,1] = 1;
# oldruns of x cannot be changed
old_x {i in OLDRUNS, v in TERMS}: x[i,v] - oldx[i,v] = 0;
distance_y {(i, j) in PAIRS}: sum {m in RESP_TYPE} (y[i,m] - y[j,m])^2 >= dmin_y;
# multiplication of y=x*beta21
x_beta {i in NEWRUNS, m in RESP_TYPE}: sum {v in TERMS} x[i,v]*beta_est[v,m] = y[i,m];
# oldruns of y cannot be changed
old_y {i in OLDRUNS, m in RESP_TYPE}: y[i,m] - oldy[i,m] = 0;
# distance cannot be negative
pos_dminx {i in TOTRUNS, v in TERMS}: dmin_x >= 0;
pos_dminy {i in TOTRUNS, m in RESP_TYPE}: dmin_y >= 0;

Figure 3.3 AMPL code to implement Drezner and Erkut (1995) Maximin method

Another finding is that, for some test functions, the multi-criteria approach used has minimal affect on the resulting performance. For example, for the second set of test
functions and for both combinations of runs and factors considered, the effect of the choice of \( \lambda \) is negligible. This occurs because the test function is such that points in the y-space are already close together after the startup design is completed. Then, there is no opportunity to affect performance, i.e., the additional runs cause a negligible drop in maximin y-space distances.

Our last finding from the numerical examples considered relates to the EOFAT procedure in the context of the SOS pathway example. Again, the EOFAT procedure offers performance advantages compared to OFAT while maintaining reasonable complexity and cost for biologists. Since the FIEM method based on an OFAT startup has no knowledge relating to the untested genes after startup, even attempting to improve the maximin distance in the y-space actually decreases the distance. In light of the result in Theorem 3 and these computational results, we again suggest that further study of the EOFAT procedure for startup testing be considered.
<table>
<thead>
<tr>
<th>Startup</th>
<th>Set</th>
<th>m</th>
<th>r</th>
<th>(n_1)</th>
<th>(n_2)</th>
<th>Form.</th>
<th>Rep.1 Min.x</th>
<th>Rep.1 Min.y</th>
<th>Rep.2 Min.x</th>
<th>Rep.2 Min.y</th>
<th>Avg. Min.x</th>
<th>Avg. Min.y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mm F1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>x</td>
<td>0.30 0.09</td>
<td>0.30 0.09</td>
<td>0.30 0.09</td>
<td>0.30 0.09</td>
<td>0.30 0.09</td>
<td></td>
</tr>
<tr>
<td>Mm F1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>x+y</td>
<td>0.25 0.13</td>
<td>0.25 0.13</td>
<td>0.25 0.13</td>
<td>0.25 0.13</td>
<td>0.25 0.13</td>
<td></td>
</tr>
<tr>
<td>Mm F1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>y</td>
<td>0.14 0.17</td>
<td>0.17 0.17</td>
<td>0.17 0.17</td>
<td>0.16 0.17</td>
<td>0.17 0.17</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>x</td>
<td>0.30 0.00</td>
<td>0.30 0.01</td>
<td>0.30 0.01</td>
<td>0.30 0.01</td>
<td>0.30 0.01</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>x+y</td>
<td>0.30 0.01</td>
<td>0.21 0.01</td>
<td>0.21 0.01</td>
<td>0.25 0.01</td>
<td>0.25 0.01</td>
<td></td>
</tr>
<tr>
<td>Mm F1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>x</td>
<td>1.07 0.27</td>
<td>1.07 0.27</td>
<td>1.07 0.27</td>
<td>1.07 0.27</td>
<td>1.07 0.27</td>
<td></td>
</tr>
<tr>
<td>Mm F1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>x+y</td>
<td>1.00 0.35</td>
<td>1.00 0.37</td>
<td>1.00 0.37</td>
<td>1.00 0.36</td>
<td>1.00 0.36</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>y</td>
<td>0.33 0.49</td>
<td>0.49 0.50</td>
<td>0.49 0.50</td>
<td>0.41 0.49</td>
<td>0.49 0.49</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>x</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>x+y</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td>1.07 0.05</td>
<td></td>
</tr>
<tr>
<td>Mm F2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>y</td>
<td>0.38 0.08</td>
<td>0.38 0.08</td>
<td>0.38 0.08</td>
<td>0.38 0.08</td>
<td>0.38 0.08</td>
<td></td>
</tr>
<tr>
<td>OFAT SOS</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>x</td>
<td>8.00 4.71</td>
<td>8.00 4.85</td>
<td>8.00 4.85</td>
<td>8.00 4.78</td>
<td>8.00 4.78</td>
<td></td>
</tr>
<tr>
<td>OFAT SOS</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>x+y</td>
<td>8.00 4.86</td>
<td>8.00 4.95</td>
<td>8.00 4.95</td>
<td>8.00 4.90</td>
<td>8.00 4.90</td>
<td></td>
</tr>
<tr>
<td>OFAT SOS</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>y</td>
<td>8.00 4.71</td>
<td>8.00 4.07</td>
<td>8.00 4.07</td>
<td>8.00 4.39</td>
<td>8.00 4.39</td>
<td></td>
</tr>
<tr>
<td>EOFAT SOS</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>x</td>
<td>8.00 4.21</td>
<td>8.00 4.03</td>
<td>8.00 4.03</td>
<td>8.00 4.12</td>
<td>8.00 4.12</td>
<td></td>
</tr>
<tr>
<td>EOFAT SOS</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>x+y</td>
<td>8.00 5.43</td>
<td>8.00 6.35</td>
<td>8.00 6.35</td>
<td>8.00 5.90</td>
<td>8.00 5.90</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4 The forward-inverse Maximin design efficiencies for the test functions

6. CONCLUSIONS

Our results establish that for relevant cases, the multi-criteria FIEM formulation can be solved to global optimality as a single forward design problem. These conditions include that a testable bound on the residuals from the first stage regression is satisfied. Also, we have established that if startup experimentation does not vary all of the factors, then, the efficacy of the FIEM formulation and method are limited. To address this in the biological systems context, we proposed the extended one-factor-at-a-time
(EOFAT) procedure. This procedure can be extended to address a follow-up or multi-stage planning by augmenting next stage design. For the biologists, the second stage formulation could be solved with additional constraints on the number of genes in each run. In our numerical studies, we verified that the EOFAT procedure offers performance advantages in the context of both D-optimality and space filling criteria compared with OFAT. Also, we verified that forward D-optimal designs are approximately forward-inverse D-optimal designs for the cases considered. Finally, we showed how the forward-inverse formulation can aid in sample size selection in the context of D-optimality.
CHAPTER 4

CONCLUSIONS AND FUTURE RESEARCH

This dissertation has proposed methods relevant to potentially all experimentation for which both forward and inverse prediction capabilities are needed. Methods have been proposed for single stage and multi-stage experimentation. Several methods have been focused on cases in which both linear models are used because of the relevance of steady state conditions and the scarcity of available data. When linear methods have been incorporated attention has further focused on cases involving no constant term motivated by the equilibrium conditions.

The following summarizes the primary findings related to the four major problem statements:

1. Relevant to first stage design and analysis, D-optimal experimental plans and ordinary least squares regression can offer important prediction advantages compared with a relevant class of alternatives from the literature (Gardner et al., 2003). These time-tested applied statistics methods can foster improved accuracy
both in the context of numerical examples from the literature and for our own proposed simulation-based evaluation methods. The proposed simulation methods are themselves relevant for evaluating any empirical model building method for estimating system resilience and other properties of the system matrix.

2. Related to test beds, random resilient matrices (RRM) are important because they are built on the plausible premise that biological systems are stable, i.e., system resilience is positive. Further, the proposed test bed in Chapter 2 offers perhaps the only way to combine plausible assumptions both about equilibrium conditions and stability. Proposed metrics summarize estimation errors and include a resilience diagnostic that can be used to assess fitted models in real cases.

3. Relevant to second stage or sequential design and analysis, two types of methods have been proposed that extend results in Barton (2005). The extensions involve addressing starved data situations and complications relating to the potential absence of constant terms. The first type includes experimental augmentation plans designed to facilitate the application of many types of empirical modeling methods. These approaches are based on two criteria maximum minimum distance formulations for both the input and predicted output spaces. Second, additional methods are also proposed based on D-optimality. While these methods have primary relevance only to linear modeling approaches, they can be
critical for addressing numerical errors that can easily make model predictions inaccurate.

4. Rigorous results are derived proving that, under certain assumptions, it is not necessary to derive D-optimal designs specifically for two criteria forward-inverse problems. D-optimal designs for the forward problem are provably optimal for the two criteria forward-inverse problem. Therefore, established augmentation heuristics can be applied. Still, as noted in Chapter 2, the optimal designs for cases in which there are no constant terms must be generated in engineering units.

5. The extended one-factor-at-a-time (EOFAT) is proposed in Chapter 3 specifically to support the Telethon Institute of Genetics and Medicine (TIGEM) project for studying the genetic network associated with Down syndrome. The methods were adapted to sequence the next set of 20 genes for testing such that the combined dataset involving 130 perturbations and 200 genes can be modeled accurately. Bayesian regression was used for prediction and a multiple choice constraint was added to the formulation. In that way, the biologists can use their preferred one-factor-at-a-time approach and approximate results will be available as soon as possible.

In general, the methods proposed and the knowledge gained from this dissertation can be adopted by the biologists. The extended one-factor-at-a-time
(EOFAT) experimentation can be used to decide which genes the biologists should perturb during the initial stage. The EOFAT design offers simplicity desired by the biologists and also as we showed in chapter 3, this design offers performance advantages over one-factor-at-a-time (OFAT) experimentation. The responses gathered from EOFAT experiments can be analyzed using FNIR to infer the underlying genetic networks as discussed in chapter 2. After the initial stage, a multi-stage design can be generated by augmenting the EOFAT used. Again the responses gathered can be used to improve the prediction of the genetic networks from previous stage. The formulation presented in chapter 3 can be used to design a multi-criteria experimental planning. To accommodate the preference of the biologists, Equation 3.2 can be solved with additional constraints on the number of genes perturbation in each experimental run.

Our research has also left important topics unaddressed including those mentioned in chapters 2 and 3. First, we suggest that the application of constrained regression to address the numerical errors encountered in Chapter 2 be investigated. Also, criteria with less model dependence than D-optimality can be investigated rigorously for forward-inverse problem. These can include Bayesian D-optimality (DuMouchell and Jones, 1994), the expected integrated mean squared error (Allen, Yu, and Schmitz, 2003) and generalized A-optimality (Ferhatosmanoglu et al., under review). Third, nonlinear modeling and test methods can be considered to address conditions when systems are no longer “close” to steady state conditions. Fourth, sequential
forward-inverse methods can be evaluated using prediction accuracy related metrics. This would combine the methods from Chapter 3 with the test methods from Chapter 2. Fifth, weighted least squares and data transformations can be investigated further to address deviation from normality assumption of the random error due to the nature of microarray experiments and human error.

With the completion of human genome project in 2003, a new chapter in the study of genetics has moved to primary importance. Now, researchers are trying to identify the relationship between genes in a network. Specifically how the genes interact with each other. Whether the presence of a particular gene inhibits or activates the existence of other genes in the network. It is our hope that the research in this dissertation will accelerate the pace of genetic network modeling and hasten the development of gene therapies and related pharmaceutical compounds.
BIBLIOGRAPHY


