FROM DATA TO INTERVENTIONS: USING SYSTEM IDENTIFICATION AND ROBUST CONTROL ALGORITHMS TO DESIGN EFFECTIVE TREATMENTS

A Dissertation in
Electrical Engineering
by
Korkut Bekiroglu

© 2015 Korkut Bekiroglu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2015
The dissertation of Korkut Bekiroglu was reviewed and approved* by the following:

Constantino Lagoa  
Professor of Electrical Engineering  
Dissertation Advisor, Chair of Committee

Minghui Zhu  
Assistant Professor of Electrical Engineering

Kenneth Jenkins  
Professor of Electrical Engineering

Stephanie T. Lanza  
Research Associate Professor of Health and Human Development

Necdet Serhat Aybat  
Assistant Professor of Industrial Engineering

Kultegin Aydin  
Professor and Department  
Head of Electrical Engineering

*Signatures are on file in the Graduate School.
Abstract

Behavioral and social scientists have demonstrated the advantages of the adaptive treatments, which usually provide better results than the fixed treatment (all patients get same type and level of treatment). Therefore, in this dissertation, we initially illustrate how one can develop dynamical models with suitable uncertainties for behaviors and identify them. Then we also demonstrate the usage of control engineering methods, such as feedback or adaptation, and robust optimization, to develop a systematic way for designing robust personalized treatment algorithms.

This new robust adaptive treatment design consists of three steps. In this dissertation, we develop different algorithms for first and second steps. For the first step, three different identification algorithms (identification with Lasso, Parsimonious model identification of single input single output systems, and Parsimonious identification of multi input multi output systems) are proposed which can utilize intensive longitudinal behavioral data to identify model parameters, interpolate the missing data, and quantify the uncertainties in the model. Then, for the second step, we provide a detailed step-by-step explanation of how control engineering methods can be used to design a robust adaptive intensive intervention. Finally, the methods are evaluated via simulation.

The performance of identification algorithms is demonstrated with synthetic behavioral data. The simulation results illustrate how the designed robust adaptive intensive intervention can produce improved outcomes with less treatment by providing treatment only when it is needed. The methods are robust to model uncertainties as well as to the influence of unobserved causes. As a result, these new methods can be used to design robust adaptive interventions that function effectively yet reduce participant burden.
Table of Contents

List of Figures viii
List of Tables x
List of Symbols xi
List of Abbreviation xv
Acknowledgments xvii

Chapter 1
Introduction 1
1.1 Model and Identification . . . . . . . . . . . . . . . . . . . . . . . 3
1.2 Robust Adaptive Treatment . . . . . . . . . . . . . . . . . . . . 5
1.3 Contributions and Outline . . . . . . . . . . . . . . . . . . . . . 7

Chapter 2
Background-Atomic Norm, System Identification, and Model Predictive Control 10
2.1 Definitions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
2.1.1 Basic Definitions . . . . . . . . . . . . . . . . . . . . . . . 11
2.1.2 Atomic Norm . . . . . . . . . . . . . . . . . . . . . . . . . 12
2.2 System Identification . . . . . . . . . . . . . . . . . . . . . . . . 14
2.3 Model Predictive Controller . . . . . . . . . . . . . . . . . . . . 15

Chapter 3
Identification Using Lasso 18
3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Model</td>
<td>19</td>
</tr>
<tr>
<td>3.3</td>
<td>Identification of Model and Perturbations</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Formulation of Identification Problem</td>
<td>22</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Lasso</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Identification Results</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusion</td>
<td>29</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Parsimonious Model Identification of SISO Systems from Multi-Run</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>31</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Literature Review</td>
<td>33</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Dynamical Model Identification of Behaviors</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Problem Statement</td>
<td>35</td>
</tr>
<tr>
<td>4.3</td>
<td>Previous Results</td>
<td>36</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Set of Atoms</td>
<td>36</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Deciding on the Scaling Factor $\gamma_p$ [1]</td>
<td>37</td>
</tr>
<tr>
<td>4.4</td>
<td>Set of Atoms for Multi-Runs</td>
<td>38</td>
</tr>
<tr>
<td>4.5</td>
<td>System Identification Via Atomic Norm Minimization</td>
<td>40</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Multi-Runs</td>
<td>41</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Nonuniform Sampling/Incomplete Data</td>
<td>43</td>
</tr>
<tr>
<td>4.6</td>
<td>Identification Algorithm</td>
<td>45</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Greedy Algorithm</td>
<td>45</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Modification of Randomized Algorithm in [2]</td>
<td>46</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Implementation-Heuristic</td>
<td>48</td>
</tr>
<tr>
<td>4.7</td>
<td>Simulation Results</td>
<td>50</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Experiment of Single Run</td>
<td>50</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Experiment-1 Multi Run</td>
<td>53</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Experiment-2 Multi Run</td>
<td>54</td>
</tr>
<tr>
<td>4.8</td>
<td>Concluding Remarks</td>
<td>58</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Parsimonious Model Identification of MIMO Systems</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction and Motivation</td>
<td>61</td>
</tr>
<tr>
<td>5.2</td>
<td>Preliminaries</td>
<td>61</td>
</tr>
<tr>
<td>5.3</td>
<td>Problem Statement</td>
<td>63</td>
</tr>
<tr>
<td>5.4</td>
<td>Convex Relaxation Using Atomic Norm</td>
<td>66</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Atomic Norm</td>
<td>66</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Convex Relaxation</td>
<td>67</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Extension to ID Problem with Missing Points</td>
<td>68</td>
</tr>
</tbody>
</table>
Appendix B
Technical Details and Proofs of Theorems in Chapter 6  122
B.1 Technical Details Concerning the Objective Function and MPC  .  122
B.2 Simulated Data for 150 Participants . . . . . . . . . . . . . . . . . 124
  B.2.1 Logistic Function . . . . . . . . . . . . . . . . . . . . . . 125
B.3 Theorem in [3, 4, 5] . . . . . . . . . . . . . . . . . . . . . . . . . 126
B.4 Proof of Theorem 14 . . . . . . . . . . . . . . . . . . . . . . . . . 126
B.5 Proof of Theorem 15 . . . . . . . . . . . . . . . . . . . . . . . . . 129

Bibliography  131
# List of Figures

1.1 Robust Adaptive Treatment ........................................... 6  
1.2 Implementation of Algorithm ........................................ 7  
2.1 Set of Standard Basis Vectors and Extremes ....................... 13  
2.2 System Identification Setup ........................................ 14  
2.3 Basic Description of MPC and Horizon Strategy: \( u(t) \) is the input applied to the system ........................................ 17  
4.1 Real and Estimated Single Run Output (SISO) .................... 51  
4.2 Impulse Response of Single Run (SISO) ............................ 52  
4.3 Initial Condition Response of Single Run (SISO) ................ 52  
4.5 Impulse Responses of Multi Run Experiment-1 (SISO) ........... 54  
4.6 Initial Condition response of Each Run Experiment-1 (SISO) ... 55  
4.7 Output of Each Run Experiment-2 (SISO) .......................... 56  
4.8 Impulse Responses of Multi Run Experiment-2 (SISO) .......... 57  
4.9 Initial Condition response of Each Run Experiment-2 (SISO) ... 57  
5.1 Real and Estimated Outputs Experiment-1 (MIMO) .............. 73  
5.2 Impulse Responses Experiment-1 (MIMO) .......................... 74  
5.3 Real and Estimated Initial Condition Responses Experiment-1 (MIMO) 75  
5.4 Real and Estimated Outputs Experiment-2 (MIMO) .............. 76  
5.5 Impulse Responses Experiment-2 (MIMO) .......................... 77  
5.6 Initial Condition Responses Experiment-2 (MIMO) .............. 77  
5.7 Multi-run Outputs of 4-Runs (MIMO) ............................... 79  
5.8 Initial Condition Response of 4-Runs (MIMO) ................... 80  
5.9 Impulse Response of Each Channel Multi-Run (MIMO) .......... 81  
6.1 Adaptive Intervention Algorithm .................................... 93  
6.2 Smoking Urge under Interventions .................................. 103  
6.3 Difference in Smoking Urge under the Adaptive Intensive Intervention and Full Treatment for One Hypothetical Individual 104
6.4 Difference in Mean Smoking Urge as Compared to No Treatment for All Individuals in Population A ........................................ 107
6.5 Difference in Mean Smoking Urge as Compared to No Treatment for Smokers with Higher than Average Smoking Urge in Population A ................................................................. 107
6.6 Smoking Urge under Adaptive Intervention ................................... 111
6.7 Sparse Disturbance ......................................................................... 112
6.8 Performance of Adaptive Intensive Intervention .............................. 113
6.9 Adaptive and Full Treatment Intervention Results ............................ 114

B.1 The Logistic Function ................................................................. 125
List of Tables

4.1 Length of Output and Initial Conditions Experiment-1 (SISO) . . 53
4.2 Length of Output and Initial Conditions . . . . . . . . . . . . . . 56

5.1 Poles and $\|\delta\|_1$ at Each Iteration . . . . . . . . . . . . . . 74
5.2 Poles and $\|\delta\|_1$ at Each Iteration . . . . . . . . . . . . . . 78
5.3 Singular Values of $c_p$ in Multi-Run . . . . . . . . . . . . . . . 79
5.4 True and Estimated Systems - Multi-Run . . . . . . . . . . . . . 82

6.1 A Partial Decision Rule Table . . . . . . . . . . . . . . . . . . . . 100
6.2 Descriptive Statistics for the 500 Evaluations . . . . . . . . . . . 108
6.3 Initial Condition of States and Input . . . . . . . . . . . . . . . 109
6.4 Intervention Performance of SDP . . . . . . . . . . . . . . . . . . . 110
List of Symbols

Sets
- $\mathbb{R}$ Set of real numbers
- $\mathbb{Z}$ Set of integers
- $\mathbb{N}$ Set of natural numbers
- $\mathbb{C}$ Set of complex numbers
- $\mathbb{D}_\rho$ Closed, origin centered disk in $\mathbb{C}$, with radius $\rho$
- $\text{conv}(A)$ Convex hull of the set $A$

Variables
- $\mathbf{g}$ a vector in $\mathbb{R}^N$
- $\mathbf{A}$ a matrix in $\mathbb{R}^{M\times N}$
- $\mathbf{A}$ a matrix in $\mathbb{C}^{M\times N}$
- $\mathbf{I}$ identity matrix of appropriate dimension
- $\mathbf{1}$ is a vector consisting of all ones in $\mathbb{R}^m$
- $\Re(p)$ Real part of $p \in \mathbb{C}$
- $\Im(p)$ Imaginary part of $p \in \mathbb{C}$
- $p^*$ Complex conjugate of $p \in \mathbb{C}$
Vector Operations

\( \mathbf{g}^T \) Transpose of a vector \( \mathbf{g} \)

\( \mathbf{g}_i \) \( i^{th} \) element of vector \( \mathbf{g} \)

\( \mathbf{g}_{(i:I)} \) extract the \( i^{th} \) through the \( I^{th} \) elements of vector \( \mathbf{g} \)

\( \text{card}(\mathbf{g}) \) Cardinality (number of the non-zeros) of vector \( \mathbf{g} \)

\( \text{dim}(\mathbf{g}) \) is the dimension of vector \( \mathbf{g} \)

\( \mathbf{g}^p \) vector for \( p \in \mathbb{D}_p \)

Matrix Operations

\( \mathbf{A}_{:,j} \) \( j^{th} \) column of matrix \( \mathbf{A} \)

\( \mathbf{A}_{i,:) \) \( i^{th} \) row of matrix \( \mathbf{A} \)

\( \sigma(\mathbf{A}) \) is the singular value vector of the matrix \( \mathbf{A} \)

\( \text{trace}(\mathbf{A}) \) trace of the matrix \( \mathbf{A} \), that is \( \text{trace}(\mathbf{A}) = \sum_i A_{i,i} \)

\( \mathcal{N}(\mathbf{A}) \) is null space of matrix \( \mathbf{A} \)

\( \text{rank}(\mathbf{A}) \) is rank of matrix \( \mathbf{A} \)

\( \text{span}_{\text{row}}(\mathbf{A}) \) vector space generated by rows of \( \mathbf{A} \)

\( \mathbf{A}^\dagger \) is pseudo inverse of matrix \( \mathbf{A} \)

\( \mathbf{A}^\perp \) spans the null space of matrix \( \mathbf{A} \)

\( \mathbf{A}^T \) is transpose of the matrix \( \mathbf{A} \)

\( \mathbf{A}^* \) complex conjugate transpose of matrix \( \mathbf{A} \)

\( \mathbf{A} \geq 0 \) denotes a positive semi definite matrix

\( \text{diag}(\mathbf{g}) \) returns a square diagonal matrix with the elements of vector \( \mathbf{g} \)
\textit{blkdiag}(A^1, \ldots, A^n) \quad \text{block diagonal matrix with the matrices } A^1, A^2, \ldots, A^n

\[ a_{i,j} \quad \text{denotes the } i\text{th row and the } j\text{th column of matrix } A \]

\textbf{Norms}

\[ ||A||_F \quad \text{is frobenious norm of the matrix } A, \text{ that is: } ||A||_F \overset{\dagger}{=} \sqrt{\text{trace}(AA^T)} \]

\[ ||A||_* \quad \text{is nuclear norm of the matrix } A, \text{ that is: } ||A||_* \overset{\dagger}{=} \sum_i \sigma_i(A) \]

\[ ||g||_{\ell_p} \quad p\text{-norm in } \mathbb{R}, \text{ that is: } ||g||_{\ell_p} \overset{\dagger}{=} \sqrt[p]{\sum_{i=1}^{N} g_i^p} \]

\[ ||g||_{\ell_2} \quad \ell_2\text{-norm in } \mathbb{R}, \text{ that is: } ||g||_{\ell_2} \overset{\dagger}{=} \sqrt{\sum_{i=1}^{N} g_i^2} = \sqrt{g^T g} \]

\[ ||g||_{\ell_\infty} \quad \ell_\infty\text{-norm of the vector } g \in \mathbb{R}, \text{ that is: } ||g||_{\ell_\infty} \overset{\dagger}{=} \text{sup}_i |g_i| \]

\[ ||g||_{\ell_0} \quad \ell_0\text{-quasinorm or zero norm } \overset{\dagger}{=} \text{number of non-zero elements in vector } g, \quad ||g||_{\ell_0} \overset{\dagger}{=} \# \{ i : g_i \neq 0 \} \]

\[ ||G(z)||_{\infty} \quad \mathcal{H}_\infty \text{ norm of a stable transfer function: } ||G(z)||_{\infty} \overset{\dagger}{=} \text{sup}_{|z|=1} |G(z)| \]

\textbf{Miscellaneous}

\[ \Upsilon^N \{ \cdot \} \quad N\text{-length impulse response vector of a discrete transfer function} \]

\[ \mathbb{D}_\rho \quad \text{Origin centered closed disc in } \mathbb{C}, \text{ with radius } \rho \]

\[ P(\Omega) \quad \text{Probability of an event } \Omega \]

\[ Z^{-1} \quad \text{Inverse of the unilateral } z\text{-transform} \]

\[ \lfloor x \rfloor \quad \text{denotes the largest integer less than or equal to } x \]

\[ \mathcal{N}(\mu, \sigma^2) \quad \text{Standard Gaussian distribution with mean } \mu \text{ and variance } \sigma^2 \]

\[ g * u \quad \text{Convolution of signal } g \text{ with } u \]

\[ U(a, b) \quad \text{Generate values from the uniform distribution on the interval } [a, b] \]
Optimization

\[ \min_x \text{ Function minimization over } x, \text{ optimal function value is returned} \]

\[ \arg\min_x \text{ Function minimization over } x, \text{ optimal value of } x \text{ is returned} \]

\[ s.t. \text{ Subject to the constraints} \]
List of Abbreviation

CHESS  Center for Health Enhancement Systems Studies
ACHESS  Alcohol-CHESS
MISDP  Mixed Integer Semi Definite Programming
SDP  Semi Definite Programming
ILD  Intensive Longitudinal Data
LTI  Linear Time Invariant
MIMO  Multi-Input Multi Output
SISO  Single Input Single Output
LTV  Linear Time Variant
LMI  Linear Matrix Inequality
ADHD  Attention-Deficit/Hyperactivity Disorder
LASSO  Least Absolute Shrinkage and Selection Operator
MPC  Model Predictive Control
EMA  Ecological Momentary Assessments
GPS  Global Positioning System
SU  Smoking Urge
SE  Self Efficacy
NA  Negative Affect

T  Treatment

I.C.  Initial Conditions
Acknowledgments

First and foremost, I would like to extend my infinite thanks to the Almighty Allah, who is the source of all knowledge and is most gracious.

It has been my great honor to work with my advisor and committee chair, Prof. Constantino Lagoa. He is a distinguished mentor and an amiable friend. He has given me the hope and encouragement to finalize my PhD over the past 5 years. Thank you for giving me a great amount of inspiration and constant support on how to become a better researcher. He will be my role model for the rest of my life with his excellent research skills and enthusiasm, and his sincerity in his research.

I would like to thank the members of my dissertation committee, Prof. Kenneth Jenkins, Prof. Minghui Zhu, Prof. Stephanie T. Lanza and Prof. Necdet Serhat Aybat, for their time actively reviewing my dissertation and providing me with valuable feedback.

I would like to continue by sincerely acknowledging Prof. Mario Sznaier, Prof. Suzan Murphy, and Dr. Burak Yilmaz, from whom I learned many research concepts and skills through their consultation, interaction, and cooperation.

I am also thankful to my lab-mates, Ashkan Jasour and Ibrahim Ekrem Bardakci (he was always there to walk with me to the coffee shop) for their help and friendship. There are also many people, whom I would like to thank, who have provided me with valuable guidance, technically as well as emotionally, which has contributed to the final making of this thesis.

I have to certainly acknowledge my family, especially my dad and mom for their love and support. Last but not the least, I am deeply indebted to my wife, Saliha, and my daughter, Halime Reyhan, for all of their love and for putting up with me through all of the stressful times. I thank you will all my heart, Saliha and my little one.
Dedication

To my mom...
Chapter 1

Introduction

Adaptive treatment strategies approximately resemble the prescription process for a real patient where a clinician alters the dosage or type of treatment according to the patient’s response and his/her current behavioral situations. The goal of the adaptation is to maximize the performance of the treatment often under some constraints. Behavioral treatments aimed at chronic diseases and illnesses such as AIDS, depression, substance abuse, and ADHD, may require sequences of treatments based on the characteristics, behaviors, and responses of individuals. Thus determining the optimum treatment sequence for individuals is not an easy task. Therefore behavioral and social scientists have started discussing the advantages of adaptive treatments [6]. They realized that adaptive interventions for patients might give better results than fixed treatment (in which all patients get the same type and level of treatment). Adaptive treatment strategies are better because they look at the current state of the patients and regulate the adaptation of behavioral interventions. In these interventions, sequences of treatments are adapted and re-adapted to individual circumstances and behaviors in order to achieve and maintain health behavior change.

Clinical and behavioral scientists have several techniques for designing adaptive treatments using statistical methods for behavioral problems [7, 8, 9]. Scientists usually prefer to use a static model for a patient’s behavior since the available behavioral data is usually non-intensive, which means that the data is gathered when the time between two consecutive data points is relatively “large”. Currently available treatment methods are usually effective when the behavioral data is non-
intensive. Behavioral data has usually been non-intensive because patients were contacted infrequently (once a month, two times a month, once a week, etc.) to change the dosage, type, and frequency of treatment. For example, if the patient starts one dosage or type of treatment in the first clinical visit, he/she uses this treatment until the second visit. However, health care providers will not know if the treatment is ineffective or unnecessary during the time between visits. Therefore, the use of portable devices might be useful. Portable devices offer real-time tracking of the health status in real time. As a result, clinical and behavioral scientists can collect intensive longitudinal data (ILD), where the time between two consecutive data points is very “small”, since patients can be easily and frequently contacted via portable devices [10, 7]. A static model can be acceptable for non-intensive data, but if the patients behavioral data is ILD, we believe that the static model might not be an accurate model for the patient’s behavior. The availability of ILD allows us to use dynamical models for behaviors. Therefore if one uses intensive data, it is natural to assume that the current value of the patient’s state is dependent on the previous value of the state [11, 12, 13].

Clinicians usually choose the treatments based on the current state of patient and their responses to the past treatments. Moreover, clinicians generally take this dynamic feature into account in a slow manner since they adjust the treatment sequence during the clinical visit. On the other hand, designing the personalized treatment with the available methods is complicated by the intensive nature of the data if one wants use all the characteristics of the ILD. Additionally, determining decision rules to guide dynamically changing the treatment dosage and type in real-time is also not an easy task.

The field of control engineering has several possible approaches which can address precisely these challenges that come from the structure of the ILD. Therefore, this dissertation introduces a new robust adaptive treatment method from the field of control engineering. *A series of treatments is called a Robust Adaptive Treatment if treatments are adapted and re-adapted to individual patient behaviors in real-time in order to improve or maintain health behavior under unmeasurable perturbations* [11, 13]. We explain ways of using available ILD collected from the patient by a portable device, to leverage the capacity of mobile devices to develop personalized adaptive treatment algorithms. We also explain how control engineering and ro-
bust optimization methods can be utilized to handle the ILD and develop these robust adaptive treatment algorithms while effectively using the portable devices. In order to have a good adaptation performance, we use the concept of feedback from control engineering. We are able to reformulate control engineering methods to develop an effective robust adaptive treatment algorithm. This is a new application area of feedback control to address very specific challenges.

The robust adaptive treatment design process has been developed in three main steps and we cover step 1 and 2 in this dissertation: 1) determine a dynamical model that accounts for uncertainty by using the study/individual data for the person’s behavior and treatment response; 2) develop algorithms based on the model, a person’s current status, and the context, in order to provide a treatment or sequence of treatments that will achieve the desired results for all probable values of the uncertainty; and 3) develop a mobile app or implementation tools for portable devices to run these algorithms or contact patient in real time in order to deliver the treatments. Mobile technology, such as the smart-phone, is central to this process of delivering interventions informed by feedback on a person’s past behavior and historical responses to different interventions.

The remainder of this introduction briefly reviews the adaptive treatment design steps in this dissertation.

1.1 Model and Identification

One can approximately model patient behavior as a dynamical model. The main reason is that the response to treatment is not instantaneous, and might stretch for a long time period (e.g., due to effect latency or behavioral maintenance) [11, 12, 13, 9]. Moreover, although ILD provides very detailed descriptions of the behavior, there are still many difficulties with modeling and identification of the dynamical model. One of the difficulties is that a precise model structure of a specific behavior is not known. Another difficulty is that the treatment responses of patients differ from one person to another due to the diverse characteristics of patients and environmental factors. In order to overcome these difficulties, a “right” dynamical model with correct type of uncertainties can provide good “coverage” of all admissible behaviors and be suitable to the robust adaptive intervention design.
In this dissertation [11, 13], for instance, affine Linear Time Invariant (LTI) model [14, pages 7-16] is preferred as a dynamical model that approximately describes most of the patients’ responses. We use a simple affine LTI model since it is easy to identify, and can be easily used to design adaptive interventions. Next, in order to cover the differences in the patients’ responses, the type of uncertainties in Linear Time Invariant model fall into two classes: i) Differences in the behavior of different participants is defined as Gaussian noise; and ii) sparse external perturbations are used to model life events that affect the response to treatment.

The model could be a difference equation [15, pages 573-578], a differential equation [15, pages 104-139] or a hidden Markov model [16, page 191] with appropriate uncertainties. In this dissertation, we use difference equation in the following format:

\[
y_{k+1} = f(y_k, y_{k-1}, \ldots, y_{k-n}, \delta_k, T_k), \quad T_k \in \mathcal{T} \text{ and } \delta_k \in \mathcal{U}
\]  

(1.1)

Hence, the model \( f \) in equation (1.1) is a difference equation with uncertainties \( \delta_k \). This uncertainty \( \delta_k \) can be noise that represents both the uncertainty in the model and the uncertainty in the measurements, or they can be sparse perturbations that represent infrequent external events (being fired, losing a relative, divorcing, etc.) which can significantly affect the behavior [11, 13]. In equation (1.1), the state at time \( k + 1 \) is a function of \( n \) previous values of the state (order of the difference equation), treatments in \( \mathcal{T} \), and uncertainties in \( \mathcal{U} \) at time \( k \).

Starting with the ILD from one person or a previously chosen study (e.g., people at risk for developing cancer due to lifestyle behaviors, people who want to quit smoking, or people who have an obesity problem), we need to identify dynamic model \( f \) and the bounds/probabilistic descriptions of the inherent uncertainty \( \delta \in \mathcal{U} \) in the equation (1.1). This dynamical model and the uncertainties together approximate the behavior and treatment response of the person. However, even if the structure of the precise model and uncertainties are available, this identification step is still difficult due to the nature of the behavioral data.

First of all, most of the behavioral data is inherently more noisy than the data from physical systems. Second, it is difficult to collect data sample at equal time intervals since patients might ignore providing the information when they are
asked. Moreover, often ecological momentary assessments (EMA), which are designed to sample time, are designed to collect data at irregular spacing and different time points across individuals. This causes having random/missing samples in the study data. Third, the behavioral data is generally collected from many different patients while each patient has different treatment and initial level of the states. Therefore we need an identification algorithm that uses all available data to identify the model parameters, and the size of the uncertainties and to interpolate the incomplete part of the data. Furthermore, we also prefer to have low order model which makes controller design step easier. A lower order model is also called as parsimonious model that accomplishes a desired level of explanation of data with as few model parameters as possible. Thus, in this dissertation, we develop a new parsimonious model identification algorithms for single-input single-output (SISO) and multi-input multi-output (MIMO) systems [1, 17, 18].

Next, we will briefly explain how robust adaptive treatment algorithms can be designed with the available dynamical model and uncertainties.

1.2 Robust Adaptive Treatment

With the identified model at hand, as well as the uncertainty description, we then develop an algorithm that provides the sequence of treatments that will achieve the desired results despite likely perturbations/uncertainties, given the current state of the person. On the other hand, the intensive nature of the intervention makes the decision process of the treatment sequence (type, dosage) highly complex. It is complex since the treatment should be provided repeatedly over time. Therefore, in order to overcome these complexities, the feedback or adaptation, as used in control, has been looked at as a way to develop a new more efficient method for the robust adaptive treatment algorithm in this dissertation [11, 12, 13].

An intuitive description of how a robust adaptive intervention works is given in Figure 1.1. Once a model and the description of the uncertainty $\delta \in \delta$ in equation (1.1) for the patients' behavior is available (as a result of previous step), it is feasible to simulate possible reactions to the treatment. In the robust adaptive treatment algorithm, either implicitly or explicitly, the algorithm simulates the range of probable responses to a treatment under the uncertainties. For this range
Figure 1.1: Robust Adaptive Treatment

of probable responses, a treatment is then selected that has the best possible outcome. The process then repeats to incorporate the data on the new state of the person.

As mentioned above, this dissertation employs control engineering methods, particularly Model Predictive Control (MPC) or more precisely robust MPC design is for the intervention design [11, 12, 13]. This method is utilized for designing a robust adaptive treatment algorithm that achieves a specified objective for all probable uncertainty values. This robust adaptive treatment design problem is formulated as a robust optimization problem with a finite time horizon. In this optimization problem, the prescribed objectives of the treatment can be mathematically defined as an objective function or performance measure. Additionally, the optimization problem might have additional constraints. For example, one can have a set $\mathcal{T}$ for the available possible treatments (treatment set) or there might be additional constraints to bound the total given treatment in the receding/finite future horizon time to prevent overburdening. One should note that such a setup is very common in control systems engineering and many approaches are available.
to solve this type of robust optimization problem.

Next, it is necessary to develop an efficient implementation of these robust adaptive intervention algorithms in portable devices. One of the possibilities is to use a central computer that solves the optimization problem and thus determines the sequence of treatments that should be given to the patient. However, it is desirable to have efficient, decentralized implementations that can be installed on portable devices. This decentralized implementation would allow for the continued application of the treatment even in the event of a failure in communications with the central computer.

Figure 1.2: Implementation of Algorithm

Figure 1.2 shows how this algorithm can be implemented for real life problems by using portable devices. Obviously, in the first step, the dynamical model and the size of the uncertainties $\delta \in \mathcal{U}$ in equation (1.1) should be identified by analyzing the data (Step 1). Once the dynamical model and size of the uncertainties are available, we can define the treatment objectives mathematically and run the robust adaptive treatment algorithm for the individual (Step 2).

1.3 Contributions and Outline

This section summarizes the contribution of this dissertation and the thesis’ organization. The contribution of this thesis is exhibited in 5 Chapters. After giving
the background section in Chapter 2, we initially explain the new identification methods for behavioral data (see Chapters 3, 4 and 5). Second, we show the design process of the robust adaptive intervention method (see Chapter 6).

First of all, we briefly give the definition of adaptive treatment and explain the importance of robust adaptive treatment and personalized treatment design. Furthermore, available methods for adaptive treatment design and their insufficiency are introduced. Then we explain how our methods in this dissertation can address the insufficiency of available methods. Finally the three main steps of the design process of the adaptive treatment are briefly described. These steps will be deeply expanded in the following chapters.

In Chapter 2, several mathematical concepts are presented. We use these concepts to explain our identification and controller design algorithms. First, some basic mathematical definitions are presented. Second, the Atomic norm, a new concept in the identification literature, is explained.

In Chapter 3, we present the Lasso (the least absolute shrinkage and selection operator) algorithm which is utilized to identify a behavioral model in this dissertation. We use Lasso since it is a selection method for linear regression which estimates model parameters and noise level at the same time for high-dimensional models.

In Chapter 4, with the behavioral treatment design under consideration, the objective is to identify a SISO LTI system from the behavioral data that is collected when the system is driven many times by different but known inputs (treatment sequence of different patients) with different unknown initial conditions at every trial. We developed a new, fast, and efficient identification algorithm to identify a low order model that explains the behavioral data.

In Chapter 5, the objective is to identify a MIMO LTI system from the behavioral data set that is again incomplete and noisy. Even there are several identification algorithms in the literature to identify models from noisy and incomplete data, these algorithms usually solve optimization problem with many extra variables to interpolate missing parts which makes the optimization problem computationally complex. Therefore, in this chapter, we develop a new computationally efficient identification algorithm to identify a low order MIMO system.

In Chapter 6, we explain that the behavioral and clinical literature have several
methods that use non-intensive data to develop personalized treatment. On the other hand, evaluating the intensive longitudinal data (ILD) with available techniques is not an easy task because of the complex nature of ILD. Therefore, in this Chapter, we explain how control engineering methods can be adapted to handle challenges arising from the intensive nature of the available data while effectively using portable devices. We also explain how control engineering methods can be reformulated in development of robust adaptive treatment design.

In Chapter 7, we give concluding remarks and some future direction of this research.
This chapter presents the mathematical background and some basic definitions as foundation for understating system identification (Chapter 4 and 5) and robust adaptive treatment design (Chapter 6). After giving a brief review of system identification, this chapter explains a new concept discussed in the literature, the Atomic Norm, which is the main tool for developing our new system identification algorithms presented in Chapter 4 and 5.

Finally, the Model Predictive Control (MPC) method from the field of control engineering is summarized in this chapter. MPC will be utilized mainly in the robust adaptive treatment design in Chapter 6.
2.1 Definitions

2.1.1 Basic Definitions

Definition 1. The Block Hankel matrix associated with any given vector sequence \( \{y_k\}_{k=0}^{n} \) (\( y_k \) is a \( m \times 1 \) column vector) is defined as:

\[
\mathcal{H}_y(r) = \begin{bmatrix}
y_0 & y_1 & y_2 & \cdots & y_{n-r} \\
y_1 & y_2 & y_3 & \cdots & y_{n-r+1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y_r & y_{r+1} & y_{r+2} & \cdots & y_n
\end{bmatrix}
\]

Definition 2. \( T^n_u \) is defined as a lower triangular block Toeplitz matrix associated with a column vector \( \{u_k\}_{k=0}^{n-1} \) (\( u_k \) is a \( p \times 1 \) column vector):

\[
T^n_u = \begin{bmatrix}
u_0 & 0 & \cdots & 0 \\
u_1 & u_0 & \ddots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
u_{n-1} & u_{n-2} & \cdots & u_0
\end{bmatrix}
\]

Definition 3. A discrete time LTI system can be represented as:

\[
\begin{align*}
x(k+1) &= Ax(k) + Bu(k) \\
y(k) &= Cx(k) + Du(k)
\end{align*}
\]

where \( x \in \mathbb{R}^{n \times 1} \) is the state vector, \( u \in \mathbb{R}^{p \times 1} \) is the input vector, \( y \in \mathbb{R}^{m \times 1} \) is the output vector, \( A \in \mathbb{R}^{n \times n} \) is the “state or system” matrix, \( B \in \mathbb{R}^{n \times p} \) is the “input” matrix, \( C \in \mathbb{R}^{m \times n} \) is the “output” matrix, \( D \in \mathbb{R}^{m \times p} \) is the “feed-through or feed-forward” matrix.

Definition 4. The relationship between input and output for a discrete time LTI n-th order SISO systems is given as \( y(z) = H(z)u(z) \), where:

\[
H(z) = \frac{y(z)}{u(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_n z^{-n}}{(1 + a_1 z^{-1} + \ldots + a_n z^{-n})}
\]
\(H(z)\) is called the transfer function of the system, which is the \(z\) transform of the impulse response sequence. \(z^{-1} \) in the transfer function (2.2) is the unit time delay. The coefficients \(a_i\)'s and \(b_j\)'s in the transfer function equation (2.2) are the system's characteristics.

### 2.1.2 Atomic Norm

In this dissertation, the Atomic norm concept is used to develop new system identification tools for dynamical model of behaviors. The atomic norm approach is preferred since minimizing the it usually results in sparse representations (low order or with a minimum number of poles) of LTI system. The idea used is based on the representation of the response of an LTI system as a linear combination of suitably chosen objects (atoms). We use this fact to develop our system identification algorithms in Chapter 4 and 5.

The definition of the Atomic norm and set are initially introduced in [19].

**Definition 5.** [19] Assume \( A \) is the Atoms that is a compact set of \( \mathbb{R}^p \). The elements of \( A \) are the extreme points of \( \text{conv}(A) \) (Convex hull of the set \( A \)). Then assume \( x \in \mathbb{R}^n \) be a given linear measurement and Atomic decomposition of the measurement vector \( x \) can be defined as:

\[
x = \sum_{i=1}^{k} c_i a_i, \quad a_i \in A \text{ and } c_i \leq 0
\]  

(2.3)

where \( A \) is Atomic set which contains the simple elements of general signals.

Then without loss of generality, assume that \( A \) is a centrally symmetric Atomic set and its center is origin then gauge (Minkowski) function of \( A \) can be written:

\[
\|x\|_A = \inf\{t > 0 : x \in t \text{conv}(A)\} = \inf \left\{ \sum_{a \in A} |c_a| : x = \sum_{a \in A} c_a a, \quad c_a \geq 0 \forall a \in A \right\}
\]

(2.4)

If set \( A \) is centrally symmetric about the origin, then it is proven in [19] that \( \|\cdot\|_A \) is a norm and is defined as an Atomic Norm and induced by \( A \). Finally,
since the Atomic norm is defined as a convex hull of the elements of Atomic sets $\mathcal{A}$, it is always a convex set.

This definition states that the measurement vector $\mathbf{x}$ is a linear combination of the elements of the Atomic set. Therefore, because of the properties of the linear system [14], a Linear Time Invariant (LTI) model of a behavior can be written as a linear combination of low order stable models if the Atomic set is defined such that it contains all of these low order stable models. As an example of an Atomic norm technique, a sparse vector can be written as a linear combination of other sparse vectors if the Atomic set contains all of these sparse vectors at once. For example, we may assume the Atomic set consists of unit vectors and its convex hull, represented in Figure 2.1, is

$$
\text{conv}(\mathcal{A}) = \{x, \|x\|_\mathcal{A} \leq 1\}.
$$

![Figure 2.1: Set of Standard Basis Vectors and Extremes](image)

There are many different application areas of the Atomic norm method. In this dissertation, the Atomic norm concept is utilized for system identification. Because of the problems in behavioral data mentioned in Chapter 1, the Atomic norm technique (in Chapters 3 and 4) is well suited to develop a new identification
algorithm for behavioral data.

### 2.2 System Identification

System identification is a crucial task in control systems that aims to find a mathematical model from a noisy and partially known input/output experimental data set supplemented with some a-priori information. System identification from experimental data is a fairly mature area in the field of control engineering especially for LTI systems [20]. There are many different methods in this area such as the Prediction Error Method (PEM) [20], which usually chooses a certain structure of the model and tries to find the parameters of a chosen model by utilizing the stochastic methods. Another method is the Expectation Maximization (EM), which utilizes iterative algorithms [21] to solve nonlinear optimization problem. A newer and more efficient method is the subspace identification method [22], which uses several matrix manipulations from Hankel norm factorization. In addition to these methods, there are deterministic worst-case controller oriented set membership methods and interpolation theory [23]. Moreover, these LTI system identification methods are performed in three different way: state-space identification, finite impulse response identification and transfer function estimation. We give a complete literature review of system identification in Chapter 3. Figure 2.2 depicts the fundamental setup of the system identification. Here $u$ represents controlled input, $y$ is the system output that is associated with input $u$, and $v$ is noise.

![System Identification Setup](image-url)
In this dissertation, we developed identification algorithms using the Atomic norm method to identify the finite impulse response or state-space model from an experimental input/output data set.

2.3 Model Predictive Controller

Model Predictive Control (MPC) is also known as Receding Horizon Control or Moving Horizon Optimal Control or Future Horizon. It is widely used in many control application problems since MPC is able to solve multivariable constrained control problems. In the MPC algorithm, an explicit model of a process is utilized to predict the future progress of this process to optimize the performance of the controller. Similar to general control algorithms, MPC also uses the explicit model of a plant to be controlled to predict the future output behavior of the plant. This prediction allows us to solve optimal control problems online to track the state or reference (mainly error) and obtain the desire output. This minimization problem is solved over a future horizon subject to constraints on the inputs, outputs, and the system states. The performance index is convex stage function $\ell$ in (2.5). For instance, assume at time $k$, we solve an optimal control problem over a finite receding horizon of $K$ steps, then we have following problem:

$$
\min_{u(k),...,u(k+K-1)} J = \sum_{t=k}^{k+K-1} \ell(x(t),u(t))
$$

$$
s.t \quad (x(t),u(t)) \in \mathcal{X}, \quad y(t) \in \mathcal{Y}, \quad t = k, k+1,...,k+K-1$$

$$
x(t+1) = f(x(t),u(t))$$

$$
y(t) = g(x(t),u(t))$$

"stability constraints"

$$
x(k-1) = z \rightarrow \text{Initial Conditions}
$$

where constraints $x(k+1) = f(x(k),u(k))$ and $y(k) = g(x(k),u(k))$ is the process model. The objective function or performance measure can be reformulated according to the control problem. We also have constraint on input and its state responses, and output where they belong given convex input, states and output
constraint sets \( \mathcal{X} \) and \( \mathcal{Y} \). Also initial state belongs to the set \( \mathcal{X} \). Additionally, we have also stability constraint. Lastly, one must ensure that the feasible solution of optimization problem at each time \( k \) must exist.

The concept of MPC can be summarized as follows. First, MPC obtains the current states \( x(k) \) at time \( k \) and solves the optimization problem along the previously decided receding horizon (input horizon) and constraints to calculate the optimal input. Second, the first element of the optimal input vector is applied to the system. After the input is applied, then current state values are calculated again and the optimization problem is solved again for the next optimal input vector. This iteration is repeated to the end of the output horizon. We also depict this iterative procedure in Figure 2.3(b).

- At time \( k \), calculate the control actions by solving an optimization problem for receding horizon.
- Obtain the first value of the control sequence and apply it to the system.
- At the next step, obtain the state value of the system after applying the control sequence and re-solve the optimization problem.
- Repeat this procedure for a pre-determined output horizon time.

Figure 2.3(a) also shows the basic feedback structure of the MPC. The main purpose of the MPC is to predict the future output of the system and to calculate the best control action for all possible outputs.

The prediction capability depends on many variables in the system. This dependency mostly affects the structure of the optimization problem. Moreover, according to the structure of the optimization problem, MPC may be able to control tracking error, output behavior, and noise attenuation by manipulating outputs and inputs online. Furthermore, the MPC algorithm is considered robust if the algorithm handles the specific uncertainties and noise range in the dynamics as well as the specific stability and performance criteria. More information about MPC can be found in [24, 25].

In this dissertation, MPC is preferred for the robust treatment design, since MPC has the capability to calculate the worst possible outcome of the future (receding horizon) and gives the best or optimal treatment at every time \( k \) to improve
Figure 2.3: Basic Description of MPC and Horizon Strategy: \( u(t) \) is the input applied to the system.

The treatment process (Chapter 6). This controller design method is well suited to the type of the problems that we are considering in this dissertation. Once a dynamical model is known or identified, the MPC approach could be utilized to optimize treatment dosage and type for different individuals in different circumstances. In this dissertation, the MPC algorithm is utilized mainly in Chapter 6.
Chapter 3

Identification Using Lasso

3.1 Introduction

Lasso (the least absolute shrinkage and selection operator), a method from the field of statistics, is one of the ways used in the behavioral model identification in this dissertation. Lasso is a selection method for linear regression which estimates model parameters and noise level at the same time for high-dimensional models. It minimizes the sum of the squared errors of measurement and estimated output with bounds on the absolute value sum of the coefficients of the regression model. More information about the Lasso can be found in literature in the field of statistics [26, 27, 28].

Furthermore, identifying a simple LTI model, noise and sparse disturbances from the behavioral data with \textit{a priori} which is order of the LTI model is possible with Lasso algorithm. Lasso is well suited for the behavioral context, in which the collection of behavioral data is often done over long periods of time and one cannot usually perform repeated experiments to improve the quality of the data. As a result, unlike most physical systems, behavioral data are generally more noisy and incomplete. Thus, designers of treatments, then, typically work with an incomplete, noisy data set from multiple patients [8, 29, 30, 31, 32, 33, 34]. Therefore, multiple participants are used to attempt to capture a more complete data set such that the gaps in one participant’s data are filled by those of another participant. In other words, to determine a model that is both meaningful and can be used to design treatment for a specific group of individuals, the identification
data is collected from several participants. The output and treatment (input) data of patients are utilized to estimate the model parameters, noise level, and structure of sparse disturbance. Finally, we prefer Lasso since the model parameters and size of the uncertainties can be simultaneously estimated [27, 28].

Moreover, choosing the structure of uncertainties is important as a part of the identification process in this setup. After discussing with scientists and clinicians in the field, we chose the following types of uncertainty. When selecting a type of uncertainty, in this dissertation, the objective is to address i) differences in the behavior of different participants and ii) sparse external perturbations that can be used to model life events that affect the response to treatment. We choose a specific structure for model uncertainty for modeling human behavior. In the approach presented here, inter-subject differences in behavior are modeled as a white noise perturbation. Finally, assuming identification data with a priori order of the model available and the model structure and uncertainty defined, the Lasso is utilized for parameter identification and uncertainty quantification because it provides a way to directly estimate a model subject to sparse perturbations. Note that this identified model and uncertainties will be used in Chapter 6 to develop robust adaptive treatment.

In this Chapter, the structure of the dynamical model that is acceptable to model a behavior is initially presented. In addition to the model, the reasons for choosing the uncertainty type and its complete mathematical description are also explained. Then once the description of the model and uncertainties are available, we show how Lasso can be formulated for behavioral data that is obtained from several patients. Finally, we present a few results from an identification process of a smoking cessation study.

### 3.2 Model

As mentioned in the previous section, data on human behavior usually involves several subjects with many gaps in information over time. Moreover, the behavior of the subjects might change with time and vary from subject to subject. Therefore, it is very unlikely that one can accurately specify a complex model of behavior; one can use a simple model to approximate patient behavior or the main dynamics
of the subject and, at the same time, highlight the fact that there is a substantial amount of uncertainty in it. Hence, a possible model for identification is a set of affine equations.

\[ \mathbf{y}_{k+1} = \mathbf{a}_f + \sum_{i=0}^{n-1} [\mathbf{A}_i \mathbf{y}_{k-i} + \mathbf{B}_i \mathbf{T}_{k-i}] + \mathbf{D} \mathbf{\varepsilon}_k + \mathbf{E} \mathbf{w}_k, \]  

(3.1)

where \( \mathbf{y}_k \in \mathbb{R}^m \) is the output vector at time \( k \), \( \mathbf{T}_k \in \mathbb{R} \), and \( \mathbf{T}_k \in \mathcal{T} \) is the control input (i.e., treatment), where \( \mathcal{T} \) is a set of all available treatments. In many cases, this set has a finite number of elements. Unmeasurable sparse exogenous perturbations are denoted by \( \mathbf{w}_k \in \mathbb{R} \) and \( \mathbf{\varepsilon}_k \in \mathbb{R}^m \) that encompasses uncertainty in the model. The affine term vector is \( \mathbf{a}_f \in \mathbb{R}^m \) and \( \mathbf{A}_i \in \mathbb{R}^{m \times m}, \mathbf{B}_i \in \mathbb{R}^m, \mathbf{D} \in \mathbb{R}^{m \times m}, \mathbf{E} \in \mathbb{R}^m \) are coefficient matrices.

Since the dynamics of behavior are very complex, the model in (3.1) can be used to handle large set of responses to behavioral treatment. In addition to LTI model in (3.1), Gaussian noise \( \mathbf{\varepsilon}_k \) and unmeasured exogenous input \( \mathbf{w}_k \) is also added to the model to represent unexpected events. This dynamical model in (3.1) with uncertainties predominantly can cover the possible responses of treatments.

We use as a motivating example a hypothetical study of smoking cessation. In this example, there are three different variables that one can estimate: smoking urge (denoted \( su \)), negative affect, \( (na) \), which is a single scale indicating an adverse mood state, and self-efficacy, \( (se) \), which represents an individual’s belief in their ability to abstain from smoking. With these variables, the following model is proposed:

\[
\begin{bmatrix}
    s\mathbf{u}_{k+1} \\
    n\mathbf{a}_{k+1} \\
    s\mathbf{e}_{k+1}
\end{bmatrix} = \begin{bmatrix}
    a_f \\
    d_f \\
    g_f
\end{bmatrix} + \sum_{i=0}^{n-1} \begin{bmatrix}
    a_i & b_i & c_i \\
    d_i & e_i & f_i \\
    g_i & h_i & r_i
\end{bmatrix} \begin{bmatrix}
    s\mathbf{u}_{k-i} \\
    n\mathbf{a}_{k-i} \\
    s\mathbf{e}_{k-i}
\end{bmatrix} + \begin{bmatrix}
    0 \\
    q_i \\
    s_i
\end{bmatrix} \mathbf{T}_{k-i}
\]

(3.2)

+ \begin{bmatrix}
    1 \\
    0 \\
    0
\end{bmatrix} \mathbf{w}_k + \begin{bmatrix}
    1 & 0 & 0 \\
    0 & 1 & 0 \\
    0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
    \mathbf{\varepsilon}^{(su)}_k \\
    \mathbf{\varepsilon}^{(na)}_k \\
    \mathbf{\varepsilon}^{(se)}_k
\end{bmatrix}. \]
Remark 1. The signs of several of these coefficients are assumed to be known in advance for this example. For example, negative affect (na) increases smoking urge (su), whereas self efficacy (se) decreases it. More information about the dynamics of the smoking urge can be found in [35, 36]. This information can be used in two different ways: i) it can be given to the identification algorithm as additional constraints or ii) it can be utilized to validate the model with the real output measurements.

This model also shows that smoking urge cannot be directly controlled, but if the negative affect can be decreased and the self efficacy can be increased by applying treatment $T$, the desired smoking urge level\textsuperscript{1} can be achieved under the uncertainties [35, 36]. This is also crucial information while we are designing an adaptive treatment. This information can again be given to the optimization problem as an additional constraint.

One should first note that the model above in (3.1) is an uncertain affine model that is called a structural equation model (SEM) and has been utilized in the social sciences [37, 38]. An affine model is preferred because the equilibrium point of the uncontrolled system is often not the origin for behavioral systems.

The model is obtained by using data from several patients and differences in patients are typically modeled as Gaussian random variables $\varepsilon$. The external perturbation $w \in L^\infty$ is also motivated by a different kind of uncertainty in human behavior. The way subjects behave is influenced by external “sparse” events that temporarily affect his/her response to treatment. Hence, one needs an uncertainty that is sparse, bounded in magnitude and with limited cumulative effect. As a result, it is assumed that the segment of perturbation signal $w \in L^\infty$ is bounded in the $\ell_1$-norm, $\ell_\infty$-norm, and $\ell_0$-norm. For this situation, a possible signal set $W$, which is used in this article, is the following. Given a “MPC control window” of length $K$ and time $k$, the set $W$ is:

$$W = \{ w \in L^\infty : \text{for all } k, \|w_{(k:k+K-1)}\|_\infty \leq \alpha, \|w_{(k:k+K-1)}\|_1 \leq \gamma, \|w_{(k:k+K-1)}\|_0 = K/\varsigma \}. \quad (3.3)$$

\textsuperscript{1}Desired smoking urge level is actually zero but because of the uncertainties, smoking urge level may not always remain zero.
\[ \alpha \in \mathbb{N} \text{ is bound on the magnitude of the perturbation, } \gamma \in \mathbb{N} \text{ is bound on cumulative effect, and } \varsigma \in \mathbb{N} \text{ enforces the sparsity constraint on vector } \mathbf{w}_{(k:k+K-1)}. \]

### 3.3 Identification of Model and Perturbations

To estimate the coefficients from study data, we start by noting that the model above can be taken to represent the relations among the variables as

\[ Y = H\beta + \varepsilon. \quad (3.4) \]

In this model, the unknown vector \( \beta \) contains the parameters of the model and vector \( w \) and \( H \) is a function of the measurements and inputs (treatment). For the smoking cessation example, all the structure and the dimension of the measurement vector \( Y \), unknown vector \( \beta \), and matrix \( H \) in (3.4) are given in next section. Note that this is a regression model.

### 3.3.1 Formulation of Identification Problem

In this section, we show how general equations (3.1) or (3.2) can be reorganized for the Lasso algorithm. Start with the data set \((su, na, se)\) of \( P \) patients. Assume that for each \( i^{th} \) patient we have collected data set \( su^i_k, na^i_k, se^i_k \) for \( k = 1, 2, \ldots, N \) and \( i = 1, 2, \ldots, P \). Note that we eliminate the missing parts of the data for each patients. Also assume the order of the model \( n \) is given. Then, for each patient \( i \), build the following matrices:

\[
X_{1,i} = \begin{bmatrix}
1 & su^i_n & \ldots & su^i_1 & na^i_n & \ldots & na^i_1 & se^i_n & \ldots & se^i_1 \\
1 & su^i_{n+1} & \ldots & su^i_2 & na^i_{n+1} & \ldots & na^i_2 & se^i_{n+1} & \ldots & se^i_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & su^i_{N-1} & \ldots & su^i_{N-n} & na^i_{N-1} & \ldots & na^i_{N-n} & se^i_{N-1} & \ldots & se^i_{N-n}
\end{bmatrix}
\]

and

\[
X_{2,i} = \begin{bmatrix}
1 & su^i_n & \ldots & su^i_1 & na^i_n & \ldots & na^i_1 & se^i_n & \ldots & se^i_1 & T^i_n & \ldots & T^i_1 \\
1 & su^i_{n+1} & \ldots & su^i_2 & na^i_{n+1} & \ldots & na^i_2 & se^i_{n+1} & \ldots & se^i_2 & T^i_{n+1} & \ldots & T^i_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & su^i_{N-1} & \ldots & su^i_{N-n} & na^i_{N-1} & \ldots & na^i_{N-n} & se^i_{N-1} & \ldots & se^i_{N-n} & T^i_{N-1} & \ldots & T^i_{N-n}
\end{bmatrix}.
\]
Also, define

\[ su^i = [su^i_{n+1} \, su^i_{n+2} \cdots su^i_N]^T \text{ where } su^i \in \mathbb{R}^{N-n} \]

\[ na^i = [na^i_{n+1} \, na^i_{n+2} \cdots na^i_N]^T \text{ where } na^i \in \mathbb{R}^{N-n} \]

\[ se^i = [se^i_{n+1} \, se^i_{n+2} \cdots se^i_N]^T \text{ where } se^i \in \mathbb{R}^{N-n} \]

and

\[ SU = \begin{bmatrix} su^1 \\ su^2 \\ \vdots \\ su^P \end{bmatrix}, \quad NA = \begin{bmatrix} na^1 \\ na^2 \\ \vdots \\ na^P \end{bmatrix}, \quad SE = \begin{bmatrix} se^1 \\ se^2 \\ \vdots \\ se^P \end{bmatrix}, \]

where \( SU \in \mathbb{R}^{P(N-n)}, \ NA \in \mathbb{R}^{P(N-n)}, \ SE \in \mathbb{R}^{P(N-n)} \).

Finally, let

\[ \theta_{su} = \begin{bmatrix} a_f & a_0 & a_1 & \cdots & a_n & b_0 & b_1 & \cdots & b_n & c_0 & c_1 & \cdots & c_n \end{bmatrix} \]

\[ \theta_{na} = \begin{bmatrix} d_f & d_0 & d_1 & \cdots & a_n & e_0 & c_1 & \cdots & b_n & f_0 & f_1 & \cdots & f_n & q_0 & q_1 & \cdots & q_n \end{bmatrix} \]

\[ \theta_{se} = \begin{bmatrix} g_f & g_0 & g_1 & \cdots & a_n & h_0 & h_1 & \cdots & b_n & r_0 & r_1 & \cdots & r_n & s_0 & s_1 & \cdots & s_n \end{bmatrix} \]

\[ \beta = \begin{bmatrix} \theta_{su} & \theta_{na} & \theta_{se} & w \end{bmatrix}^T, \]

where \( w \in \mathbb{R}^{P(N-n)} \). Then, if all the patients satisfy the model provided in (3.2), then we have

\[ Y = H \beta, \quad (3.5) \]
where \( Y \in \mathbb{R}^{3P(N-n)} \), \( H \in \mathbb{R}^{3P(N-n) \times 3(3n+1)+N} \), and coefficient vector \( \beta \in \mathbb{R}^{(11n+3+P(N-n))} \).

\[
Y = \begin{bmatrix} SU \\ NA \\ SE \end{bmatrix}, \quad H = \begin{bmatrix} X_1 & 0 & 0 & I \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & X_2 & 0 \end{bmatrix},
\]

and \( \varepsilon \) is a vector containing all noise. Note that if we do not have treatment (input), we can easily update matrix \( H \) and coefficient vector \( \beta \) as follows:

\[
\tilde{\theta}_{na} = \begin{bmatrix} d_f & d_0 & d_1 & \cdots & a_n & e_0 & e_1 & \cdots & b_n & f_0 & f_1 & \cdots & f_n \end{bmatrix}, \\
\tilde{\theta}_{se} = \begin{bmatrix} g_f & g_0 & g_1 & \cdots & a_n & h_0 & h_1 & \cdots & b_n & r_0 & r_1 & \cdots & r_n \end{bmatrix}, \\
\tilde{\beta} = \begin{bmatrix} \theta_{su} & \tilde{\theta}_{na} & \tilde{\theta}_{se} & w \end{bmatrix}^T,
\]

and

\[
Y = \tilde{H}\tilde{\beta}, \quad \tilde{H} = \begin{bmatrix} X_1 & 0 & 0 & I \\ 0 & X_1 & 0 & 0 \\ 0 & 0 & X_1 & 0 \end{bmatrix}.
\]

### 3.3.2 Lasso

There are several ways to estimate coefficient vector \( \beta \). In the examples provided in this Chapter, Lasso [27, 28] is used as an identification tool. Assuming the order of the model \( n \) is given, this algorithm provides a way to determine an estimate of the parameters that balances the size of the noise \( \varepsilon \in \mathcal{N}(0, \sigma^2) \) and the sparsity of the exogenous perturbation \( w \) in the model. It is also noted that there is a balance between the order of the system and size of the perturbation (magnitude, sparsity, etc.). Sparsity of \( w \) is crucial in this context, as it represents the (sparse) perturbations that a patient faces infrequently.

For the Lasso algorithm, the optimization problem in (3.6) will be a convex minimization of a penalized joint loss function with a certain regularization parameter \( \lambda \) given for the regression coefficient, noise level and \( w \). The usual penalized loss function takes the following form:
\begin{equation}
L(\beta, \sigma) = \frac{\|Y - H\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda\|w\|_1. \tag{3.6}
\end{equation}

However, (3.6) can be re-expressed as (3.7) because the regularization parameter \(\lambda\) and \(\ell_1\) constraint \(\mu\) are one-to-one correspondence by Lagrangian Duality [39].

\[
\{\hat{\beta}, \hat{\sigma}\} = \arg\min_{\beta, \sigma, w} \frac{\|Y - H\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2}
\text{subject to}
\|w\|_1 \leq \mu
\] \tag{3.7}

Again the design matrix \(H\) and the response vector \(Y\) are constructed by measurements from \(\mathcal{P}\) different patients that are explained in previous section. Moreover the penalized loss function (3.6) is jointly convex with respect to the regression coefficient vector \(\beta\) and the noise level \(\sigma\) [27, 28]. Since the problem is jointly convex with respect to \(\beta\) and \(\sigma\), if the initial \(\beta\) is known or given, then \(\sigma\), which minimizes the objective function, can be calculated by differentiating the objective function (3.7) with constant \(\beta\) vector (Step 3 in algorithm 1). The \(\sigma\) is calculated as follows:

\[\hat{\sigma}^2 = \frac{\|Y - H\beta_{j+1}\|_2^2}{n}\]

Then the optimization problem in Step 4 in algorithm 1 is solved to estimate coefficient vector \(\beta\):

\[
\hat{\beta} = \arg\min_{\beta, w} \frac{\|Y - H\beta\|_2^2}{2n\hat{\sigma}} + \frac{\hat{\sigma}}{2}
\text{subject to}
\|w\|_1 \leq \mu
\]

The following Lasso algorithm shows the identification procedure in [27], where \(j\) shows the iteration number.
Algorithm 1 Lasso Iterative Identification Algorithm

1: Given $\mu$, tolerance $\delta$, iteration $j = 0, 1, ..., \beta_0 = 0$, and randomly chosen $\beta_1 = \beta_{\text{initial}}$ \hfill $\triangleright$ Initialization

2: while $|\beta_{j+1} - \beta_j| \leq \delta$ do

3: $\hat{\sigma}^2 \leftarrow \frac{\|Y - H\beta_{j+1}\|^2}{n}$

4: $\hat{\beta} \leftarrow \text{argmin}_{\beta, w} \frac{1}{2} \left( \frac{\|Y - H\beta\|^2}{2n\sigma^2} + \frac{\hat{\sigma}^2}{2} \right)$ subject to $\|w\|_1 \leq \mu$

5: $\beta_{j+1} \leftarrow \hat{\beta}$

6: end while

7: $\sigma \leftarrow \hat{\sigma}$ and $\beta \leftarrow \hat{\beta}$

8: $\alpha$, $\gamma$ and $\varsigma$ $\leftarrow$ Parameters of set $W$ from statistic of vector $w$,

The terms in the objective function $(3.7)$ aim at finding the maximum likelihood estimate [40] of the parameters of the model and the variance of the noise $\varepsilon$ while the $\ell_1$ constraint searches for the sparsest exogenous input. Here the $\ell_1$ norm is used as a convex approximation of the $\ell_0$ norm [41]. Thus given the data and the order of the model $n$, at the end of the identification algorithm, the parameters of the model in (3.2), the noise levels of $\varepsilon \in \mathcal{N}(0, \sigma^2)$, and unmeasurable exogenous input $w \in \mathbb{R}^{P(N-n)}$ are estimated. Then the statistic of identified vector $w \in \mathbb{R}^{P(N-n)}$ is used to define signal set $W$ in (3.3).

Since an uncertain model is searched for all $P$ individuals (patients) in the smoking cessation example, all observations of $su_k^i, na_k^i$ and $se_k^i$ ($i = 1, 2, ..., P$) are used to construct the regression model (3.4). Note that the data from all $P$ individuals are used for system identification.

Finally, given a time interval of length $K$, output vector $y$ is a function of uncertainties $w \in \mathcal{L}^\infty$, $\varepsilon \in \mathcal{N}(0, \sigma^2)$, and the state of the system. As mentioned before $w$ is a bounded sparse uncertainty and $\varepsilon$ is Gaussian noise that models differences in between different subjects. Now, to be able to design robust controllers, we need to have a bounded set for $\varepsilon$.

**Remark 2.** Since Gaussian distributions have unbounded support, we choose a set of high probability. More precisely, given a window of size $K$, the density of $\varepsilon_{k:k+K-1}$ has hyper-spherical contours and we take a set of high probability of the
form
\[ \|\varepsilon_{k,k+K-1}\|_2 \leq \rho \] (3.8)

where \( \rho \) is chosen based on the identified standard deviation of the noise. This also has the added advantage of leading to a formulation of the robust treatment design that is computationally tractable. We refer the reader to [42] for a more in-depth discussion on the advantages of using this approach to describe this type of noise.

Given a time interval of length \( K \), output vector \( \mathbf{y} \) is a function of uncertainties \( \mathbf{w}, \varepsilon \), and the state of the system. This is a crucial information while we are designing robust adaptive treatment in Chapter 6.

### 3.4 Identification Results

In order to test the performance of the approach presented in this Chapter, we emulate a real application. We develop a second order \( (n=2) \) model for smoking urge that approximates the behavior observed in empirical studies such as [43, 44]. Note that in this study, subjects smoke heavily.

Here is the true model of smoking urge:

\[
\begin{bmatrix}
a_f \\
a_0 \\
a_1 \\
b_0 \\
b_1 \\
c_0 \\
c_1
\end{bmatrix} = \begin{bmatrix} .50 \\ .35 \end{bmatrix}
\begin{bmatrix}
d_f \\
d_0 \\
d_1 \\
e_0 \\
e_1 \\
f_0 \\
f_1
\end{bmatrix} = \begin{bmatrix} .15 \\ .01 \\ 0 \\ .49 \\ .25 \\ -.05 \\ 0
\end{bmatrix}
\begin{bmatrix}
g_f \\
g_0 \\
g_1 \\
h_0 \\
h_1 \\
r_0 \\
r_1
\end{bmatrix} = \begin{bmatrix} .52 \\ -.01 \\ 0 \\ -.02 \\ 0 \\ .49 \\ .36
\end{bmatrix}
\]

The studies mentioned above do not contain treatment. Hence, we augmented the model with a nonlinear effect of treatment in order to represent treatment burden. In the literature, treatment burden is defined as increasing the level of treatment to the extent that it may cause suboptimal adherence and even negative outcomes. More information about treatment burden in social and medical sciences
can be found in [45, 46, 47, 48]. More precisely, the following terms were introduced:

\[ q_0 T_{k-1} + q_1 \text{sig} \left[ \sum_{u=k-44}^{u=k-2} T_u \right] T_{k-2} \quad (3.9) \]

and

\[ s_0 T_{k-1} + s_1 \text{sig} \left[ \sum_{u=k-44}^{u=k-2} T_u \right] T_{k-2}, \]

where

\[ \text{sig}(x) = \frac{1}{1 + e^{-(x-17)}}. \]

The sigmoid function or special case of logistic function is represented as \( \text{sig}(x) \). This function explained in Appendix B.2.1. Then the following parameter values are introduced:

\[
\begin{bmatrix}
q_0 & q_1 & s_0 & s_1
\end{bmatrix} = \begin{bmatrix}
-.2 & .2 & .2 & -.2
\end{bmatrix}.
\]

This model for the effect of treatment is designed to emulate the case where the treatment has a positive impact, but its effectiveness decreases when the treatment is applied too frequently.

For this highly uncertain model, 500 different trajectories of the system are generated, with random initial conditions and random realizations of uncertainty. This is done to simulate the behavior of 500 subjects in a study where the treatment was provided at random times. Note that this is a reasonable surrogate of a study since the model is designed to approximate real data, except for the influence of treatment. To validate the model, we can use the information in [35, 36] showing that negative affect (\( na \)) increases smoking urge (\( se \)), where the sign of the coefficient of negative affect (\( na \)) is positive and self-efficacy (\( se \)) decreases the smoking urge (\( su \)), where the sign of the coefficient of self-efficacy (\( se \)) is negative.

To do this, a model of the form (3.2) is used with order \( n = 2 \). The model coefficients obtained for a randomly chosen virtual patient are shown below:
In addition, the description of the uncertainty is, for robust MPC window of \( K = 8 \), the following: \( w(k:k+K-1) \) has 25\% non-zero terms with an \( \|w(k:k+K-1)\|_1 \leq 5 \) and \( \|w(k:k+K-1)\|_\infty \leq 2.5 \). In order to obtain these parameters, the properties of the identified \( w \) in Section 3.3 is used. As for Gaussian uncertainty, \( \varepsilon^{su} \sim \mathcal{N}(0, 1.3) \), \( \varepsilon^{na} \sim \mathcal{N}(0, 0.78) \), and \( \varepsilon^{se} \sim \mathcal{N}(0, 0.55) \) are identified. The sampling period here is 8 hours and, hence, data are collected three times per day. The simulation is run for 50 days, and 150 data points are collected. Finally, the coefficients of negative affect (\( na \)) and self efficacy (\( se \)) indicate that the effect of treatment is very small.

### 3.5 Conclusion

In this section, we show that one can use the Lasso with a given \textit{a priori}, order of the model, to identify a behavioral model and the size of the perturbation. In this identification setup, we concatenate the data from several individuals to identify a model and uncertainties that explain the whole concatenated signal. Since we obtain a vector from all patients’ data, even identified model adequately explain whole combined input/output data, the model might not validate some of the patients while it perfectly explains the others. Therefore we need an algorithm that can evaluate each patient’s data separately to identify a model which independently explains and validates each patient’s data.

Thus, in this thesis, we develop an identification algorithm in Chapter 4 and 5 that independently evaluates the data of each individual and finds a model for each
individual’s data. This identification algorithm also appraises the missing parts of
the each individual data and interpolates these missing measurements.
Chapter 4

Parsimonious Model Identification of SISO Systems from Multi-Run

4.1 Introduction

During the past few years, a considerably research effort has been devoted to the problem of identifying parsimonious models\footnote{Parsimonious model that accomplishes a desired level of explanation of data with as few model parameters as possible.} from experimental data. On the other hand, identification of a parsimonious model of Linear Time Invariant (LTI) systems can be difficult depending on the nature of the data and/or priors available. It is difficult since very often, the available data is noisy and has missing samples. Therefore, parsimonious model identification of the SISO system is a crucial step for robust adaptive treatment design due to the structure of behavioral data that is frequently noisy, incomplete, and collected from several patients.

Thus, in this Chapter, we develop an identification algorithm that can utilize given \textit{a priori} information (the pole location constraints) on Single-Input/Single-Output (SISO) systems, and input/incomplete output data from several patients to identify the lowest order (parsimonious) model that is compatible with \textit{a priori} assumption and the collected data. Additionally, the developed algorithm can simultaneously interpolate the missing part of the data.

Parsimonious model identification problem is generically a non-convex problem. Presently available approaches in the literature are limited to relatively small sized
identification problems, due to its computational complexity. Motivated by these difficulties, a new approach has recently been proposed based on the concept of representing the response of an LTI system as a linear combination of suitably chosen objects (atoms) and the observation that minimizing the atomic norm leads to sparse representations. In this Chapter, we cover the fundamentals of this new approach and show that it leads to a very efficient algorithm to identify dynamical model of behaviors.

First, the new approach (atomic norm) can be employed to accommodate non-uniform sampling and (unknown) initial conditions of LTI systems. Second, the algorithm proposed in this Chapter can exploit noisy/incomplete data from different runs of the same system to determine a low order model that is consistent with the data collected. This algorithm is motivated by the problem addressed that comes from behavioral sciences where studies performed usually involve several patients.

In the behavioral context, we need to determine a dynamical model of behavior that describes the main dynamics in the response to treatment of several patients. The data from one individual is not sufficient to provide a reliable model for all patients. Thus, we consider the data of each patient as the result of non-uniform noisy samples of the response of the same system to different (known) inputs and different (unknown) initial conditions. Therefore, the algorithm needs to consider simultaneously all the data available to determine a “good enough” model that covers all responses of these patients. As a result, we developed an algorithm that is well suited for the identification of the behavioral model from study data.

In this chapter, we include a brief background on previous works and the central problem. Then, the identification problem and the description of the atomic-norm constrained convex optimization problem is presented in detail for single run and/or multiple runs of the same LTI system. We also expand one of the available greedy algorithms to solve the identification problem for multi-run cases. Finally, results of proposed algorithms are illustrated with several identification examples from time domain measurements in this chapter. We then conclude with some remarks.
4.1.1 Literature Review

In the past two decades, researchers have extensively studied non-parametric identification of linear time invariant (LTI) systems based on sets of noise-corrupted measurements accompanied by a priori information. These research has led to numerous approaches, which can be roughly divided into subspace identification [49, 50] and control oriented methods [51, 52, 23].

Subspace-based methods, based on matrix factorization, are computationally attractive, can handle unknown initial conditions, and (approximately) enforce bounds on the order of the resulting model. On the other hand, these methods are not well suited for enforcing the consistency of the model with existing a priori information, such as bounds on the time constants, pole locations (including stability) or interpolation error. Alternatively, control oriented methods, based on interpolation theory, generate a nominal model consistent with both the observed experimental data and the a priori information [52, 23, 53, 54, 55], or show that none exists. The main drawback of these methods is the order of the resulting models (roughly the number of data points used in the identification), usually resulting in the need to perform a model reduction step. Unfortunately, this step may lead to models that are no longer consistent with the data. In addition, since these methods are based on semi-definite optimization, computational complexity grows roughly as the cube of the number of data points, limiting the size of problems that can be solved.

In the past few years, several approaches [56, 57, 58, 59, 60] sought to obtain a framework that combines elements from both approaches in order to obtain parsimonious models while explicitly controlling the approximations error, and in some cases [58, 59, 60] enforcing stability margin and maximum gain constraints. A salient feature of these approaches is the reduction of the original (non-convex) problem to a sequence of convex semi-definite programs, typically through the use of the nuclear norm as a surrogate for rank. However, while these approaches usually work well in practice, there is no guarantee that using these surrogates

\[58\] pursues a different approach where convexification of the original problem is achieved via reduction to a sequence of SDPs via moments-based arguments. While this approach is asymptotically convergent, in practice only a few elements of the sequence of approximations can be computed due to the entailed computational complexity.
will lead to the simplest model explaining the experimental data. In addition, incorporating stability constraints into the formalism entails a substantial increase in the computational complexity. Finally, even though [57, 59, 60] avoid the use of interior point methods by resorting to first order methods, they still require performing singular value decompositions, and potentially require many iterations to converge. Thus, while able to handle substantially larger problems than earlier formulations, they are still somewhat limited in the size of the problems than can be handled. However, none of these methods in the literature could not identify behavioral model from several patients data by evaluating all of them separately and interpolate the missing parts of the data set. Therefore we develop this algorithm for the systems of behaviors.

4.1.2 Dynamical Model Identification of Behaviors

One of the popular application areas of identification is finding simple models to explain certain behavioral traits in humans. In this context, finding a parsimonious model of patient behavior is needed to design adaptive interventions. Assuming that there are underlying governing dynamics that relate certain behaviors of the patient (outputs/tailoring variables) to the treatment (inputs), the identification step becomes vital to designing efficient treatments.

However, the collection of behavioral data is usually done over long periods of time and, hence, one cannot perform repeated experiments in order to improve the quality of the data collected. In this Chapter, we formulate the problem of identifying behavioral treatment responses in humans as a problem of identifying a LTI system from a set of noise-corrupted measurements as it is done in many practical applications of control system design. Therefore, we develop a new tool that uses all the data available to determine reliable behavioral models.

To summarize the problem tackled in this chapter in system theory language, the objective is to identify an LTI system (the behavioral process under consideration) a priori from data collected when the system is driven by different but known inputs (treatments corresponding to different patients) with different unknown initial conditions of the system at every run. The data collected is mostly noisy with possibly missing samples. The objective of the chapter is to develop
a method to identify a low order model that explains the data and complete the missing part of the data from all of the runs (patients) simultaneously.

The proposed algorithm is motivated by the work in [19, 61]. In [19, 61], it is shown that a model which explains data can be written as a linear combination of simple objects; i.e., atoms. Then, by enforcing the algorithm to use as few of these simple objects as possible, a simple representation of the system is obtained. Based on the idea of representing the response of an LTI system as a linear combination of suitably chosen objects (atoms) [19], minimizing the atomic norm leads to sparse representations. In this Chapter, we cover the fundamentals of this new approach and show that it leads to a very efficient algorithm for behavioral data. Indeed, at each step, the proposed algorithm requires computing only inner products and thus computational time scales linearly adapted to the size of the data. The results are illustrated with several examples.

4.2 Problem Statement

Given a finite set of input/noisy fragmented output data and the region $D_\rho$ where the poles of the system are contained, as a priori information, determine a parsimonious SISO LTI model $G(z)$ that is compatible with both the a priori information and the collected data.

$$y_{t_j} = g^{ic}_{t_j} + (g * u)_{t_j} + \eta_{t_j}, \ j = 0, \ldots, m,$$

(4.1)

where $*$ denotes convolution. Here, the signals $u$ and $y$ represent a known input and its corresponding time domain output, corrupted by some noise $\eta$. Moreover, $g^{ic}$ represents the initial condition response and $g$ represents the impulse response of the model $G(z)$ in time domain. We will further assume that the time domain data $y_{t_j}$ has been obtained by sampling at not-necessarily equally spaced, but commensurate time instants:

$$0 \leq t_1 < t_2 < \cdots < t_m = n; \ m \leq N + 1.$$

$^3D_\rho$ is a origin centered closed disc in $\in \subset C$, with radius $\rho$
Without loss of generality, we assume that the sampling instants $t_j$ are integers.

Second, we extend the problem and formalize it for multi-run. More precisely, the second problem addressed is the following: Consider a plant $G(z)$ with input $u$ and output $y$ whose poles are again in $\mathbb{D}_\rho$. Given $i \in \{1, 2, \ldots, P\}$, let $y_i(t_j)$ be the response to a known input $u_i(t)$ and unknown initial conditions. For each of the responses, again assume that one collects noisy measures of $y_i(t_j)$ at commensurate times. The objective is to find a parsimonious model $G(z)$ with all of its poles in $\mathbb{D}_\rho$ that is compatible with the collected (noisy-patients) data and a priori information.

4.3 Previous Results

4.3.1 Set of Atoms

In this section, we give set of atoms in [2, 1] that is suitable for solving the identification problem addressed in this dissertation.

In general, every strictly proper transfer function with poles in $\mathbb{D}_\rho$ can be written as a linear combination of first order strictly proper transfer functions with poles chosen from $\mathbb{D}_\rho$. Following on this fact, the atoms used in [61] are defined as

$$\left\{ \Psi_p(z) = \frac{1 - |p|^2}{z - p} : p \in \mathbb{D}_\rho \right\},$$

(4.2)

where $p$ is the pole of this first order transfer function. The numerator is chosen so that the Hankel norm of each atom is 1. With this choice of atoms, the coefficients of the linear combination turn out to be complex numbers, and the optimization problem forces the solution to have real valued outputs.

These two requirements can be eliminated by choosing a different set of atoms. Then the following atomic set is defined in [2, 1], where $\mathcal{A}$ is defined that takes in a set of complex poles and produces a set of transfer functions as

$$\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3 \cup \mathcal{A}_4,$$
where each sub-set is defined as follows:

\[
\mathcal{A}_1 = \left\{ \pm \gamma^1_p \left( \frac{1}{z-p} + \frac{1}{z-\bar{p}} \right) : p \in \mathbb{D}_\rho \right\}
\]

\[
\mathcal{A}_2 = \left\{ \pm \gamma^2_p \left( -\frac{j}{z-p} + \frac{j}{z-\bar{p}} \right) : p \in \mathbb{D}_\rho \right\}
\]

\[
\mathcal{A}_3 = \left\{ \pm \frac{\gamma_p}{z-p} : p \in \mathbb{D}_\rho, p \text{ real} \right\}
\]

\[
\mathcal{A}_4 = \{+1, -1\},
\]

with numerator chosen as \(\gamma_p\)'s that are scaling factors defined in the next section and [1].

### 4.3.2 Deciding on the Scaling Factor \(\gamma_p\) [1]

Scaling factor is important in the minimization problem to weight each pole in \(\mathbb{D}_\rho\). Since this is a zero norm minimization problem that is approximated by \(\ell_1\) norm, then the scaling has a significant effect on weighting the poles and on the sparsity.

It is shown in [61] that the McMillan degree of an LTI system is equal to the Hankel operator of the system. Then, they chose the numerator of the atom so that the Hankel norm of each atom is 1. On the other hand, the response of our atoms in (4.3) is a finite number of data points for strictly stable plants (and unbounded otherwise), so the same idea in [61] might be problematic from a numerical point of view. For example, if the systems has poles close to the unit circle, then the Hankel nuclear norm becomes very large and approaches infinity as at least one pole approaches the unit circle. This might be a problem for our algorithm when the experiment horizon is short and the system is lightly damped.

Therefore, a scaling for our Atoms can be proposed both on the pole and on the data horizon. Given a horizon length \(N\), we normalize the nuclear norm of the finite dimensional square (or near square) Hankel matrix constructed by the first \(N\) Markov parameters of our Atoms. A complete description and calculation steps can be found in [1] for the Atoms in set \(\mathcal{A}\). We define the closed form of \(\gamma_p\)'s as follows:
\[ a = \frac{1 - p^{2N}}{1 - p^2} \quad \text{and} \quad c = \frac{1 - |p|^{2N}}{1 - |p|^2} \]

and

\[ \text{prod} = 2 \left( |a|^2 - c^2 \left[ \frac{\Re(a) - c - \Re(p^2(p^*)^{2N}a) + |p|^2|p|^{2N}c)}{1 - |p|^2} \right] \right) \]

then

\[ \gamma_p^1 = \frac{1}{\sqrt{2(\Re(a^2) + c^2)} + 2\sqrt{\text{prod}}} \]
\[ \gamma_p^2 = \frac{1}{\sqrt{2(c^2 - \Re(a^2))} + 2\sqrt{\text{prod}}} \]
\[ \gamma_p = \frac{1 - p^2}{1 - p^{2N}} \]

We also use these scaling factors for multi-run’s Atoms.

### 4.4 Set of Atoms for Multi-Runs

The set of Atoms above in (4.3) is defined assuming that the system started with zero initial conditions. Given that, in the problem addressed in this Chapter, one needs to take into account that data was generated by a system started at different initial conditions. Then these sets need to be modified. The Atoms provided in this Chapter are just an extension of the Atoms in section 4.3.1 and contain two components: i) response to initial conditions of several different runs and ii) impulse response of a LTI system. Here, the impulse response is just shifted up one step time to form the initial condition response. We will use the following notations for simplicity.

**Definition 6.** Given an Atom \( a \in \mathcal{A} \), let’s define \( \Upsilon^N\{a(z)\} \) being \( N \)-length impulse response vector of a discrete transfer function \( a(z) \in \mathcal{A} \).

We now introduce some additional notation aimed at simplifying the exposition
to follow. Consider first $p \in \mathbb{D}_\rho$ with imaginary part is equal to zero. Define the set of “Atomic” responses to initial conditions form the set $\mathcal{A}_3$ as

$$\text{g}^{ic(R)}(p) = \left\{ \gamma^N \{za(z)\} = \gamma_p \frac{\alpha z}{z - p} \right\}$$

where $\alpha \in \{1, -1\}$ and $a \in \mathcal{A}_3$

and the set of “Atomic” impulse responses again from the set $\mathcal{A}_3$ as:

$$\text{g}^R(p) = \left\{ \pm \frac{\gamma_p}{z - p} : p \in \mathbb{D}_\rho, p \text{ real} \right\}$$

Now consider the case of $p \in \mathbb{D}_\rho$ with imaginary part different from zero. In this case the “Atomic” responses to initial conditions from the set $\mathcal{A}_1$ and $\mathcal{A}_2$ are

$$\text{g}^{ic(C)}(p) = \left\{ \gamma^N \{za(z)\} \right\}$$

where $p^*$ denotes the complex conjugate of the pole $p$. Also, define the sets of “Atomic” impulse responses associated with $p$ from set $\mathcal{A}_1$ and $\mathcal{A}_2$ as:

$$\text{g}_1^C(p) = \left\{ \pm \gamma_1 \left( \frac{1}{z - p} + \frac{1}{z - p^*} \right) : p \in \mathbb{D}_\rho \right\}$$

$$\text{g}_2^C(p) = \left\{ \pm \gamma_2 \left( \frac{-j}{z - p} + \frac{j}{z - p^*} \right) : p \in \mathbb{D}_\rho \right\}$$

Now again the set in equation (4.3) can be extended for multi run with unknown initial conditions. To define the Atoms, recall that one has $\mathcal{P}$ different runs of the system, each with different initial conditions. Hence, an Atom must contain elements corresponding to the initial conditions of each of the runs and, also, to the impulse response of the LTI system. Therefore, define

$$\tilde{\mathcal{A}}_1 = \{(g_1^{ic}, g_2^{ic}, \ldots, g_\mathcal{P}^{ic}, g) : g_i^{ic} \in \text{g}^{ic(C)}(p) \text{ and } g \in \text{g}_1^C(p) \text{ for some } p \in \mathbb{D}_\rho \text{ with } \mathcal{I}(p) \neq 0\}$$

$$\tilde{\mathcal{A}}_2 = \{(g_1^{ic}, g_2^{ic}, \ldots, g_\mathcal{P}^{ic}, g) : g_i^{ic} \in \text{g}^{ic(C)}(p) \text{ and } g \in \text{g}_2^C(p) \text{ for some } p \in \mathbb{D}_\rho \text{ with } \mathcal{I}(p) \neq 0\}$$

$$\tilde{\mathcal{A}}_3 = \{(0, 0, \ldots, 0, \pm 1)\}$$
and
\[ \tilde{\mathcal{A}}_4 = \{(g_1^{ic}, g_2^{ic}, \ldots, g_P^{ic}, g): g_i^{ic} \in g^{ic(R)}(p) \text{ and } g \in g_2^R(p) \text{ for some } p \in \mathbb{D}_\rho \text{ with } \mathcal{I}(p) = 0 \} \]

The set of Atoms then can be defined as
\[ \tilde{\mathcal{A}} = \left\{ \pm \mathcal{Y}^N\{za(z)\}^T \pm \mathcal{Y}^N\{za(z)\}^T \pm \cdots \pm \mathcal{Y}^N\{a(z)\}^T, a(z) \in \mathcal{A} \right\} \]
\[ = \tilde{\mathcal{A}}_1 \cup \tilde{\mathcal{A}}_2 \cup \tilde{\mathcal{A}}_3 \cup \tilde{\mathcal{A}}_4. \] (4.4)

At this point, one should note that, given any finite dimensional linear time invariant system with non-repeated poles in \( \mathbb{D}_\rho \), its response can be put as a linear combination of a finite number of Atoms in \( \tilde{\mathcal{A}} \).

Now, take \( \mathcal{P} \) responses \( y_i(t) \) of a system to inputs \( u_i(t) \), \( i \in \{1, 2, \ldots, \mathcal{P}\} \) and to initial conditions. Then, there exist real constants \( c_a \) such that, for all \( i \in \{1, 2, \ldots, \mathcal{P}\} \)
\[ y_{i,t} = \sum_{a \in \mathcal{A}} c_a \left\{ \mathcal{Z}^{-1}\{\mathcal{Y}^N\{za(z)\}^T\}_t + \left[ \mathcal{Z}^{-1}\{\mathcal{Y}^N\{a(z)\}^T\} * u_i \right]_t \right\} \]
where \( \mathcal{Z}^{-1} \) denotes the inverse of the unilateral z-transform.

### 4.5 System Identification Via Atomic Norm Minimization

In this section, a convex relaxation of the problem of finding the lowest order model is provided. Indeed, this problem is known to be a hard non-convex one. Since the data is collected in the time domain, in order to develop efficient algorithms one needs to “convert” Atoms into the time domain.

One of the main strengths of the approach proposed in this dissertation is its ability to extract a model from fragmented data. More precisely, in this section, we show how one can extract a parsimonious LTI model from several runs (covers single run) of the same system with different initial conditions and inputs. Moreover, we show how to address the case of non-uniform sampling/missing data. Recall
that one has measurements of $P$ responses of the system $y_{i,t}$ to $u_{i,t}$ at times

$$0 \leq t_{i,1} < t_{i,2} < \cdots < t_{i,m_i} = N; m_i \leq N + 1.$$ 

### 4.5.1 Multi-Runs

More precisely, given $P$ inputs $u_i$, assume that, for each $i = 1, 2, \ldots, P$, the following data is available

$$y_{i,t} = g_{i}^{ic} + (g * u_i)_t + \eta_{i,t}, \quad t = 0, 1, \ldots, N_i$$

where $N_i$ is the number of samples collected and $g_{i}^{ic}$ is the initial condition response for the $i$-th run. As before, $g$ is the impulse response of the system. For simplicity of exposition, we assume from now on that all the runs have same length $N$ and we denote the vector containing the measurements of the $i$-th run as

$$y_i = [y_{i,0} \ y_{i,1} \ \cdots \ y_{i,N_i}]^T,$$

and the concatenations of all the responses as

$$y = [y_1^T \ y_2^T \ \cdots \ y_P^T]^T.$$ 

We have seen that, for any finite dimensional system with non-repeated poles \(^4\), the impulse response can be defined as:

$$g(a) = \sum_{a(z) \in A} c_a \mathcal{T}^N\{a(z)\}, \quad c_a \in \mathbb{R}.$$ 

A similar result holds for the initial condition responses; i.e.,

$$g_{i}^{ic}(a) = \sum_{a(z) \in A} c_{i,a}^{ic} \mathcal{T}^N\{za(z)\}, \quad c_{i,a}^{ic} \in \mathbb{R}, \quad i = 1, 2, \ldots, P.$$ 

For simplicity, we use $g_{i}^{ic} \doteq g_{i}^{ic}(a)$ and $g \doteq g(a)$. The objective is to find the most parsimonious model that is compatible with the data; i.e., we want to

\(^4\)A system with repeated poles can be approximated by a system with non-repeated poles by recurring to arbitrarily small perturbations [62].
explain the data using the system with the least number of poles. In terms of the coefficients $c_a$ and $c_{i,a}^c$, this objective corresponds to maximizing the “right” block sparsity measure. More precisely, we define vectors as

$$c_a = [c_{1,a}^c, c_{2,a}^c, \ldots, c_{P,a}^c, c_a]^T.$$  

Hence, the search for the simplest model that explains the data is equivalent to having as many of these vectors equal to zero as possible. In terms of Atomic norms, this leads to the following problem: Define the concatenation of the initial condition responses and the impulse response in time domain as

$$\tilde{g} = [g_1^{icT} g_2^{icT} \cdots g_P^{icT} g^T]^T \quad (4.5)$$

and recall that

$$\tilde{g}(a) = \sum_{a \in A} [c_{1,a}^c \mathcal{Y}^N\{za(z)\}^T c_{2,a}^c \mathcal{Y}^N\{za(z)\}^T \cdots c_{P,a}^c \mathcal{Y}^N\{za(z)\}^T c_a \mathcal{Y}^N\{a(z)\}^T]^T. \quad (4.6)$$

For simplicity, again we use $\tilde{g} = \tilde{g}(a)$. Then, in this case, the Atomic norm of interest is

$$\|\tilde{g}\|_{\tilde{A}} = \inf \sum \|c_a\|_{\infty} \quad \text{subject to } (4.6) \quad (4.7)$$

Define also

$$T = \begin{bmatrix} I & 0 & \cdots & 0 & \mathcal{T}_{u_1} \\ 0 & I & \cdots & 0 & \mathcal{T}_{u_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & \mathcal{T}_{u_p} \end{bmatrix}$$

where the Toeplitz matrix $\mathcal{T}_{u_i}$ is:
\[ T_{u_i}^n = \begin{bmatrix} u_{i,0} & 0 & \ldots & 0 \\ u_{i,1} & u_{i,0} & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ u_{i,n-1} & u_{i,n-2} & \ldots & u_{i,0} \end{bmatrix} \]  

(4.8)

and

\[ y = [y_1^T, y_2^T, \ldots, y_P^T]^T. \]

Given this, the identification problem in this Chapter can be loosely stated as follows: given measurements \( y \), determine \( c_a \) and \( c_{i,a} \) above that i) approximates \( y \) with a small error and ii) use only Atoms associated with a small number of poles. In order to solve this problem, Atomic norm \( \| \cdot \|_{\tilde{A}} \) in (4.7) is defined as the norm with unit ball \( B \subset \mathbb{R}^{(P+1)N} \)

\[ B \doteq \text{conv}\{\tilde{g} : a \in \tilde{A}\}. \]

The norm above is used in this chapter as a convex surrogate to enforce block sparsity, where each of the blocks is associated with a specific pole \( p \in \mathbb{D}_\rho \). The problem to be solved then becomes

\[
\begin{align*}
\min_{\tilde{g}} & \quad \|T\tilde{g} - y\|_2^2 \\
\text{subject to} & \quad \|\tilde{g}\|_{\tilde{A}} \leq \tau
\end{align*}
\]

(4.9)

Now, we need to identify the Atoms associated with the Atomic norm above. We can utilize the set of Atoms in (4.4) to approximate the impulse response and associated initial condition responses of the most parsimonious model that is compatible with the data collected.

### 4.5.2 Nonuniform Sampling/Incomplete Data

Next, consider the identification problem described in the previous section but where one has nonuniform sampling/missing samples. Without loss of generality, we assume that the sampling instants \( t_{i,j} \) are integers. Again, the objective for
this multi-run problem is to find a parsimonious model $\bar{G}(z)$ with all of its poles in $\mathbb{D}_\rho$ that is compatible with the collected (noisy) data. This problem can be formulated in a similar way to the one discussed in the previous section. We now elaborate on this. Let

$$y = [y_1(t_{1,1}) \ldots y_1(t_{1,m_1}) \ y_2(t_{2,1}) \ldots y_2(t_{2,m_2}) \ldots y_P(t_{P,1}) \ldots y_P(t_{P,m_P})]^T.$$  

be the vector containing all the available measurements. Define also measurement matrix, which “extracts” the output at the measured times from the overall response

$$M_{\text{meas}} = \begin{bmatrix}
  M^1 & 0 & \cdots & 0 \\
  0 & M^2 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & \cdots & 0 & M^P 
\end{bmatrix}$$

where, for $i = 1, 2, \ldots, P$, $M^i$ is a $m_i \times n_i$ whose $(j, k)$ entry is given by

$$M^i_{jk} = \begin{cases} 
1 & \text{if } k = t_{i,j}; \\
0 & \text{otherwise}.
\end{cases}$$

Then, and recalling that the matrix $T$ defined in the previous section contains all the information on the inputs applied to the system, we can modify the problem described in the previous section to address the case where one has nonuniform sampling/missing data as

$$\arg\min_{\bar{g}} \ |M_{\text{meas}}(T\bar{g} - y)|^2_2$$

subject to $\|\bar{g}\|_A \leq \tau$.  

(4.10)

At this point, we should note that this approach provides the means to determine a system that interpolates in between “fragmented data,” in the presence of both nonzero initial conditions and noise. This problem, to the best of our knowledge, has not been successfully addressed in the literature.
4.6 Identification Algorithm

4.6.1 Greedy Algorithm

The optimization problem (4.9) and (4.10) are special cases of the class of problems studied in [63] where the authors consider problems involving Atomic norms of the form

\[
\arg\min_{\tilde{g}} f(\tilde{g}) \quad \text{subject to } \|\tilde{g}\|_A \leq \tau
\]

where \(f(\tilde{g})\) is a convex and smooth function. A direct adaptation of the algorithm provided in [63] leads to the following greedy optimization algorithm:

**Algorithm 2** A general greedy Frank-Wolfe (1956) [64] algorithm to minimize a convex function \(f\) over the \(\tau\)-scaled Atomic norm ball

1: \(\tilde{g}_{initial} \leftarrow \tau \tilde{g}(a_0)\) for arbitrary \(a_0 \in \mathcal{A}\) \(\triangleright\) Initialization
2: for \(t = 0,1,2,3,...\) do
3: \(a_t \leftarrow \arg\min_{a \in \mathcal{A}} \langle \nabla f(\tilde{g}_t), \tilde{g}(a) \rangle\)
4: \(\alpha_t \leftarrow \arg\min_{\alpha \in [0,1]} f(\tilde{g}_t + \alpha[\tau \tilde{g}(a_t) - \tilde{g}_t])\)
5: \(\tilde{g}_{t+1} \leftarrow \tilde{g}_t + \alpha_t[\tau \tilde{g}(a_t) - \tilde{g}_t]\)
6: end for

Since function \(f(\tilde{g})\) in the optimization problem (4.11) is convex and continuously differentiable at domain \(\mathcal{A}\), then one of the earliest and simplest iterative algorithm to such optimization problem is developed by the Frank-Wolfe in [64] in algorithm 2. On the other hand, step 3 is computationally hard step in this algorithm. Therefore, the work in [2] adds a randomized step to this iterative algorithm 3 to have a computationally more efficient identification algorithm.
4.6.2 Modification of Randomized Algorithm in [2]

Algorithm 2 is the well-known Frank-Wolfe Algorithm and its proof of convergence can be found in [63, 64, 65]. In the algorithm above the “hard step” is step 3. A slight modification of the reasoning about a similar algorithm in [2, 1] leads to the conclusion that the optimization problem in step 3 is a polynomial optimization problem in two variables which can be (approximately) solved using existing tools for multi-run problem in (4.10) and (4.9).

However, even for such a small number of variables, solving this polynomial optimization problem can be time consuming. Hence, we propose the following modification of the randomized algorithm in [2, 1].

In this algorithm 3, a random number of Atoms is generated at each step. At each step \( k \) and given \( \epsilon_k \), it is possible to calculate number of the randomly generated poles over the set \( \mathbb{D}_\rho \), in order to have at least one pole within \( \epsilon_k \) neighborhood of the optimum pole with probability \( p \). This number is called \( N(\epsilon_k, p) \) [1]. Again all calculation of this number \( N(\epsilon_k, p) \) and convergence proof of Algorithm 3 is given in [1]. In the algorithm 3 authors in [2, 1] have replaced the search for the Atom that provides the steepest descent (step 3 in Algorithm 2) by a random search for an Atom that provides a descent direction (steps 3–5 in Algorithm 3).

\textbf{Algorithm 3} Modified Randomized Frank-Wolfe algorithm [2] to minimize a convex function \( f \) over the \( \tau \)-scaled Atomic norm ball

\begin{center}
\begin{tabular}{ll}
1: & \( \tilde{\mathbf{g}}_0 \leftarrow \tau \mathcal{T} \{ a_0(z) \} \) for arbitrary \( a_0(z) \in \mathcal{A} \) & \( \triangleright \text{ Init.} \) \\
2: for & \( k = 0,1,2,3,\ldots, k_{max} \) do \\
3: & Pick \( N(\epsilon_k, p) \) poles uniformly distributed over \( \mathbb{D}_\rho \), \\
4: & \( \tilde{\mathbf{g}}(a_k) \leftarrow \arg\min_{a \in \mathcal{A}(\mathcal{N}(\epsilon_k, p))} \langle \nabla f(\tilde{\mathbf{g}}_k), \tilde{\mathbf{g}}_k(a) \rangle \) \\
5: & \( \alpha_k \leftarrow \arg\min_{\alpha \in [0,1]} f(\tilde{\mathbf{g}}_k + \alpha[\tau \tilde{\mathbf{g}}(a_k) - \tilde{\mathbf{g}}_k]) \) \\
6: & \( \tilde{\mathbf{g}}_{k+1} \leftarrow \tilde{\mathbf{g}}_k + \alpha_k[\tau \tilde{\mathbf{g}}(a_k) - \tilde{\mathbf{g}}_k] \) \\
7: & end for
\end{tabular}
\end{center}

First, let’s look at the complexity of each step. Step 4 in Algorithm 3 can be recast as an optimization of the form

\[
\min_{l} \ l^T v
\]
subject to \( l_i \in \{-1, 1\} \)

which has an optimum

\[ l_i = -\text{sign}(v_i). \]

In other words, step 4 in Algorithm 3 admits a closed form solution. One should note at this point that, given the fact that the set of Atoms is centrally symmetric, steps 4-5 will find a descent direction with probability of one [2, 1].

As for the optimization problem in step 5, this is a second order polynomial minimization problem in \( \alpha \in [0, 1] \) where optimal \( \alpha^* \) has the following closed form:

\[ \alpha^* = \max(0, \min(\alpha_r, 1)) \]

where

\[ \alpha_r = -\frac{s^T q}{q^T q}; \]

\[ s = T\tilde{g}_k - y \quad \text{and} \quad q = T[r\tilde{g}(a_k) - \tilde{g}_k]. \]

Therefore, it is also efficiently computable. As far as convergence is concerned, one has the following results in [1].

**Lemma 7.** [1] Algorithm 3 has convergence in expectation and almost surely.

Proof of the Lemma can be found in [1]. It is also proven in [1, 2] that the cost function in Algorithm 3 decreases with probability 1 at each step, although the rate of descent is typically not optimal. Therefore, Algorithm 3 has a linear rate of convergence \( O(1/k) \) in expected value. More precisely, one states the following

**Theorem 8.** [1] There exists constants \( C_1, C_2, L \) and a sequence \( \gamma_k = 0.25C_2/(k + L) \) such that the iterates \( \bar{g}_{k+1} \) of Algorithm 3 satisfies, for \( k \geq 0 \):

\[ E[f(\bar{g}_{k+1}) - f(\bar{g}^*)] \leq \frac{C_1}{k + L} + \frac{C_2}{k + L + 1} \]

where \( \bar{g}^* \in \|\bar{g}\|_A \leq \tau \) is an optimal solution to the Problems (4.9) and (4.10).
4.6.3 Implementation-Heuristic

In this section, the heuristics for the problems (4.9) and (4.10) are given. We start with a heuristic for multi-run case, which obviously covers single run.

When applying Algorithm 3 to problems (4.9) or (4.10), especially for large volumes of data, we are faced with two main challenges:

i) Large size of the gradients used to choose the best Atom among the sampled ones.

ii) Differences in the magnitude of impulse and initial condition responses.

In this section, we propose a modified version of randomized Algorithm 3 that addresses the points above. This is done according to following steps: i) First, compute the gradient with respect only to the impulse response, and choose the best Atom as if the initial condition response is fixed. Then we use only the Atoms with the same pole to update the initial conditions responses (our heuristic forces to have the “same” poles for the impulse response and initial condition responses of the system). ii) Use difference scalings for impulse and initial condition responses, denoted in the algorithm by \( \tau \) and \( \tau_{ic} \) respectively. Different \( \tau \) and \( \tau_{ic} \) put a bound on the gain of the impulse response and initial condition responses. Different \( \tau \)'s for the impulse response and initial condition responses are used to increase the degree of freedom if the magnitude of impulse response and initial condition responses are very different. Moreover, this is a very efficient heuristic in terms of calculation since we only calculate part of the gradient which corresponds \( g \) in step 4 or in \( \tilde{g} \) equation (4.5) and calculate only part of gradient which corresponds \( g_{ic} \) in step 8 in algorithm 4. In order to clarify the heuristic, we just remind the initial condition responses and impulse response vectors separately as:

\[
g(a) = \sum_{a \in A} c_a \mathcal{T}^N\{a(z)\}, \quad c_a \in \mathbb{R}
\]

\[
g_{ic}(a) = \sum_{a \in A} c^i_a \mathcal{T}^N\{za(z)\}, \quad c^i_a \in \mathbb{R}
\]

\[
\tilde{g} = [g_{ic}^T \ g_{ic}^T \cdots \ g_{ic}^T \ g^T]^T
\]
\|\hat{g}\|_A = \sum_{i=1}^{P} \left\| \begin{bmatrix} c_{1,i} \cr c_{2,i} \cr \vdots \cr c_{P,i} \cr c_a \end{bmatrix} \right\|_\infty \quad \text{subject to } \hat{g}

where \( i = 1, 2, \ldots, P \) in the algorithm.

The description of this algorithm is provided in Algorithm 4. The notation is the same used in Section 4.4 and 4.6.

**Algorithm 4** Randomized algorithm to solve LTI identification problem with multi runs

1. \( \hat{g} = [g_{i,0}^{ic}, \ldots, g_{P,0}^{ic}, g_0] \leftarrow 0 \) \quad \triangleright \text{Init.}
2. \textbf{for} \( k = 0, 1, 2, 3, \ldots, k_{\text{max}} \) \textbf{do}
3. \quad \text{Pick } N(\epsilon_k, p) \text{ poles uniformly random over } D_{\rho}
4. \quad a_k \leftarrow \arg\min_{a(z) \in A\{N(\epsilon_k, p)\}} \langle \nabla f(g_k), g(a) \rangle
5. \quad \text{Store optimum pole that forms } a_k \rightarrow p^*
6. \quad g_{k+1} \leftarrow g_k + \alpha[\tau a_k - g_k] \text{ such that } \alpha \leftarrow \arg\min_{\alpha \in [0, 1]} f(g_{k+1})
7. \quad \textbf{for} \ i = 1, \ldots, P \ \textbf{do}
8. \quad a^*(z) \leftarrow \arg\min_{a(z) \in A\{p^*\}} \langle \nabla f(g_{i,k}^{ic}), g_{i,k}^{ic}(a) \rangle
9. \quad a_{i,k}^{ic} \leftarrow \Upsilon_N \{ za^*(z) \}
10. \quad g_{i,k+1}^{ic} \leftarrow g_{i,k}^{ic} + \alpha[\tau_{i,k} a_{i,k}^{ic} - g_{i,k}^{ic}] \text{ such that } \alpha \leftarrow \arg\min_{\alpha \in [0, 1]} f(g_{k+1})
11. \quad \textbf{end for}
12. \textbf{end for}

where \( f(\hat{g}) \) represents the objective function in (4.9) or (4.10). In order to compute the gradients, we first define the output estimation error for the \( i \)th run as \( e_i = M_{\text{meas}}^i (T_u g + g_{i,0}^{ic} - y_i) \). The gradients are given by:

\[
\nabla f(g) = \sum_{i=1}^{i=P} T_u^T (M_{\text{meas}}^i)^T e_i \quad \text{and} \quad \nabla f(g_{i,0}^{ic}) = (M_{\text{meas}}^i)^T e_i \quad (4.12)
\]

Note that by using the pole found in Step 5 in update steps for initial condition responses, we are favoring a solution where all the responses tend to share same poles, hence resulting in a row-sparse \( c \) in principle.
4.7 Simulation Results

4.7.1 Experiment of Single Run

We start with a synthetic data example, where the objective is to estimate the model in the presence of noise and missing data. The system used in this example is

\[ G(z) = \frac{-18.77z^3 + 21.13z^2 + 4.509z - 6.038}{z^4 - 0.04449z^3 + 0.09372z^2 + 0.08008z - 0.3015} \]

The input signal was generated using a uniform distribution on the interval [0, 1]. The initial conditions were also generated randomly with a uniform distribution on the interval [0, 4]. We then added Gaussian noise \( \eta \sim \mathcal{N}(0, 1.5) \) to the responses. Finally, 20\% of the data was removed at random.

The proposed algorithm was applied to the data collected with \( \tau = 76.6 \). Once the impulse response is identified then a realization of the system can be obtained from this impulse response via Ho’s algorithm (or via subspace methods) [20]. The plant model obtained is

\[ \hat{G}(z) = \frac{-10.36z^3 + 13.34z^2 + 5.519z - 7.309}{z^4 + 0.0239z^3 + 0.0914z^2 + 0.1541z - 0.303} \]

The fact that the identified model is not exactly equal to the true one is a consequence of the fact that the data collected has a substantial amount of noise. Moreover, as mentioned above, the algorithm only has access to 80\% of the data. Note, however, that the obtained system model fits the data quite well, as shown in the plots provided below. In Figure 4.1 we show both the real and the estimated outputs. As one can see, they are very similar. In Figure 4.2, we compare the estimated and the true impulse responses and, in Figure 4.3, the true and estimated response to initial conditions. As the presented results show, the algorithm proposed performs well when identifying the system from noisy/fragmented data.
Figure 4.1: Real and Estimated Single Run Output (SISO)

It should be noted that the outcome of the algorithm is highly dependent on the choice of $\tau$. If $\tau$ is chosen to be “too large,” there is no emphasis on sparsity and one is modeling not only the system but also the noise.

On the other hand, if we choose a value of $\tau$ which is “too small,” the true model is excluded from the feasible set and, obviously, one is not able to identify a system that provides a good explanation of the data. In practice, one chooses the value of $\tau$ iteratively until an acceptable compromise between model complexity and modeling error is reached.
Figure 4.2: Impulse Response of Single Run (SISO)

Figure 4.3: Initial Condition Response of Single Run (SISO)
### 4.7.2 Experiment-1 Multi Run

We illustrate the performance of the proposed algorithm in a non-trivial scenario with two different experiment: multiple runs of an LTI plant with different initial conditions and random missing data. A random 4\textsuperscript{th} order system was generated:

\[
g(z) = \frac{0.113z^2 - 0.281z - 0.176}{z^4 - 0.887z^3 + 0.515z^2 + 0.233z + 0.032} \tag{4.13}
\]

The system was excited with PRBS input of length 250 for each run, starting at a random initial condition. The output of each run was corrupted by 10\% additive white noise, and at least 60\% of the measurements were kept at each run. Four different runs were conducted with horizons as \(y_1 \sim 224\), \(y_2 \sim 195\), \(y_3 \sim 245\), \(y_4 \sim 159\). The ground truth \(\tau = 1.43\) was supplied to the algorithm. A handpicked \(\tau_{ic} = 2.87\) was used to bound all of the initial condition responses. The results are given in Figs. 4.4, 4.5 and 4.6. Table 4.1 below shows the length of the data record and initial conditions used for identification in each of the runs.

<table>
<thead>
<tr>
<th>Run</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of y</td>
<td>224</td>
<td>195</td>
<td>245</td>
<td>159</td>
</tr>
<tr>
<td>(y_0)</td>
<td>-2.244</td>
<td>-1.38</td>
<td>0.8369</td>
<td>-1.726</td>
</tr>
<tr>
<td>(y_0)</td>
<td>-1.386</td>
<td>-1.25</td>
<td>0.055</td>
<td>-0.646</td>
</tr>
<tr>
<td>(y_0)</td>
<td>2.469</td>
<td>1.978</td>
<td>2.07</td>
<td>-0.5452</td>
</tr>
<tr>
<td>(y_0)</td>
<td>0.0764</td>
<td>-0.0337</td>
<td>-2.259</td>
<td>2.3685</td>
</tr>
</tbody>
</table>

**Table 4.1: Length of Output and Initial Conditions Experiment-1 (SISO)**

Note that, if the system has no initial condition and it is SISO system, then some heuristic is developed on adjusting \(\tau\) iteratively in [1]. If one has information about the noise characteristics, one can adjust \(\tau\) to find the lowest order system that provides a response whose difference from the data satisfies the characteristics of the noise process. As the results presented suggest, with a suitable \(\tau\), the algorithm proposed performs well in identifying the actual poles of the system from noisy data.
Figure 4.4: Output of Each Run Experiment-1 (SISO)

Figure 4.5: Impulse Responses of Multi Run Experiment-1 (SISO)

4.7.3 Experiment-2 Multi Run

Again multiple runs of an LTI plant with different initial conditions and random missing data with $2^{\text{th}}$ order system was generated. To this effect, we considered
the following system:

\[ g(z) = \frac{-0.0673z - 0.0303}{z^2 - 0.0673z - 0.0303} \]

Generated data is obtained by exciting the system with PRBS input of length 250 for each run, starting from random initial conditions. The noise is uniformly distributed \( \eta \sim 0.1 \max(y) \ast \mathcal{U}(-0.5, 0.5) \) which is assumed to be 10% noise. \( \mathcal{U}(-0.5, 0.5) \) generates values from the uniform distribution on the interval \([-0.5, 0.5]\). Finally, at most 30% of the output data of each run is removed randomly.

For the example presented here, we choose 2 different runs. Two different runs were conducted with horizons \( y_1 \sim 227, \ y_2 \sim 185 \). Table 4.2 below shows the length of the data record and initial conditions used for identification in each of the runs.
Table 4.2: Length of Output and Initial Conditions

<table>
<thead>
<tr>
<th></th>
<th>Run₁</th>
<th>Run₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of (y)</td>
<td>227</td>
<td>185</td>
</tr>
<tr>
<td>I.C (y_0)</td>
<td>-1.854</td>
<td>2.059</td>
</tr>
<tr>
<td>(y_0^2)</td>
<td>-0.158</td>
<td>0.9165</td>
</tr>
</tbody>
</table>

The ground truth \(\tau = 0.13\) was supplied to the algorithm. A handpicked \(\tau_{ic} = 3\tau = 0.3928\) was used to bound all the initial condition responses. Identification results are presented by several plots. Figure 4.7 shows the measurement and estimated output and 4.9 shows the real and estimated initial condition response of each run. As the results presented suggest, with a suitable \(\tau\), the algorithm proposed performs well in identifying the impulse response of the system from noisy and incomplete data.

Figure 4.7: Output of Each Run Experiment-2 (SISO)
Figure 4.8: Impulse Responses of Multi Run Experiment-2 (SISO)

Figure 4.9: Initial Condition response of Each Run Experiment-2 (SISO)
4.8 Concluding Remarks

First of all, in this Chapter we consider the problem of identifying parsimonious model from noisy time domain data, in the presence of unknown initial conditions. The main idea is to enforce sparsity by recasting the problem into an optimal approximation form subject to Atomic norm constraints. As shown in this Chapter, this leads to a problem that can be solved very efficiently by combining a Frank-Wolfe type algorithm with randomization ideas, avoiding the need to performing singular value decompositions for system identification. Indeed, since this randomized algorithm only involves inner products, it scales linearly with the number of experimental data points, as opposed to interior point or augmented Lagrangian methods.

Then we also considered the problem of identifying a linear system from noisy and incomplete measurements produced by multiple runs of the system with unknown initial conditions. Our results show that we can denoise and estimate the impulse response of the system reliably and efficiently. This can then be used to identify the model of the system using well known methods. Other a priori information, if available, can be further incorporated at this stage, e.g. maximum system order or bounds on certain system norms etc. The proposed approach provides a systematic way of estimating a model from data sets such as the ones obtained in patient behavior studies.

On the other hand, extending this approach to the MIMO system is still not an easy task since if one works with transfer function, one ends up with a non-convex optimization problem. Additionally, the parsimonious model identification of an LTI, MIMO systems is still a challenging task depending on the nature of the experimental data and/or priors available. The main difficulty comes from the nature of the identification problem and available data. Motivated by these difficulties and identification of behavioral models where behavioral data has many missing points, in Chapter 5, we propose an algorithm that uses noisy and/or incomplete data from a MIMO system to determine a low order model that is consistent with existing a priori information and measurements. The approach, based on concepts of Atomic norms and subspace identification methods, provides an efficient way to use noisy and fragmented data together to determine an approximated, low order
model and interpolate the missing part of the measurements. This algorithm can also be extended for multi-run case of MIMO systems.
Chapter 5

Parsimonious Model Identification of MIMO Systems

In this Chapter, we provide results aimed at solving the following problem: Given *a priori* information on a Multi-Input/Multi-Output (MIMO) plant, namely constraints on the pole location, and scattered input/output data, find the lowest order model that is compatible with both the *a priori* assumptions and the collected data. By combining concepts from signal sparsification and subspace identification, algorithms are developed that can determine a low order model from data that is both corrupted by measurement noise and has missing measurements.

As it is previously stated for SISO identification in Chapter 4, the identification of a parsimonious model for a MIMO system is also a challenging task. Again, the main difficulty comes from the nature of the identification problem and the available data.

The identification algorithm for a SISO system and multiple run of the SISO system are previously presented in Chapter 4. In this Chapter, we will develop a parsimonious model identification algorithm for MIMO systems. Identification algorithm needs to estimate the parsimonious model while interpolating the missing parts of the data set. The formulation for multi-run case for a MIMO system is also given in this Chapter. Finally, the effectiveness of the proposed approach is demonstrated through a few academic examples.
5.1 Introduction and Motivation

The complete literature review of the system identification for this Chapter is already given in section 4.1.1. Therefore, here, we will just briefly review the available methods needed to develop our new identification algorithm.

In particular, system identification tools based on the concept of the atomic norm [61, 17, 2] have shown to be very efficient in the identification of low order models from data with “large” levels of noise while still being able to incorporate \textit{a priori} information on the location of the modes of the plant. Moreover, it has been shown in Chapter 4 that this approach can be applied to cases where one has scattered data and multiple runs of the same system with different initial conditions. However, application has been limited to single-input/single output systems. One could try to apply the available algorithms to the identification of multi-input/multi-output systems by identifying models for each of outputs one at a time, but the fact that one has noisy measurements can easily lead to high order models since there is no systematic way to enforce the same poles and the same initial conditions for all models obtained.

The work presented in this Chapter is integrated in this line of research. We present results on how to extend the atomic norm approach to the problem of identification multi-input/multi-output systems. The algorithms developed are able to incorporate \textit{a priori} information on the location of the modes of the plant and are applicable to the case where one has scattered data; i.e., they can handle missing data. Additionally, this algorithm can be extended for multi-run of MIMO systems. The effectiveness of the algorithm is demonstrated through application to a few academic examples.

5.2 Preliminaries

Assume, a linear system is given as:

\[ x(k + 1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k), \]

where \( A \in \mathbb{R}^{n \times n} \), \( B \in \mathbb{R}^{n \times m_i} \), \( C \in \mathbb{R}^{m_o \times n} \), and \( D \in \mathbb{R}^{m_o \times m_i} \). Then, let us define \( M_r \) that contains Markov parameters of the system and \( O_r \) that is an extended
observability matrix:

\[
M_r = \begin{bmatrix}
D & 0 & 0 & \ldots & 0 \\
CB & D & 0 & \ldots & 0 \\
CAB & CB & D & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
CA^{r-1}B & CA^{r-2}B & CA^{r-3}B & \ldots & D
\end{bmatrix},
\]

\[
O_r = \begin{bmatrix}
C \\
CA \\
CA^2 \\
\vdots \\
CA^{r-1}
\end{bmatrix}.
\]

Define \(X_r\) as:

\[
X_r \doteq \begin{bmatrix} x_0 & x_1 & x_2 & \ldots & x_{N-r} \end{bmatrix}.
\]

We use \(M := M_r, O := O_r, \) and \(X := X_r\) to simplify the notation throughout the Chapter. Then, given \(H_u\), the Hankel matrix of the inputs \(u(k) \in \mathbb{R}^{m_i}\), and \(H_y\), the Hankel matrix of the outputs \(y(k) \in \mathbb{R}^{m_o}\), one can rewrite equation (5.1) as [66]:

\[
H_y = OX + M H_u, \tag{5.2}
\]

where \(H_y(r)\) is a \(rm_o \times (N-r)\) Hankel matrix associated with any finite sequence \(\{y(k) \in \mathbb{R}^{m_o}, k = 0, 1, \ldots, N\}\):

\[
H_y(r) \doteq \begin{bmatrix} y(0) & y(1) & y(2) & \ldots & y(N-r) \\
y(1) & y(2) & y(3) & \ldots & y(N-r+1) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y(r) & y(r+1) & y(r+2) & \ldots & y(N) \end{bmatrix},
\]

and we use \(H_y := H_y(r)\) to simplify the notation. Now define an orthonormal matrix \(H_u^\perp\) where the columns of \(H_u^\perp\) span the nullspace of \(H_u\). It clearly satisfies \(H_u H_u^\perp = 0\) and can be calculated as:

\[
H_u^\perp = I - H_u^T (H_u H_u^T)^T H_u.
\]

Furthermore, if one multiplies the equation (5.2) on the right with a \(H_u^\perp\) by using the idea in subspace identification [66, 67], the forced response of the system can be eliminated in output Hankel \(H_y\). Therefore \(H_y H_u^\perp\) and \(OX H_u^\perp\) are formed by
the response of natural modes that is proven in section 5.3.

\[ \mathcal{H}_y \mathcal{H}_u^\perp = OX \mathcal{H}_u^\perp \]  \hspace{1cm} (5.3)

**Assumption 9.** Assume \( X \) is persistently excited by input and there is no rank cancellation in \( X \mathcal{H}_u^\perp \).

Assume that the rank of the extended observability matrix is \( n_o \) which is equal to the order of the system \( n_o = n \), if the linear system is observable. Let us denote the observable excited order by \( n_{oe} \). Under the persistent excitation Assumption 9, inputs excite all modes of the system, and thus the result is that \( n = n_o = n_{oe} \). Furthermore, note that the phenomenon of rank cancellation can occur if the inputs are outputs of the linear system or if the input Hankel matrix \( \mathcal{H}_u \) is not full row rank [67].

**Theorem 10.** [67] Assume noiseless input output data are available, then \( \mathcal{H}_y \mathcal{H}_u^\perp \) has rank \( n_{oe} = n \) provided that Assumption 9 holds.

### 5.3 Problem Statement

Given a MIMO, a discrete-time LTI system with unknown state space model in (5.1), and a finite collection of time domain noisy and incomplete output measurements \( y(k) \in \mathbb{R}^{m_o} \) and known input \( u(k) \in \mathbb{R}^{m_i} \) where \( k = 0, \ldots, N \), the problem considered in this Chapter is to identify the lowest order model that approximates the true plant and, at the same time, interpolates missing measurements. We start by stating one of the assumptions used in this Chapter.

**Assumption 11.** The system to be identified does not have repeated poles and all the poles are located in the set \( \mathbb{D}_p \).

One should note that the assumption above does not limit the applicability of the approach proposed in this Chapter since we can always approximate a system with repeated poles by a system with non-repeated poles by recurring to arbitrarily small perturbations [62].
Since we can only identify a model up to a transformation matrix, we define the parameters of the estimated model as follows:

$$\dot{x}(k+1) = \hat{A}x(k) + \hat{B}u(k), \quad y(k) = \hat{C}x(k) + \hat{D}u(k)$$  \hspace{1cm} (5.4)

where $\hat{A} = T^{-1}AT$, $\hat{B} = T^{-1}B$, $\hat{C} = CT$, $\hat{D} = D$ and $T$ is non-singular transformation matrix.

To motivate the approach presented in this Chapter, recall that the response of an LTI system can be decomposed as

$$y(k) = y_n(k) + y_u(k)$$

where $y_n$ is the natural response of the system which contains only the part of the response associated with the modes of the system and $y_u$ is the forced response whose poles of its Z-transform coincide with the poles of the Z-transform of the input $u$. Given this partition of the response of the system, we have the following result whose proof is provided in Appendix A.

**Lemma 12.** If Assumption 9 holds, then

$$\mathcal{H}_y \mathcal{H}_u^\perp = \mathcal{H}_{y_n} \mathcal{H}_u^\perp$$

and, hence, $\mathcal{H}_y \mathcal{H}_u^\perp$ is only a function of the modes of the system.

Hence, the result above together with the one in Lemma 12, we have

$$\text{rank}(\mathcal{H}_y \mathcal{H}_u^\perp) = \text{rank}(\mathcal{H}_{y_n} \mathcal{H}_u^\perp) = \text{rank}(\mathcal{O}X \mathcal{H}_u^\perp) = n$$  \hspace{1cm} (5.5)

if the system is observable. Again $n$ is the system order.

The results above lead to the following way of determining the order of the system and a realization of $A$ and $C$ matrices from input/output data. No proof is provided since it is an immediate consequence of previous results.

**Proposition 13.** Assume that Assumptions 9 and 11 hold. Moreover, assume that the modes of the system are in the set $\mathbb{D}_\rho$. Given input $u(k)$ and output $y(k)$ data,
Consider the following (infinite dimensional) optimization problem:

\[ n^* = \min \| \delta \|_0 \]  \hspace{1cm} \text{(5.6)}

s. t.

- \( \mathbf{c}^p \in \mathbb{R}^{m_o} \) for all \( p \in \mathbb{D}_\rho \)
- \( \| \mathbf{c}^p \|_\infty \leq \delta(p) \) for all \( p \in \mathbb{D}_\rho \)
- \( y_n(k) = \sum_{p \in \mathbb{D}_\rho} \mathbf{c}^p p^k \) for \( k = 0, \ldots, N \)
- \( \mathcal{H}_y \mathcal{H}_u^\perp = \mathcal{H}_{y_n} \mathcal{H}_u^\perp \)

where \( \| \delta \|_0 = \text{card}\{p \in \mathbb{D}_\rho : \delta(p) \neq 0 \} \). Then, \( n^* \) is the order of the system that generated the data and a realization of the \( \mathbf{A} \) and \( \mathbf{C} \) can be computed as follows: Let, \( \delta^* \) and \( \mathbf{c}^{p,*} \) be a solution of the optimization problem above and let \( p_i \in \mathbb{D}, \) \( i = 1, 2, \ldots, n^* \) be all the values of \( p \) satisfying \( \delta(p_i) \neq 0 \). Then, take

\[ A = \text{diag}(p_1, p_2, \ldots, p_{n^*}); \quad \mathbf{C} = [\mathbf{c}^{p_1,*} \mathbf{c}^{p_2,*} \cdots \mathbf{c}^{p_{n^*},*}] \]

With this result at hand, we can now describe our approach for parsimonious system identification from noisy data. It consists of two steps:

**Step 1:** Estimate the output vector \( \hat{\mathbf{C}} \) and system matrix \( \hat{\mathbf{A}} \) while enforcing to have least number poles by solving the following optimization problem:

\[ \{ \mathbf{c}^p, \delta \} = \text{argmin} \| \delta \|_0 \]  \hspace{1cm} \text{(5.7)}

s. t.

- \( \mathbf{c}^p \in \mathbb{R}^{m_o} \) for all \( p \in \mathbb{D}_\rho \)
- \( \| \mathbf{c}^p \|_\infty \leq \delta(p) \) for all \( p \in \mathbb{D}_\rho \)
- \( y_n^{\text{est}}(k) = \sum_{p \in \mathbb{D}_\rho} \mathbf{c}^p p^k \) for \( k = 0, \ldots, N \)
- \( \| \mathcal{H}_y \mathcal{H}_u^\perp - \mathcal{H}_{y_n^{\text{est}}} \mathcal{H}_u^\perp \|_F \leq \varepsilon \)

In (5.7), we minimize the number of poles used (order of the system) and included a Frobenius norm constraint that enforces fidelity to the data collected. The matrices \( \hat{\mathbf{C}} \) and \( \hat{\mathbf{A}} \) are constructed using the modes used in the estimated natural
response; i.e., if $\delta$ and $c^p$ are the solutions of the problem above and letting $p_i \in D$, $i = 1, 2, \ldots, n^*$ be all the values of $p$ satisfying $\delta(p_i) \neq 0$ we have

$$\hat{A} = \text{diag}(p_1, p_2, \ldots, p_{n^*}); \quad \hat{C} = \begin{bmatrix} c^{p_1} & c^{p_2} & \cdots & c^{p_{n^*}} \end{bmatrix}$$

**Step 2:** Once $\hat{C}$ and $\hat{A}$ are identified, then the input matrix $\hat{B}$, feed-forward matrix $\hat{D}$ and the initial state $\hat{x}_0$ are computed by solving

$$\{\hat{B}, \hat{D}, \hat{x}_0\} = \arg\min_{\hat{B} \in \mathbb{C}^{m_o \times m_i}, \hat{D} \in \mathbb{R}^{m_o \times m_i}, \hat{x}_0 \in \mathbb{C}^n} \sum_{k=1}^{N} \left( y(k) - \hat{C}\hat{A}^k\hat{x}_0 + \sum_{t=0}^{k-1} \hat{C}\hat{A}^{k-t}\hat{B}u(t) + \hat{D}u(k) \right)^2 \quad (5.8)$$

Although the setup above would produce the most parsimonious model that is compatible with the data, there are two reasons why it cannot be solved directly:

- Minimization of zero norm is an NP-hard problem [68].
- Number of the poles in $D_p$ is infinite.

Therefore in the remainder of this Chapter we provide tractable relaxations to the procedure above.

### 5.4 Convex Relaxation Using Atomic Norm

#### 5.4.1 Atomic Norm

As mentioned in the previous section, the problem of finding the system of lowest order that is compatible with the data and a priori assumptions requires finding an estimate of the natural response

$$y_n^{est}(k) = \sum_{p \in D_p} c^p p^k \quad (5.9)$$

that uses the least number of modes. This is a “hard problem” and in this Chapter we use the convex envelope of the sparsity measure; i.e., we use the “right” atomic norm.
For the problem discussed in this chapter, the atomic norm that promotes the sparsity measure of interest is

$$\|y_n^{\text{est}}\|_A = \min \left\{ \sum_{p \in \mathbb{D}_\rho} \|c^p\|_\infty : y_n^{\text{est}}(k) = \sum_{p \in \mathbb{D}_\rho} c^p(\alpha_p p^k) \text{ for } k = 0, \ldots, N \right\} \quad (5.10)$$

where the constant $\alpha_p$ is a regularization constant that aims at giving the “same weight” to all the terms involved in the sum. With the same reasoning in section 4.3.2, scaling factor $\alpha_p$ has a significant effect on weighting the poles. In section 4.3.2, the scaling factor is calculated based on the finite response of Atoms (Atomic set in section 4.3.1 has both second order and first order systems) to prevent numerical problems. Note that, in this Chapter, Atoms are finite impulse response of first order systems which poles in $\mathbb{D}_\rho$. As a result, in this Chapter, we use $\alpha_p = (1 - |p|^2)/(1 - |p|^{2N+2})$ which leads to a Hankel matrix associated with $\alpha_p p^k$ of unit nuclear norm.

### 5.4.2 Convex Relaxation

Given the atomic norm above, the convex relaxation of the problem of finding the most parsimonious model that is compatible with the collected data and the a\ priori assumptions is

$$c^p = \arg\min \|y_n^{\text{est}}\|_A \quad (5.11)$$

s. t.

$$y_n^{\text{est}}(k) = \sum_{p \in \mathbb{D}_\rho} c^p(\alpha_p p^k) \text{ for } k = 0, \ldots, N$$

$$\|\mathcal{H}_y \mathcal{H}_u^\perp - \mathcal{H}_{y_n^{\text{est}}}^\perp \mathcal{H}_u^\perp\|_F \leq \varepsilon$$

However, this is still an infinite dimensional problem; i.e., the number of variables is uncountably infinite. To (approximately) solve this, one can take two approaches: i) use a randomized version of the Frank-Wolfe algorithm [2] or ii) one can grid the set $\mathbb{D}_\rho$ and use a finite number of samples [61].

For simplicity of exposition, in this Chapter, we have chosen to use the second
approach; i.e. we start with a finite set \( \mathcal{F} \subset \mathbb{D}_\rho \) and solve the problem

\[
\{c^p, \delta\} = \text{argmin} \|\delta\|_1
\]

subject to

\[
\|c^p\|_\infty \leq \delta^p \text{ for all } p \in \mathcal{F}
\]

\[
y^\text{est}_n(k) = \sum_{p \in \mathcal{F}} c^p \alpha_p p^k \text{ for } k = 0, \ldots, N
\]

\[
\|\mathcal{H}_y \mathcal{H}^\perp_u - \mathcal{H}^\text{est}_n \mathcal{H}^\perp_u\|_F \leq \varepsilon
\]

Once, the optimization problem in (5.12) is solved, as before, the estimated matrix \( \hat{A} \) is a diagonal matrix whose elements are the values of \( p \) for which \( \delta^p \) is non-zero and the estimated matrix \( \hat{C} \) is a matrix whose columns are again, the \( c^p \) corresponding to non-zero \( \delta^p \).

Finally, the estimate the input matrix \( \hat{B} \), the feed-through matrix \( \hat{D} \), and initial state \( \hat{x}(0) \) by solving the equation in (5.8).

### 5.4.3 Extension to ID Problem with Missing Points

One of the most interesting characteristics of the approach proposed is that it can easily handle missing data, without introducing extra variables. The modification of the problem to be solved is quite straightforward.

Define a measurement matrix \( \Omega^{\text{meas}} \), which “extracts” the missing points in the equation (5.3). \( \Omega^{\text{meas}} \in \mathbb{R}^{rmo \times rm_0} \) is a diagonal matrix with 0 entries in the row corresponding to the rows of the matrix \( \mathcal{H}_y \) if it has any missing measurements and 1 otherwise.

\[
\Omega^{\text{meas}}_{i,i} = \begin{cases} 
1 & \text{if } i^{th} \text{ row of } \mathcal{H}_y \text{ has no missing data; } \\
0 & \text{if } i^{th} \text{ row of } \mathcal{H}_y \text{ has at least a missing data }
\end{cases}
\]

where \( i = 1, \ldots, rm_0 \). Then, we solve the problem:
\[ \{c^p, \delta\} = \arg\min \|\delta\|_1 \]  
subject to 
\[ \|c^p\|_\infty \leq \delta^p \text{ for all } p \in \mathcal{F} \]  
y_{n}^{\text{est}}(k) = \sum_{p \in \mathcal{F}} c^p \alpha_p p^k \text{ for } k = 0, \ldots, N 
\[ \|\Omega_{\text{meas}} \mathcal{H}_y \mathcal{H}_u^\perp - \Omega_{\text{meas}} \mathcal{H}_{y_{\text{est}}} \mathcal{H}_u^\perp \|_F \leq \varepsilon \]

Again, once we have the estimate of output vector \( \hat{C} \) and system matrix \( \hat{A} \), we can again solve the optimization problem (5.8) after the elimination of the missing parts in the output vector and estimate \( \hat{B} \), \( \hat{D} \), and \( \hat{x}_0 \).

### 5.4.4 Extension to Multi-Run

Assume one collects data from multi-run of the same system. More precisely, given \( \mathcal{P} \) inputs \( u_i(k) \in \mathbb{R}^{m_i} \), assume that, for each \( i = 1, 2, \ldots, \mathcal{P} \), the following data is available

\[ y_i(k) = y_{n_i}(k) + y_{u_i}(k) \]  

where \( y_{n_i} \) is the natural response of the system for \( i^{th} \) run which again has only the part of the response associated with the modes of the system and \( y_{u_i} \) is the forced response for \( i^{th} \) run. Further, without lost of generality, we assume that all the runs have same length \( N \) for each output. The approach can be easily extended to runs of different length. For the given input \( u_i(k) \in \mathbb{R}^{m_i} \) and output \( y_i(k) \in \mathbb{R}^{m_o} \), we can construct the following matrix for multi-run:

\[ \mathcal{H}_y = \text{blkdiag}(\mathcal{H}_{y_1}, \mathcal{H}_{y_2}, \ldots, \mathcal{H}_{y_\mathcal{P}}) \text{ and } \mathcal{H}_u^\perp = [\mathcal{H}_{u_1}^\perp, \mathcal{H}_{u_2}^\perp, \ldots, \mathcal{H}_{u_\mathcal{P}}^\perp]^T. \]

First of all, since the same system is run with different inputs, we need to identify a single matrix \( C \). On the other hand, each run has different natural responses with
different initial conditions\(^1\). Since the natural response associated with a given mode is a constant times a fixed exponential, the part of the natural responses of each of the runs associated with a given mode only differ by a multiplying constant. Since, in our approach, we assume all initial conditions of natural response for each run are 1, then the identified matrices \(c^p_i\) will be different but only up to multiplication by a scalar. Hence the matrix \([c^p_1, c^p_2, \ldots, c^p_P]\) has to have rank 1. We can then estimate a state matrix \(\hat{A}\) and output matrix \(\hat{C}\) by solving the following optimization problem (5.15). In this optimization problem, nuclear norm constraint on \(c^p_i\) forces rank 1 matrix \([c^p_1, c^p_2, \ldots, c^p_P]\). Nuclear norm constraint is employed as tightest convex approximation of the rank [69].

Thus we assume all initial conditions for each run are 1. Then this scalar term which comes from different initial conditions goes to the \(c^p_i\). It means each \(c^p_i\) are a scalar times of each other. Finally, we can estimate a state matrix

\[
\{c^p_i, \delta\} = \text{argmin} \|\delta\|_1 \quad (5.15)
\text{subject to}
\begin{align*}
\|\begin{bmatrix} c^p_1, c^p_2, \ldots, c^p_P \end{bmatrix}\|_* &\leq \delta^p \text{ for all } p \in F \\
y^\text{estn}_{ni}(k) &= \sum_{p \in F} c^p_i \alpha_p p^k \text{ for } k = 0, \ldots, N \\
\|\mathcal{H}_y \mathcal{H}_u^\perp - \mathcal{H}_y^\text{estn} \mathcal{H}_u^\perp\|_F &\leq \varepsilon
\end{align*}
\]

Once, one solves the optimization problem in (5.15), as demonstrated in the previous section, the estimated matrix \(\hat{A}\) is a diagonal matrix whose elements are the values of \(p\) for which \(\delta^p\) is non-zero. The estimated matrix \(\hat{C}\) is a matrix whose columns are again, one of the \(c^p_i\) corresponding to non-zero \(\delta^p\). We can use one of the \(c^p_i\) since the others are just scalar multiplication of this chosen one and the output of the each run will be fitted by solving the (5.8).

We can obviously extend this for the case of missing measurements. We only need to define a measurement matrix \(\Omega^{\text{meas}}\), which “extracts” the missing points in the equation (5.15) such that:

\(^1\)This is the value where the natural response starts at time 0. Note that this is different than initial condition of the system
\[
\{c^p_i, \delta\} = \text{argmin} \|\delta\|_1 \tag{5.16}
\]
subject to
\[
\|\{c^1_i, c^2_i, \ldots, c^p_i\}\|_* \leq \delta^p \text{ for all } p \in \mathcal{F}
\]
\[
y_{n_i}^{\text{est}}(k) = \sum_{p \in \mathcal{F}} c^p_i \alpha_p p^k \text{ for } k = 0, \ldots, N
\]
\[
\|\Omega_{\text{meas}} \mathcal{H}_y \mathcal{H}_u^\perp - \Omega_{\text{meas}}^{\text{est}} \mathcal{H}_u^\perp\|_{F} \leq \varepsilon
\]

Finally, once we have the estimate of output vector \( \hat{C} \) and system matrix \( \hat{A} \), one can again solve the optimization problem (5.8) to estimate \( \hat{B}, \hat{D}, \) and \( \hat{x}_0 \).

## 5.5 Identification Algorithm - Heuristics

In this section, we provide some heuristics aimed at improving the performance of the proposed approach. To improve the sparsity of the solutions obtained, we use the approach in [69] and iteratively solve the optimization problem with weight matrix \( W \) which depends on the previous solution. More precisely, given the value of \( \delta_{\text{pre}} \) obtained in the previous iteration, we take \( W = \text{diag}(1/(|\delta_{\text{pre}}| + \gamma)) \) with \( \gamma \) small and solve the optimization problem

\[
\{c^p, \delta\} = \text{argmin} \|W\delta\|_1 \tag{5.17}
\]
subject to
\[
\|c^p\|_\infty \leq \delta^p \text{ for all } p \in \mathcal{F}
\]
\[
y_n^{\text{est}}(k) = \sum_{p \in \mathcal{F}} c^p \alpha_p p^k \text{ for } k = 0, \ldots, N
\]
\[
\|\mathcal{H}_y \mathcal{H}_u^\perp - \mathcal{H}_{y_n}^{\text{est}} \mathcal{H}_u^\perp\|_{F} \leq \varepsilon
\]

Next, we briefly explain how a further heuristic that can increase sparsity. Recall that the proposed identification algorithm starts by gridding \( \mathcal{D}_\rho \) and obtaining a finite set \( \mathcal{F} \). If one wants a high precision algorithm, the number of grid points should be “large”. Since this first step sometimes only gives an approximation of the sparsest solution, after the first iteration, we eliminate the values of \( p \) with
numerically zero $\delta^p$. Then we use only the remaining poles to perform the next iteration of the algorithm. And eventually one will converge. Here is the complete outline of algorithm.

**Algorithm 5 ReWeighted Sparsification Algorithm**

1: Pick finite (large set) $\mathcal{F} \subset \mathbb{D}_p$ \hspace{1cm} $\triangleright$ Initialization
2: Construct matrix $\mathcal{H}^\perp_u$
3: Set $W$ to identity matrix
4: Set $\varepsilon$ to noise level, $\gamma$ and $\alpha$ to small number
5: Set $\delta_{\text{currval}}$ all zeros and Set $\delta_{\text{pervval}}$ all ones
6: while $- (\|\delta_{\text{pervval}}\|_1 - \|\delta_{\text{currval}}\|_1) \geq \alpha$ or $\text{card}(\delta_{\text{pervval}}) - \text{card}(\delta_{\text{currval}}) \neq 0$ do
7: Set $\delta_{\text{pervval}} = \delta_{\text{currval}}$
8: Solve (5.17)
9: Set $\delta_{\text{currval}} = \delta(\neq 0)$ update the set $\mathcal{F}$ be removing the points with numerically zero weight.
10: Set $W = \text{diag}(1/(|\delta_{\text{currval}}| + \gamma))$
11: end while
12: Construct $\hat{C}$ and $\hat{A}$ by using the poles $p$ and associated $c^p$ for which $\delta(p) \neq 0$ and solve (5.8).

It is worth emphasizing that noise level $\varepsilon$ is kept same at each iteration.

### 5.6 Results and Performance of Algorithm

#### 5.6.1 Experiment-1

In this section, the performance of the proposed algorithm is illustrated using an academic example. We consider a randomly generated following 2-input 3-output 3$^{rd}$ order system. Output measurements are obtained by exciting the system with an input $u$ that is a random sum of sinusoids and random initial conditions. The noise is uniformly distributed noise where $\varepsilon \in 0.08 \max(y) * U(-0.5, 0.5)$. Blocks of each output signal are randomly eliminated to simulate missing measurements. Finally, we used the following parameters for Algorithm 5, and cvx is used as a solver [70].

- True poles of the system $-0.9120, -0.6770 + 0.5902i, -0.6770 - 0.5902i$
- Number of data $N = 330$ or $k = 0, 1, \ldots, 329$
• Tolerance on $\|\mathcal{H}_y \mathcal{H}_u^\perp - \mathcal{H}_{y_{\text{est}}} \mathcal{H}_u^\perp\|_F \leq \varepsilon$, $\varepsilon = 24$

• Noise $\sim 0.08 \max(y) * \mathcal{U}(-0.5, 0.5)$

• Hankel Matrix Parameters $N - r = 30$

• 900 poles in $\mathbb{D}_1$

Figure 5.1: Real and Estimated Outputs Experiment-1 (MIMO)

Since we do not have infinitely many samples of the set $\mathbb{D}_1$, when solving problem (5.12) we cannot expect to obtain the exact location of the poles of the system. However, as we can see in the results presented, the approach proposed can closely approximate the true poles of the system even when one both has a significant amount of noise and missing data.
Figure 5.2: Impulse Responses Experiment-1 (MIMO)

<table>
<thead>
<tr>
<th>$|\delta|_1$</th>
<th>Iteration-1</th>
<th>Iteration-2</th>
<th>Iteration-3, 4, 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p \in \mathbb{D}$</td>
<td>-0.8485 + 0.0000i</td>
<td>-0.6667 - 0.6667i</td>
<td>-0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.9091 - 0.0606i</td>
<td>-0.6667 + 0.6667i</td>
<td>-0.6667 + 0.6061i</td>
</tr>
<tr>
<td></td>
<td>-0.9091 + 0.1212i</td>
<td>-0.6667 - 0.7273i</td>
<td>-0.6667 - 0.6061i</td>
</tr>
<tr>
<td></td>
<td>-0.9091 + 0.1212i</td>
<td>-0.6667 + 0.7273i</td>
<td>-0.6667 + 0.6061i</td>
</tr>
<tr>
<td></td>
<td>-0.9097 + 0.2424i</td>
<td>-0.6061 - 0.7879i</td>
<td>-0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.9097 + 0.2424i</td>
<td>0.6061 + 0.7879i</td>
<td>-0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.6667 - 0.6061i</td>
<td>0.3030 - 0.9091i</td>
<td>0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.6667 + 0.6061i</td>
<td>0.3030 + 0.9091i</td>
<td>0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.6667 - 0.5455i</td>
<td>-0.9697i</td>
<td>0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.6667 + 0.5455i</td>
<td>-0.9697i</td>
<td>0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.6061 + 0.7879i</td>
<td>-0.6667 + 0.6061i</td>
<td>-0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.9091 - 0.1212i</td>
<td>-0.6667 - 0.7273i</td>
<td>-0.6667 - 0.6061i</td>
</tr>
<tr>
<td></td>
<td>0.9091 + 0.1212i</td>
<td>-0.6667 + 0.7273i</td>
<td>-0.6667 + 0.6061i</td>
</tr>
<tr>
<td></td>
<td>0.9097 + 0.2424i</td>
<td>-0.6061 - 0.7879i</td>
<td>-0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.9097 + 0.2424i</td>
<td>0.6061 + 0.7879i</td>
<td>-0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.6667 - 0.6061i</td>
<td>0.3030 - 0.9091i</td>
<td>0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>0.6667 + 0.6061i</td>
<td>0.3030 + 0.9091i</td>
<td>0.3030 - 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.6667 - 0.5455i</td>
<td>-0.9697i</td>
<td>0.3030 + 0.9091i</td>
</tr>
<tr>
<td></td>
<td>-0.6667 + 0.5455i</td>
<td>-0.9697i</td>
<td>0.3030 - 0.9091i</td>
</tr>
</tbody>
</table>

Table 5.1: Poles and $\|\delta\|_1$ at Each Iteration
Figure 5.3: Real and Estimated Initial Condition Responses Experiment-1 (MIMO)

In this example, the algorithm stopped after 5 iterations of steps $7 - 12$. Complete information about the $\delta$ and the identified poles at each iteration can be found in Table 5.1.

Finally, the identification results are presented by several plots in this section. Figure 5.2, we have true and estimated impulse responses. Figure 5.3, we again compare true and estimated initial condition responses. Moreover, noisy and incomplete measurements and estimated outputs of the system are depicted in Figure 5.1.
5.6.2 Experiment-2

Another example is presented to show the effectiveness of the algorithm. Again a randomly generated 2-input 2-output 2\textsuperscript{nd} order system is considered. Similar to the previous example, the output are obtained by exciting the system with an input \( u \) (sum of sinusoids) and random initial conditions. The noise is again uniformly distributed noise \( \varepsilon \in 0.08 \max(y) \ast \mathcal{U}(-0.5, 0.5)^2 \). Missing measurements are randomly eliminated in the output measures. Finally, we used the following parameters for the algorithm 5.

- True poles of the system \( 0.5661 + 0.4949i, 0.5661 - 0.4949i \)
- Number of data \( N = 300 \) or \( k = 0, 1, \ldots, 299 \)
- Tolerance on \( \| \mathcal{H}_y \mathcal{H}_u^+ - \mathcal{H}_{y_{est}} \mathcal{H}_u^+ \|_F \leq \varepsilon, \varepsilon = 2.1 \)
- Noise \( \sim 0.08 \max(y) \ast \mathcal{U}(-0.5, 0.5) \)
- Hankel Matrix Parameters \( N - r = 30 \)
- 700 poles in \( \mathbb{D}_1 \)

\[ \mathcal{U}(a,b) \text{ Generate values from the uniform distribution on the interval } [a,b] \]
As one can see in the results below, the approach proposed can closely approximate the true poles of the system even when one has a significant amount of noise and also missing data.

Figure 5.5: Impulse Responses Experiment-2 (MIMO)

Figure 5.6: Initial Condition Responses Experiment-2 (MIMO)
Table 5.2: Poles and $\|\delta\|_1$ at Each Iteration

<table>
<thead>
<tr>
<th></th>
<th>Iteration-1</th>
<th>Iteration-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|\delta|_1$</td>
<td>3.3828</td>
<td>3.3829</td>
</tr>
<tr>
<td>$p \in \mathbb{D}_p$</td>
<td>$0.5517 - 0.4828i$</td>
<td>$0.5517 - 0.4828i$</td>
</tr>
<tr>
<td></td>
<td>$0.5517 + 0.4828i$</td>
<td>$0.5517 + 0.4828i$</td>
</tr>
</tbody>
</table>

In this example, the algorithm stopped after 2 iterations of steps 7 – 12. Complete information about the $\delta$ and the identified poles at each iteration can be found in Table 5.2. Several figures are depicted to show real and estimated signals.

### 5.6.3 Multi-Run Experiment

In the following example, a MIMO system (32-input 2-output) is run 4 times with a different sinusoidal input, and also some points in the data are randomly deleted as seen in the figures. Additionally, each run has different unknown initial condition response. We use the following parameters for this multi-run example.

- True poles of the system $-0.4895, 0.3649 + 0.6904i, 0.3649 - 0.6904i, 0.5709$
- Number of data $N_i = 300$ for all run or $k = 0, 1, \ldots, 99$
- Tolerance on $\|\mathcal{H}_y \mathcal{H}_u^+ - \mathcal{H}_{y^{est}} \mathcal{H}_u^+\|_F \leq \varepsilon$, $\varepsilon = 6$
- Noise $\sim 0.05 \max(y) \cdot \mathcal{U}(-0.5, 0.5)$
- Hankel Matrix Parameters $N - r = 30$
- 1900 poles in $\mathbb{D}_1$

In Figure 5.7, we present the output of each run and its estimate. Figure 5.7 demonstrates that, missing parts of each run are interpolated while keeping a certain fidelity. Next, Figure 5.8 presents the true and estimated initial condition responses. Finally, we present the impulse response of estimated and true system in Figure 5.9 with their transfer function.

In the Table 5.3, we show the results of the nuclear norm constraint in (5.16).

For the first and second poles ($-0.4898$ and $0.5306$), the rank of matrix $[c_1^p \ c_2^p \ c_3^p \ c_4^p]$ is 1. For the others, since nuclear norm minimization is an approximation of rank
minimization and the ratio of singular values is 16, then we can assume that the rank is 1 for third and forth poles.
Figure 5.8: Initial Condition Response of 4-Runs (MIMO)
Figure 5.9: Impulse Response of Each Channel Multi-Run (MIMO)
In following, we present the “Estimated and True System” matrices:

<table>
<thead>
<tr>
<th></th>
<th>True System</th>
<th>Estimated System</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.4267 0.3295 0.0793 0.3691</td>
<td>-0.4898 0 0 0</td>
</tr>
<tr>
<td></td>
<td>-0.4901 0.2806 0.4625 0.1928</td>
<td>0 0.5306 0 0</td>
</tr>
<tr>
<td></td>
<td>0.0653 -0.5456 0.4350 -0.0627</td>
<td>0 0 0.3673 - 0.6939i 0</td>
</tr>
<tr>
<td></td>
<td>0.0816 0.2916 0.2927 -0.3310</td>
<td>0 0 0 0.3673 + 0.6939i</td>
</tr>
<tr>
<td>B</td>
<td>2.0167 0 0</td>
<td>-0.2326 -0.0517 -1.2252</td>
</tr>
<tr>
<td></td>
<td>-1.7571 -0.0077</td>
<td>-6.2845 -4.7289 0.2333</td>
</tr>
<tr>
<td></td>
<td>0.0914</td>
<td>-0.0074 + 0.0760i -0.0637 - 0.0656i -0.0142 + 0.0047i</td>
</tr>
<tr>
<td></td>
<td>-0.1847 -1.0016</td>
<td>-0.0074 - 0.0760i -0.0637 + 0.0656i -0.0142 - 0.0047i</td>
</tr>
<tr>
<td>C</td>
<td>0.6818 0 -0.5413 -0.3071</td>
<td>-0.3066 -0.0252 -0.2611 + 10.0265i -0.2611 - 10.0265i</td>
</tr>
<tr>
<td></td>
<td>0 0 -0.7394 0.5675</td>
<td>0.4201 0.1064 3.4789 + 6.4046i 3.4789 - 6.4046i</td>
</tr>
<tr>
<td>D</td>
<td>0 0 0</td>
<td>-0.0052 0.0227 0.0054</td>
</tr>
<tr>
<td></td>
<td>-0.8792 0 -0.1730</td>
<td>-0.8321 -0.0106 -0.1791</td>
</tr>
<tr>
<td>I.C.</td>
<td>2.4577 0.0957 1.8583 6.6465</td>
<td>1.3508 1.9784 5.1595</td>
</tr>
<tr>
<td></td>
<td>0.9244 -2.1667 2.3958 1.1734</td>
<td>0.3123 + 0.0019i 0.2314 + 0.0504i</td>
</tr>
<tr>
<td></td>
<td>4.5396 3.2005 3.3061 -3.2428</td>
<td>0.3123 - 0.0019i 0.2314 - 0.0504i</td>
</tr>
</tbody>
</table>

Table 5.4: True and Estimated Systems - Multi-Run

As a result, the proposed algorithm successfully manages to identify a close system to the true system and interpolates the missing part of the data.
5.7 Concluding Remarks

In this Chapter, we have provided results aimed at the identification of MIMO LTI systems from noisy fragmented data. The main goal of the algorithm is to find a parsimonious model of the system (i.e., lowest order) that is compatible with the collected data, while at the same time, allowing for unknown initial conditions. The proposed approach combines ideas from both atomic norm methods and subspace identification to develop an algorithm that is able to exploit both a priori information on the system and the collected fragmented data to determine a low order model that explains the data collected and interpolates the missing measurements. In terms of future work, more effort will be devoted to improving the performance of the developed algorithms.

In Chapters 3, 4 and 5, we have mainly focused on identification algorithms where these algorithms could handle the difficulties arising from the nature of the behavioral data. Next, in Chapter 6, we will introduce the adaptive treatment design since if one has a dynamical model, then developing a control algorithm is always possible. In Chapter 6, we will mainly introduce how control engineering methods can be utilized to develop personalized treatment. All mathematical steps and an engineering perspective of adaptive treatment design are given. In addition, behavioral science perspective concerning engineering methods is introduced, followed by some experimental results.
Designing Adaptive Intensive Interventions

Human behavior (e.g., smoking addiction, exercise behavior) and reactions to treatment are complex and depend on many unmeasurable external stimuli. Since some stimuli are unknown, it is crucial to model human behavior over the data from many patients. When choosing a model, identifying the type of uncertainty that best describes the variations in measurements is critical. We propose a simple (low order) uncertain affine model that is subject to uncertainties; its response covers the most probable behavioral responses. This Chapter then provides a detailed step-by-step explanation of how control engineering methods can be used with intensive longitudinal data to design an adaptive intensive intervention. The methods are evaluated via simulation, and the results illustrate how the designed adaptive intensive intervention can result in improved outcomes with less treatment by providing treatment only when it is needed. Furthermore, the methods are robust to model misspecification as well as the influence of unobserved causes. Finally we will conclude that the new methods can be used to design adaptive interventions that are effective yet reduce participant burden.

6.1 Introduction

One of the main challenges clinical psychologists and psychiatrists increasingly face today is the complexity of the decision process when treating individuals
for chronic problems such as drug addiction, binge eating disorders, and non-suicidal self-injury. Decisions may include the best pharmaceutical intervention at the most appropriate dosage, the most promising therapeutic intervention, and the optimal duration of various treatments. In addition, clinical and behavioral scientists are increasingly collecting intensive longitudinal data (ILD) and more recently providing a variety of behavioral interventions intensively, often via web- or smartphone-based interventions [7, 10]. With this unparalleled opportunity for health behavior change comes the important new responsibility of deciding how best to adapt intensive behavioral interventions to the person.

Adaptive interventions operationalize the adaptation and re-adaptation of behavioral interventions; they are sequences of treatments that are adapted and re-adapted to individual circumstances and behaviors in order to achieve and maintain health behavior change. They adapt to the individual using decision rules, which input individual information such as symptoms, adherence to prior treatment, and biological measures, and output treatment recommendations, such as the most promising intervention or the appropriate dose of the intervention. An adaptive intensive intervention uses the intensive longitudinal data collected on an individual to deliver the appropriate intervention when the individual is in need. An example is provided by Gustafson and colleagues [10], who developed Alcohol-CHESS (ACHESS [CHESS stands for Center for Health Enhancement Systems Studies]), which is adaptive in that it provides recovering individuals access to social support when the GPS on their smartphone detects that they are near a high-risk location (e.g., an alcohol retail outlet). In this case the decision rule inputs the geographical location, checks whether this location is within a certain distance of established high-risk locations, and then alerts the individual that they may be at increased risk, giving them an opportunity to call someone for support. Individuals can also set up ACHESS to automatically alert a family member, peer, or professional when they are in a high-risk location for immediate social support. Thus this intervention provides treatment at the time when the individual most needs support. This adaptive intensive intervention provides a low-cost, context-sensitive way to augment more traditional treatment (e.g., in-person psychotherapy).

Scientists developing adaptive interventions typically use behavioral theory, literature reviews, and clinical experience to fully form decision rules specifying how
the intervention will be adapted to the person. However this approach to adaptive intervention design is difficult in an ILD context because the intensive nature of the intervention makes the decision process highly complex. Furthermore because the intervention may be provided repeatedly over time, it is crucial to provide treatment only when needed to avoid overburdening participants. Even when there is no monetary cost associated with additional treatment, more treatment may have other hidden costs, such as treatment fatigue. Identifying the intervention frequency that optimally balances effectiveness with participant burden can be difficult; further, the ideal frequency is likely to differ across individuals.

This dissertation discusses a possible control-engineering-based approach and explains how tools from robust control engineering can provide a systematic way to design adaptive intensive interventions using ILD. In these approaches algorithms for adapting treatment are developed using existing observational ILD data. Then at each time point during the implementation of the intervention for a new individual, the developed algorithm uses the most recent data collected on the individual to recommend an intervention for this individual at this time. As we demonstrate, when treatment fatigue is an issue, this approach is superior to full treatment in which every individual is treated at every instance. The adaptive intensive intervention will use less total treatment than a standard protocol, yet it will result in stronger effects because it uses treatment only when it is needed.

To make the ideas and tools concrete we illustrate the design of an adaptive intensive intervention for smoking-cessation. Combined treatments involving both a pharmacological and a behavioral component have been most successful in achieving smoking cessation [71]. Yet even with effective smoking cessation aids, less than 30 percent of smokers achieve long-term abstinence in a given quit attempt [71]. Thus we focus on designing an adaptive intensive behavioral intervention for use in conjunction with a pharmacological treatment.

Smoking lapse is highly associated with intra-individual phenomena such as craving [72]. Higher self-efficacy has been associated with lower self-reported nicotine craving among smokers [73], and there is extensive evidence that negative affect is closely linked to craving [74, 75]. Both self-efficacy and negative affect have been the focus of previous smoking cessation treatments [76]. Therefore, our hypothetical intervention is designed to improve self-efficacy and decrease negative
affect, and thus decrease smoking urge. In the developed intervention the smartphone will collect ecological momentary assessments (EMA) data three times per day (approximately morning, midday and evening); following each data collection, the smartphone will either not prompt or prompt the participant to access a smartphone behavioral application. This hypothetical smart-phone application will contain a variety of intervention components such as messages designed to maintain motivation, improve coping skills, provide positive reinforcement from the participants primary care doctor, and remind the participant why he/she wants to quit smoking.

For illustrative purposes we assume we have ILD from a study, n = 150 in which participants are prompted by their smart-phone to provide data in the morning, midday and in the evening for 50 days. This results in data from time points \( k = 1 \) through 150 for each participant At each time \( t \), participants report their momentary self-efficacy \( (s_{ek}) \), negative affect \( (n_{ak}) \), and smoking urge \( (s_{uk}) \), each on a continuous scale. Furthermore, participants may receive treatment by accessing the smart-phone behavioral application following each point of data collection (i.e., the participant may access the behavioral application up to three times per day). Whether or not the participant accessed the behavioral application at a particular time is indicated by \( T_k \) (coded by 1 if the participant receives the treatment and coded by 0 otherwise).

### 6.1.1 Modeling of Behavior

Dynamical models of behavior are critical in developing control algorithms [77, 78, 79, 80]. However, according to our best knowledge, very little attention has been paid to the fact that human behavior has a significant amount of uncertainty and that this uncertainty should be addressed in a systematic way. This dissertation provides an approach for modeling the patient behavior and designing a robust adaptive treatment that takes into account both the present state of the patient and the probable perturbations to expected behavior. Furthermore the affine linear system is commonly used in the literature for modeling behavior (affine equation, structural equation models) [38]. In this context, one could consider more complex models, such as polynomial ones, only “simple” affine models will be considered
here since very little is known about the structure of models of human behavior. Nevertheless, the approach here can be extended to more complex models such as nonlinear and time-varying models.

The type of the model and uncertainties which is used in this Chapter to design robust adaptive treatment are already explained in Chapter 3. Moreover, Lasso as an identification for these model and uncertainties is also given in Chapter 3.

6.1.2 Previous Work

Since adaptive intervention adapts to individual progress, it is effective to prevent insufficient response and react immediately against unexpected shock on behavior. Therefore, scientists in social or behavioral science have been working on finding systematic ways to use portable devices such as smart phones to change health behavior or maintain healthy behavior in real time. Consequently, adaptive interventions (in control jargon, considered as feedback control) are being developed by social and behavioral scientists for different areas such as criminal justice [81], mental health [29, 82], hypertension [31], substance abuse [32, 33, 34], and Alzheimer’s disease [83].

Previous researchers have also applied some control concepts [77, 78, 84, 85, 86] in behavioral research, but the use of feedback is still a novel approach. Additionally engineering concepts such as dynamical modeling have been applied to the modeling and controlling of behavior [80]. [77, 78, 79] present special issues that arise in applying such approaches to behavioral research. These studies employ both time-invariant and time-varying models, depending on the specific problem. Some of the literature also uses controller design methods to design adaptive interventions for special problems in behavioral science [84, 85, 86]. In contrast to these studies, this thesis defines uncertainties in the behavior more explicitly in the model that affect the robustness of the algorithm. In general, according to the intervention design problem, the model predictive controller design method is preferred.
6.2 Designing an Adaptive Intensive Intervention Using Control Engineering

In this section, we provide a step-by-step explanation of control engineering methods that can inform the development of an adaptive intensive intervention using ILD. The methods are illustrated by considering the design of the adaptive intensive, smoking-cessation intervention discussed in the introduction. First we fit a dynamical model, which in this case is a system of regression equations, to the data. We then provide an example of how the scientific team can quantify the treatment goal. We then describe how to use the dynamical model with estimated regression coefficients and the treatment goal to develop the decision rules comprising the adaptive intensive intervention. We use two different types of methods to solve optimization problem of adaptive intervention. Finally, we describe how to implement the adaptive intensive intervention with new individuals.

6.2.1 Dynamical Model of Smoking Cessation

The first step is to fit a dynamical model to the data, that is, a model for the variables at each time point $k$. Recall that three continuous dependent variables, smoking urge, negative affect, and self-efficacy ($su_k, na_k, se_k$), are assessed at each time $t$ on each participant. Here we use linear models in which each variable at time $k$ depends linearly on both the variables and treatments at times $k-1$ and $k-2$. For example the linear models (for smoking cessation) might be:

$$su_k = a_0 + a_1 su_{k-1} + a_2 su_{k-2} + b_1 na_{k-1} + b_2 na_{k-2} + c_1 se_{k-1} + c_2 se_{k-2} + \varepsilon_k$$  \hspace{1cm} (6.1)

$$na_k = d_0 + d_1 su_{k-1} + d_2 su_{k-2} + e_1 na_{k-1} + e_2 na_{k-2} + f_1 se_{k-1} + f_2 se_{k-2} + q_1 T_{k-1} + q_2 T_{k-2} + \varepsilon_{na}^k$$ \hspace{1cm} (6.2)

$$se_k = g_0 + g_1 su_{k-1} + g_2 su_{k-2} + h_1 na_{k-1} + h_2 na_{k-2} + r_1 se_{k-1} + r_2 se_{k-2} + s_1 T_{k-1} + s_2 T_{k-2} + \varepsilon_{se}^k$$ \hspace{1cm} (6.3)
where $\sigma(\varepsilon_k) = \sigma^2$, $\sigma(\varepsilon_{na}^k) = \sigma_{na}^2$, $\sigma(\varepsilon_{se}^k) = \sigma_{se}^2$ for $k = 1, \ldots, 150$ and \{\varepsilon(1), \varepsilon_{na}^1(1), \varepsilon_{se}^1(1), \varepsilon(2), \varepsilon_{na}^2(2), \varepsilon_{se}^2(2), \ldots, \varepsilon(150), \varepsilon_{na}^{150}(150), \varepsilon_{se}^{150}(150)\} \text{ are independent, normal, mean zero errors}. Note the model for $su_k$ (see equation (6.1)) does not include treatment, as our hypothetical behavioral intervention is assumed to directly impact self-efficacy and negative affect. Equation (6.1) also does not include sparse external disturbance which will be covered in the next sections. Note that this model can be easily formatted similar to equation (3.1) and model parameters can be identified by using Lasso algorithm in Chapter 3.

Treatment could be included in this model as an independent variable if it is expected to impact smoking urge directly. Furthermore, (6.2) and (6.3) permit treatment fatigue; for example, if in model (6.2) the regression coefficients of the terms involving treatment are of opposite sign, the effect of treatment $T_{k-1} = 1$ will be washed out if treatment has been applied at time $k - 2$; i.e., if $T_{k-2} = 1$. In all of these models data from even earlier time points, say time $k - 3$, may be included. Because the error terms are assumed independent, three separate standard regression models can be fit, one each for self-efficacy, negative affect, and smoking urge, corresponding to equations (6.1), (6.2), and (6.3) above. Appendix B.2 (available also at http://dx.doi.org/10.1037/a0037736.supp) provides hypothetical data from 150 individuals to demonstrate how to structure the data. Together, fitting these three models will provide estimates of all regression coefficients and error variances, $\hat{\sigma}^2$, $\hat{\sigma}_{na}^2$, and $\hat{\sigma}_{se}^2$.

### 6.2.2 Objectives of Adaptive Treatment Design

The second step is to quantify the treatment goal for a new individual. In our example, the treatment goal is to reduce smoking urge. We translate this treatment goal into an “objective function” that will be used to determine the treatment, if any, that will be provided at each time point for each individual. An objective function that reflects this goal for this individual at each time point $k$, is the sum of squares of smoking urge:

---

1A more sophisticated model could include additional features such as diurnal trends that allow for within-day patterns in nicotine craving or characteristics such as level of dependence at baseline. Furthermore, the error terms could be correlated across time.
\[
\sum_{t \geq k+2} [su_t - \theta]^2 \quad (6.4)
\]

We want to provide an adaptive intervention so as to minimize this individual's sum of squares of smoking urge. First, often a tuning parameter \(\theta\) is subtracted from each smoking urge; \(\theta\) can be interpreted as the average level of smoking urge achievable by particular treatment(s). Equation (6.4) is implicitly a function of treatment at time \(k\). This is easiest to see by considering only the first term in the sum, namely \([su_{k+2}]^2\). We “unpack” \([su_{k+2} - \theta]^2\) as follows. First consider equation (6.1) with \(k + 2\) in place of \(k\). We see that \(su_{k+2}\) and thus \([su_{k+2}]^2\) is a function of \(su_{k+1}, su_k, se_{k+1}, se_k\) and \(na_{k+1}, na_k\). Next consider equations (6.2), (6.3) with \(k + 1\) in place of \(k\). Unpack \([su_{k+2} - \theta]^2\) further by substituting \(se_{k+1}, na_{k+1}\) with their respective formula from equations (6.2) and (6.3). After making these substitutions, we see that \([su_{k+2}]^2\) is actually a function of treatment at time \(k, T_k\), as well as smoking urge at time \(k\) and time \(k - 1, su_k, su_{k-1}\), self-efficacy at time \(k\) and time \(k - 1, se_k, se_{k-1}\), negative affect at time \(k\) and time \(k - 1, na_k, na_{k-1}\) and past treatment, \(T_{k-1}\). Thus, the equation (6.4) is a function of treatment at time \(k\), present and past smoking urge, present and past self-efficacy, present and past negative affect and past treatment. Suppose we have observations of present and past smoking urge, present and past self-efficacy, present and past negative affect and past treatment on this new individual then we can compare the value of equation (6.4) when \(T_k\) is set to 1, to the value of equation (6.4) when \(T_k\) is set to 0. If \(T_k = 1\) leads to a smaller value of equation (6.4), then the optimal decision is to provide treatment to this new individual at time \(k\). Now equation (6.4) contains more than just the first term, \([su_{k+2} - \theta]^2\); resulting in a more complex objective function. However, none-the-less, once unpacked, equation (6.4) is a function of treatment at time \(k\), present and past smoking urge, present and past self-efficacy, present and past negative affect and past treatment.

There are three aspects of objective function in equation (6.4) that require explanation.

1. If two new individuals have the same observed values of present and past smoking urge, present and past self-efficacy, present and past negative affect
and past treatment then the results of the MPC algorithm will be the same (i.e., the two individuals will be recommended the same treatment at that time).

2. Since equation (6.4) includes a sum over all time points beginning at \( k + 2 \), this objective function adjusts for the presence of delayed effects of time \( k \) treatment. Thus, for a treatment at time \( k \) to be recommended, the treatment must not only be predicted to reduce smoking urges at time \( k + 2 \), it must also be predicted to lead to a reduction in smoking urge at later time points, at least as compared to no treatment at time \( k \).

3. Each term in equation (6.4) squares the smoking urge, \( su_t \); the square acts to prevent treatment at time \( k \) if treatment is predicted to lead to too many future occasions of very high smoking urge. In particular this means that if no treatment at time \( k \) is likely to lead to a constant moderate smoking urge in the future, but treatment at time \( k \) is likely to lead occasionally to small smoking urges but often to high smoking urge, then no treatment is better.

### 6.2.3 Robust Model Predictive Control for Adaptive Treatment Design

Now, we assume that \( w_k \neq 0 \) in equation (3.1) in this section. \( w_k \) is a sparse disturbance.\(^2\) In this section, we will develop a more general robust adaptive interventions against the disturbances \( w_k \) and \( \varepsilon_k \) in the model (3.1). All mathematical details of optimization problem and the objective function in (6.4) to design robust algorithm are given. Moreover, a more compact form of the objective function is also calculated in this section. For the completeness, we restate the model (3.1) and the set \( \mathcal{W} \) which is presented in Chapter 3.

\[
y_k = a_f + \sum_{i=1}^{n} [Ay_{k-i} + BT_{k-i}] + Cw_k + D\varepsilon_k, \quad (6.5)
\]

\(^2\)\(w_k\) is used to model life events such as loosing job etc. that affect the treatment response.
and set $\mathcal{W}$.

$$\mathcal{W} = \{ \mathbf{w} \in L^\infty : \text{for all } k, \| \mathbf{w}(k,k+K-1) \|_\infty \leq \alpha, \| \mathbf{w}(k,k+K-1) \|_1 \leq \gamma, \| \mathbf{w}(k,k+K-1) \|_0 = K/\varsigma \}.$$ 

Again the model parameters and description of the uncertainties identified by using the Lasso algorithm in Chapter 3.

Here, the usual approach in robust MPC is taken: one estimates the present value of the state variable and determines the value of the control variables over the horizon that minimize a given objective function under the uncertainties. Similar to previous section 6.2.2, the first of these control signals is applied, and the process is then repeated. The control algorithm is summarized in Figure 6.1. In step 1, we collect the initial conditions or the information $y_{k-1} \ldots y_{k-n}$ (in equation (6.5)) that the control algorithm will use. In the smoking cessation study, this information is $su_{k-1} \ldots su_{k-n}$, $na_{k-1} \ldots na_{k-n}$, $se_{k-1} \ldots se_{k-n}$. Additionally, after the researchers discussed with the practitioner in the field, the total given treatment to the patient is limited with additional constraint on treatments. Therefore, more treatment data is needed for this constraint. As a result, more control input information $T_{k-1} \ldots T_{k-l}$ ($l \geq n$) is collected in order to enforce constraint on the total treatment provided.

In line with the usual MPC approach, in step 2 one minimizes a given cost function subject to constraints on the total applied control in certain range. Then in step 3, the result of the control algorithm is applied to the patient. Therefore in this section a decision rule will be developed that will dictate whether treatment will be applied to each individual at each time point under the uncertainties.

The difference equation (3.1) is used to determine a closed-form objective func-
tion for MPC formulation with a given receding horizon $K$. In the objective function, we represent $\mathcal{Y}_{k+1} \in \mathbb{R}^{Km}$ as:

$$\mathcal{Y}_{k+1} = A_f + \tilde{A}\mathcal{Y}_0 + \tilde{B}\tilde{T}_k + \tilde{D}\tilde{\epsilon}_k + \tilde{E}\tilde{w}_k.$$  \hfill (6.6)

where $\mathcal{Y}_0 \in \mathbb{R}^{nm}$ (remember that $y_k \in \mathbb{R}^m$ in equation (3.1)) is the vector containing the state of the model at time $k$. $\tilde{T}_k \in \mathbb{R}^{K+n-1}$ is the treatment. $\tilde{w}_k \in \mathbb{R}^K$ is sparse disturbance and $\tilde{\epsilon}_k \in \mathbb{R}^K$ is noise. $A_f$, $\tilde{A}$, $\tilde{B}$, $\tilde{D}$, and $\tilde{E}$ are matrices which are calculated recursively from difference equation (3.1).

Then, in this Chapter the objective function of the following form is considered:

$$((\mathcal{Y}_{k+1} - \theta)X(\mathcal{Y}_{k+1} - \theta)^T$$

although more complex convex functions can be addressed by the framework presented here. For the smoking cessation example, since one minimizes the smoking urge, matrix $X$ is defined to choose smoking urge measure in vector $\mathcal{Y}_{k+1}$ as:

$$X = \begin{cases} 
1 & \text{if } x_{i,i} = k, k+m, k+2m, \ldots, k + (K-1)m; \\
0 & \text{otherwise.}
\end{cases}$$
the algorithm;

$$\min_{T \in T} \max_{\tilde{\epsilon}_k \in \mathcal{W}_k} \min_{\tilde{\epsilon}_k \in \mathcal{W}_k} \| \tilde{\epsilon}_k \|_2 \leq \rho \quad (6.7)$$

which is subject to the system dynamics described in equation (6.6). We also define $\mathcal{W}_k$ from the set $\mathcal{W}$ as:

$$\mathcal{W}_k = \{ \text{vector } \tilde{\epsilon}_k \in \mathbb{R}^K \text{ satisfies : } \| \tilde{\epsilon}_k \|_\infty \leq \alpha, \| \tilde{\epsilon}_k \|_1 \leq \gamma, \| \tilde{\epsilon}_k \|_0 = K/\varsigma \}.$$  \( (6.8) \)

Then it should be noted that the set $\mathcal{W}_k$ is a union of the convex polytopes; this fact is explored later when solving the robust optimization problem resulting from the MPC formulation of the control problem.

Next, the set $\mathcal{T}$ that is the set of allowable control signals is defined for receding horizon $K$.

**Smoking Cessation Example:** In the case of the smoking cessation example;

$$\mathcal{T} = \left\{ \text{T : } T_\eta \in \{0, 1\} \text{ for } \eta = k, \ldots, k + K - 1 \text{ and } \sum_{\eta = k-l}^{k+K-1} T_\eta \leq T_{total} \right\} \quad (6.9)$$

represent the fact that one can either apply or not apply treatment at each time. Additionally the number of treatments applied in the last $K + l$ instances is constrained to be less than or equal to $T_{total}$. This constraint in the number of treatments is aimed at addressing the problem of treatment burden [45, 46, 47, 48], which is decreasing effectiveness as the number of applied treatments increases. In this specific example, treatment $T$ is binary, but in reality it can have a different level and type of treatment. One should note that the optimization problem described above is complex.

In this study, we use two different method to solve equation (6.7) in MPC. i) Scenario Approach and ii) Robust Approach. Results of both methods will be presented in result section.

**Scenario Approach:** This approach is usually preferred to deal with uncertainty in the optimization problem based on the constraint samples. In this method, the algorithm receives several random samples of the uncertainty and
find the optimal solution of the problem for these sampled subset of uncertainty [87, 88, 89]. This theory gives a solid level to the randomization step for robust or chance-constrained optimization. We also utilized this approach to solve the optimization problem in equation (6.7). The scenario based finite horizon MPC problem is recently studied in [90]. This scenario approach is also preferred as one of the methods to solve MPC for adaptive treatment.

**Robust Approach:** The robust approach is designed to deal with all set of uncertainties in the optimization problem. This approach provides guarantees about the performance of the solution.

For the robust approach, a problem equivalent to the original problem (6.7) is now presented that is suitable for implementation in Theorem 14. The minimax formulation of robust least square problems with the uncertain data presented in [3, 4, 5] is used to reformulate our minimax problem in (6.7). Theorem 16 is used to handle the uncertainty \( \| \tilde{e}_k \|_2 \leq \rho \). This result is simply restated for completeness in Appendix B.3. For the uncertainty \( \tilde{w}_k \in W_k \) in LMI, the following Theorem 14 is defined.

**Theorem 14.** The results of the LMI in following optimization problem (6.10) is an upper bound on the original optimization problem in (6.7) and is exact if only one of the elements in the set \( W_{ext} \) leads to an active constraint at the optimum.

\[
\begin{align*}
\min_{\tau, T_k \in T, \lambda} & \quad \tau \\
\text{subject to} & \quad \begin{bmatrix}
\tau - \lambda \rho^2 & 0 & (A_f - \theta + \tilde{A}Y_0 + \tilde{B}T_k + \tilde{E}\tilde{w}_k)^T \\
0 & \lambda I & (\tilde{D}X)^T \\
(A_f - \theta + \tilde{A}Y_0 + \tilde{B}T_k + \tilde{E}\tilde{w}_k)X & \tilde{D}X & I
\end{bmatrix} \geq 0,
\end{align*}
\]

for all \( \tilde{w}_k \in W_{ext} \) where \( W_{ext} \) are the extremes of the polytopes whose union is \( W_k \).

**Proof.** See Appendix B.4.

We now characterize the set of extremes, \( W_{ext} \).

**Theorem 15.** Consider set \( W_k \) defined in (6.8), where \( \alpha, \gamma \) and \( \varsigma \) define the bounds on the \( \ell_\infty \) norm, \( \ell_1 \) norm, and sparsity respectively. Let \( n_{nz} = \text{dim}(\tilde{w}_k)/\varsigma \).
Then

- If $\alpha \geq \gamma$, then $W_{ext}$ is the set of all signed permutations of
  \[
  \begin{bmatrix}
  \gamma & 0 & \cdots & 0 \\
  & & & \\
  & & & \\
  & & & \\
  \end{bmatrix};
  \]

- If $n_{nz}\alpha \leq \gamma$, then $W_{ext}$ is the set of all signed permutations of
  \[
  \begin{bmatrix}
  \alpha & \cdots & \alpha & 0 & \cdots & 0 \\
  & & & & & \\
  & & & & & \\
  & & & & & \\
  \end{bmatrix};
  \]

- If $\gamma < n_{nz}\alpha < n_{nz}\gamma$, then $W_{ext}$ is the set of all signed permutations of
  \[
  \begin{bmatrix}
  \alpha & \cdots & \alpha & h & 0 & \cdots & 0 \\
  & & & & & & \\
  & & & & & & \\
  & & & & & & \\
  \end{bmatrix},
  \]

where

\[
  n_+ = \left\lfloor \frac{\gamma}{\alpha} \right\rfloor \quad \text{and} \quad h = \gamma - n_+\alpha.
\]

Proof. See Appendix B.5. \qed

After finding the extreme points of the set $W_k$, a semi-definite programming or mixed integer semi-definite programming solver could be used to solve problem (6.10), but for a large receding horizon $K$, the number of extremes and, thus the number of LMI constraints, might be fairly large.

The next section presents a simulated application of the algorithm using a model that mimics data that might have been collected in smoking cessation studies such as [43, 44].

### 6.3 Implementation of the Adaptive Intervention

In this section, two different possible implementation methods are explained, one with a smart phone app and one with a decision rule table.
6.3.1 With Smart Phone App

As an implementation of the adaptive treatment algorithm that is proposed by using portable devices, we consider individuals who wish to quit smoking and we present a hypothetical treatment algorithm that provides treatment through a smart phone app. The participant provides information on his/her level of smoking urge, negative-affect, and self-efficacy on the morning of day 1; the portable device then provides the behavioral treatment. Robust adaptive treatment is begun at midday after the participant again provides his/her level of smoking urge, negative-affect, and self-efficacy. Right now, $su^i_0$, $na^i_0$, $se^i_0$, $T^i_0$ and $su^i_1$, $na^i_1$, $se^i_1$ are known, where $T^i_1$ is treatment provided in the morning. The proposed adaptive treatment algorithm is run and decides $T^{i\text{optimal}}_i$ and the set $T^i_2 = T^{i\text{optimal}}_i$. That is, if $T^{i\text{optimal}}_i = 1$, treatment is provided at midday; otherwise it is not. Afterwards, new $su^i_3$, $na^i_3$, $se^i_3$ are collected, then by using these new values and $su^i_2$, $na^i_2$, $se^i_2$, $T^i_2$ the process is repeated to obtain $T^i_3 = T^{i\text{optimal}}_i$. This process is repeated throughout the entire process of the treatment. Here are the steps:

Step 1: Decide on message frequency, (assume contact with the patient three times a day morning, midday, evening).

Step 2: Patient provides information on his/her measurements $su^i_0$, $na^i_0$, $se^i_0$, $T^i_0$ in the morning.

Step 3: At midday message collect $su^i_1$, $na^i_1$ and $se^i_1$.

Step 4: Build measurement set $\mathcal{M}$ (e.g. at midday, it has $su^i_0$, $na^i_0$, $se^i_0$, $T^i_0$, $su^i_1$, $na^i_1$ and $se^i_1$)

Step 5: Run MPC algorithm with identified model $f$, calculate $T^{i\text{optimal}}_i$, update measurement set $\mathcal{M}$, and message $T^i_1 = T^{i\text{optimal}}_i$ to the patient.

Step 6: Collect new state values $su^i_2$, $na^i_2$, $se^i_2$ at evening and update measurement set $\mathcal{M}$.

Step 7: Repeat this process throughout the entire process of the treatment.
6.3.2 Decision Rule Table

In practice, prior to providing any treatment for a new individual, we can run the MPC algorithm for each possible value of smoking urge, negative affect, and self-efficacy at past and present time (times $k - 1$, $k$ respectively) and the treatment indicator at the past time (time $k - 1$; e.g., all values of $se_k$, $na_k$, $su_k$, $T_{k-1}$, $se_{k-1}$, $na_{k-1}$ and $su_{k-1}$) to obtain the best values for the treatment $T_k$. We could use this to produce a large decision rule table in which each row corresponds to a different value of $se_k$, $na_k$, $su_k$, $T_{k-1}$, $se_{k-1}$, $na_{k-1}$ and $su_{k-1}$ and the last column of the table provides the best treatment at the present time $k$. Table 6.1 provides a small subset of the rows that would be found in the full table. The full decision rule table provides the decision rules for the adaptive intensive behavioral intervention.
Table 6.1: A Partial Decision Rule Table

<table>
<thead>
<tr>
<th>Description of Individual Observations</th>
<th>Input</th>
<th>Output Recommended Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-Efficacy $(k-1)$</td>
<td>Negative Affect $(k)$</td>
</tr>
<tr>
<td>Low, improving self-efficacy; High, improving negative affect; High, improving smoking urge; No prior treatment</td>
<td>1.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Low, improving self-efficacy; High, improving negative affect; High, improving smoking urge; No prior treatment (more extreme than the above)</td>
<td>0.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Low, stable self-efficacy, High deteriorating negative affect, High, deteriorating smoking urge; Prior Treatment</td>
<td>2.0</td>
<td>1.8</td>
</tr>
<tr>
<td>Moderate, stable self-efficacy; Moderate, stable negative affect; High, stable smoking urge; Prior treatment</td>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td>High, improving self-efficacy; High, stable negative affect; High, deteriorating smoking urge; No prior treatment</td>
<td>3.8</td>
<td>4.5</td>
</tr>
<tr>
<td>High, deteriorating self-efficacy; Low, stable negative affect; High, stable smoking urges; Prior treatment</td>
<td>4.9</td>
<td>4.3</td>
</tr>
<tr>
<td>Low deteriorating self-efficacy; Low, deteriorating negative affect; Moderate stable smoking urge; Prior treatment</td>
<td>1.69</td>
<td>0.80</td>
</tr>
<tr>
<td>High deteriorating self-efficacy; Low, stable negative affect; Low deteriorating smoking urge; No prior treatment</td>
<td>3.47</td>
<td>2.97</td>
</tr>
</tbody>
</table>

Note: $t$ refers to current time for which treatment recommendation is being made, in this case, $t = 75$. In this population, the mean self-efficacy is 3.2 with a standard deviation of 1.0, the mean smoking urge is 2.8 with a standard deviation of 1.8 and the mean negative affect is −0.08 with a standard deviation of 1.1.
6.4 An Illustrative Example of the Development and Implementation of an Adaptive Intensive Intervention using MPC

In this section, we illustrate how the above steps for designing an adaptive intensive intervention can be carried out. We do this using a simulated intensive longitudinal data set that loosely resembles the data from a smoking cessation study [44, 91, 43]. The simulated data set contains 150 participants from a hypothetical smoking cessation study conducted over a period of 50 days; measurements were collected three times per day. At each occasion treatment was provided with probability 1/2. See the appendix section B.2 for the model used to generate the simulated data; this data set is available at http://dx.doi.org/10.1037/a0037736.supp.

The first step is to fit equations (6.1), (6.2), and (6.3) to the data by using Lasso in Chapter 3. This results in the fitted regression models,

\[
\begin{align*}
su^i_k &= 1.12 + 0.48su^i_{k-1} + 0.34su^i_{k-2} + 0.016na^i_{k-1} - 0.015NA^i_{k-2} \\
&- 0.015SE^i_{k-1} - 0.008SE^i_{k-2} + \varepsilon^i_k \\
na^i_k &= .14 + 0.003su^i_{k-1} + 0.0009SU^i_{k-2} + 0.49na^i_{k-1} + 0.25na^i_{k-2} \\
&- 0.05SE^i_{k-1} + 0.005SE^i_{k-2} - 0.22T^i_{k-1} + 0.15T^i_{k-2} + \varepsilon^{na^i}_k \\
se^i_k &= 0.54 - 0.008su^i_{k-1} - 0.004su^i_{k-2} - 0.018na^i_{k-1} + 0.001na^i_{k-2} \\
&+ 0.49se^i_{k-1} + 0.36se^i_{k-2} + 0.18T^i_{k-1} - 0.15T^i_{k-2} + \varepsilon^{se^i}_k
\end{align*}
\]  

(6.11) (6.12) (6.13)

with estimated error variances \(\hat{\sigma}^2 = 1.1\), \(\hat{\sigma}_{na}^2 = 0.6\), and \(\hat{\sigma}_{se}^2 = 0.3\). Equation (6.11) indicates that smoking urge at any time \(k\) is positively correlated with past smoking urge, and negatively correlated with past self-efficacy. Furthermore equation (6.11) shows a mixed relationship between smoking urge at time \(k\) and past negative affect, depending on the time lag.

The next step is to design an adaptive intensive intervention algorithm using the above fitted regression models. We use the MPC algorithm as described in the subsection 6.2.3. For further details concerning the objective function and param-
eters of the MPC algorithm see Appendix B.1 at http://dx.doi.org/10.1037/a0037736.supp. Data from the 150 participants are used to fit the dynamical models; there is no further use of the data from the 150 participants. The dynamical models are then used to specify the objective function as described earlier. Use of the MPC algorithm results in decision rules which, based on observations of past and present self-efficacy, past and present negative affect and past treatment, provide the best treatment at the present time $k$. The (partial) decision rule table provided in Table 6.1 was constructed using the MPC algorithm.

Consider row one. This row corresponds to observations indicating poor self-efficacy, negative affect, high smoking urge and no treatment at prior time, yet improvement in each of these variables from time $k - 1$ to time $k$; the MPC algorithm recommends no treatment for an individual with these observations. The observations in row two are similar to row one, but with an excessively large observed smoking urge; the algorithm recommends treatment. In row three, the observations correspond to stable self-efficacy, but poor levels of negative affect and smoking urge and treatment at the prior time. Here the MPC algorithm does not recommend treatment because treatment had been given at the previous time.

In Figures 6.2 and 6.3, we illustrate the implementation of the above adaptive intensive intervention (corresponding to the full decision rule table) on a simulated individual. This figure shows the evolution of the individuals smoking urge for two different interventions: in blue we depict the response to the adaptive intensive intervention (using the decision rule table), and in green we show the response to full treatment intervention (treatment three times per day). Moreover, the red “stars” indicate when treatment is provided by the adaptive intervention. As one can see, full treatment works better in the beginning, when the individual has not yet experienced treatment fatigue. However, the proposed adaptive intervention is more effective in the long term. In particular, when the model forecasts that smoking urge is expected to be excessive, then treatment is provided. As a result treatment is only provided as needed, with the decision based on present and past information about the patient. Otherwise, treatment is not provided. As shown in Figures 6.2 and 6.3, treatment is provided when an increasing trend in smoking urge occurs.
6.5 A Simulation Study Evaluating the MPC Approach

To evaluate the MPC approach to developing an adaptive intensive intervention we conducted a simulation study. We consider two different populations of individuals labeled A and B, defined below. For each population, 500 data sets are simulated, each containing 150 virtual participants. Each data set represents a smoking cessation study conducted over a period of 50 days; as before, measurements are collected three times per day. At each measurement occasion treatment is provided with probability $1/2$. The models used to generate the data are

$$su_k^i = 1.1 + .50su_{k-1}^i + .35su_{k-2}^i + .01na_{k-1}^i - .12se_{k-1}^i + .01se_{k-2}^i + w_k^i + \varepsilon_k^i \quad (6.14)$$

$$na_k^i = .15 + .01su_{k-1}^i + .49na_{k-1}^i + .24na_{k-2}^i - .05se_{k-1}^i + treat^na(T_k) + \varepsilon_k^{na} \quad (6.15)$$
Figure 6.3: Difference in Smoking Urge under the Adaptive Intensive Intervention and Full Treatment for One Hypothetical Individual

\[ se_k^i = 0.52 - 0.01su_{k-1}^i - 0.02na_{k-1}^i + 0.49se_{k-1}^i + 0.36se_{k-2}^i \]  
\[ \text{treat}^{se}(T_k) + \varepsilon_k^{se} \]  

where \( \alpha(\varepsilon_j) = 1.2, \alpha(\varepsilon_{ja}^j) = 0.6, \alpha(\varepsilon_{se}^j) = 0.3 \) for \( j = 1, \ldots, 150 \) and \( \{ \varepsilon_1^i, \varepsilon_1^{na}, \varepsilon_1^{se}, \varepsilon_2^{na}, \varepsilon_2^{se}, \ldots, \varepsilon_n^{na}, \varepsilon_n^{se} \} \) are independent, normal, mean zero errors. The term \( \varepsilon_k^{se} \) in equation (6.14) reflects random disturbances (i.e., unobserved momentary events that affect urge to smoke at that moment, e.g., family stress, health issues).  

From equation (6.14) we see that higher smoking urge, more negative affect, and lower self-efficacy at the previous time predicts higher smoking urge at time \( k \). In equation (6.15) we see that higher smoking urge, more negative affect, and lower

\(^3\)In general scientists attempt to collect good measures of relevant constructs; however due to complexity of behavior change, there are likely unknown, and thus unobserved, momentary events.
self-efficacy at past time points predicts more negative affect at time t. Finally, in equation (6.16), lower smoking urge, less negative affect, and higher self-efficacy at past time points predicts higher self-efficacy at time \( k \).

The treatment terms \( \text{treat}^\text{na}(\bar{T}_k) \) in equation (6.15) and \( \text{treat}^\text{se}(\bar{T}_k) \) in equation (6.16) are given by

\[
\text{treat}^\text{na}(\bar{T}_k) = -0.2T_{k-1} + 0.2\text{logistic}(\sum_{u=k-2}^{u=k-2} T_{u}^{i})T_{k-2} \quad (6.17)
\]

\[
\text{treat}^\text{se}(\bar{T}_k) = 0.2T_{k-1}^i - 0.2\text{logistic}(\sum_{u=k-44}^{u=k-2} T_{u}^{i})T_{k-2}^i \quad (6.18)
\]

The Appendix B.2 provides the form of the logistic function; the value of this logistic function is close to 1 if the total amount of treatment, \( \sum_{u=k-2}^{u=k-44} T_{u}^{i} \), is large; it is close to 0 if the total amount of past treatment is close to 0. The coefficients of \( T_{k-1}^i \) imply that providing treatment at time \( k-1 \) reduces negative affect and increases self-efficacy. However, providing treatment many times during the last 2 weeks (e.g., \( \sum_{u=k-44}^{u=k-2} T_{u}^{i} \) is large) produces treatment fatigue, to the extent that providing treatment at time \( k-2 \) greatly reduces the effect of treatment at time \( k-1 \) on both negative affect and self-efficacy.

In virtual population A, \( w_{k}^{i} \) is set to 0 (i.e., there were no unmeasured momentary events that affected urge to smoke). In virtual population B, \( w_{k}^{i} \) is zero for 75\% of randomly selected time points; for the remaining time points it is uniformly distributed across a range of momentary positive and negative effects between -2.5 and 2.5. Thus the impact of the momentary events is sporadic as it is present at only 25\% of the time points.

### 6.6 Results of Scenario Approach

The dynamical model in equations (6.1), (6.2), (6.3) is fit to each of the 500 data sets from each of the two populations. For each population, the MPC algorithm is applied using each of the 500 fitted models resulting in 500 decision rule tables (500 adaptive intensive interventions). Next for each of the 500 adaptive interven-
tions, we record the smoking urge, negative affect and self-efficacy of a simulated individual under three scenarios:

1. if provided the adaptive intensive intervention,

2. if provided a full treatment intervention (treatment three times per day for all days) as a comparison, and

3. if not provided treatment.

Results are below.

6.6.1 Population A

The results are summarized in Table 6.2 (upper panel) and Figures 6.4 and 6.5. First note from Table 6.2 that using the adaptive intensive intervention results in the lowest average level of smoking urge, lowest average negative affect and highest average self-efficacy. Also note that when the adaptive intervention is used, treatment is provided on average only once every two days (mean treatments per day = 0.54). Furthermore, the mean differences between smoking urge under no treatment and smoking urge under the adaptive intervention or full treatment intervention are 0.28 ($se = 0.04$) and 0.20 ($se = 0.02$), respectively. Both the full treatment intervention and the adaptive intervention reduce smoking urge, but the adaptive intervention is more effective and uses less treatment. Figure 6.4 provides a more complete picture of the differences in average smoking urge between no treatment, adaptive treatment, and full treatment. Note that high values (i.e., greater improvement) are preferred in this Figure.

An adaptive intensive intervention should provide treatment only when needed when, in the absence of treatment, the smoking urge is expected to be large at subsequent times. To see this, consider 233 of the 500 individuals with a mean smoking urge above 3.04 in the absence of treatment. For this subgroup the mean differences between smoking urge for an individual under no treatment and this individuals smoking urge under the adaptive intervention or full treatment intervention are 0.34 ($se = 0.05$) and 0.21 ($se = 0.02$), respectively. In other words, the adaptive intervention performs better than the full treatment intervention among these higher-risk individuals. Also, treatment is provided, on average, about twice
every three days (\(0.65/\text{day}\)) more frequently than for the whole population, which received the treatment, on average, once per two days. Figure 6.5 provides the histograms for this subgroup.

**6.6.2 Population B**

Recall that the virtual participants in population B are similar to those in population A, except that they experience unmeasured momentary events that affect
the urge to smoke (these are the $w_k^j$s). Further recall that we fit the dynamical models given in equations (6.1), (6.2) and (6.3) which do not contain $w_k^j$s; that is, the data analyst is unaware of these momentary events.

The results of this simulation are summarized in Table 6.2 (lower panel). As before, the adaptive intensive intervention results in the lowest level of mean smoking urge, the lowest mean negative affect, and the highest mean self-efficacy. Thus even though the data analyst is unaware of the momentary events, the intervention algorithm produces improvements in smoking urge with less treatment. The average number of treatments in the adaptive intervention is slightly higher than in the scenario with no unmeasured momentary events, as can be seen by comparing the lower and upper panels of Table 6.2. Of course the algorithm cannot completely adjust for the unobserved momentary events, as can be seen by comparing the mean smoking urge reported in the lower panel of Table 6.2 to that in the upper panel.

<table>
<thead>
<tr>
<th>Variable</th>
<th>No Treatment</th>
<th>Full Treatment</th>
<th>Adaptive Intensive Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean(se)</td>
<td>SD</td>
<td>Mean(se)</td>
</tr>
<tr>
<td>Smoking Urge</td>
<td>3.04(0.027)</td>
<td>1.64</td>
<td>2.84(0.026)</td>
</tr>
<tr>
<td>Negative Affect</td>
<td>0.09(0.012)</td>
<td>1.03</td>
<td>-0.02(0.012)</td>
</tr>
<tr>
<td>Self-Efficacy</td>
<td>3.01(0.014)</td>
<td>0.94</td>
<td>3.18(0.014)</td>
</tr>
<tr>
<td>Treatments per day</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>No Treatment</th>
<th>Full Treatment</th>
<th>Adaptive Intensive Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking Urge</td>
<td>3.2(0.028)</td>
<td>2.01</td>
<td>3.03(0.027)</td>
</tr>
<tr>
<td>Negative Affect</td>
<td>0.1(0.011)</td>
<td>1.03</td>
<td>-0.02(0.011)</td>
</tr>
<tr>
<td>Self-Efficacy</td>
<td>3.03(0.013)</td>
<td>0.92</td>
<td>3.19(0.013)</td>
</tr>
<tr>
<td>Treatments per day</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

SD:Standard Deviation

In this setting, for a new individual, the mean differences between smoking urge under no treatment and smoking urge under the adaptive intervention or full treatment intervention are .27 ($se = .004$) and 0.17 ($se = .002$), respectively, which are very similar to those in population A. That is, both the full treatment intervention and the adaptive intervention reduce smoking urge; however, the adaptive intervention is more effective and uses less treatment. The histograms of the mean
difference in smoking urge between no treatment and adaptive treatment or the full treatment intervention are quite similar to that for population A and thus are not shown.

6.7 Results of Robust Approach

We simulate the application of our robust MPC algorithm to a patient. The parameters used are specified for this example as

- MPC window size: $K = 8$;
- Constraints in control: $T_k \in \{0, 1\}$ and
  \[
  T_{total} = \sum_{\eta=k-3}^{k+K-1} T_\eta \leq T_{total} = 10/3.
  \]

In other words, one either applies or does not apply treatment at each time point, and there is a constraint of a maximum of approximately one treatment per 3 sample times. Full treatment is defined as $T_k = 1$ for all $k$.

- Objective function to be minimized:
  \[
  \sum_{\eta=k+1}^{k+K} (su(\eta) - 2.2)^2,
  \]
  where the target value 2.2 was chosen so that one has a significant decrease in the mean smoking value.

- Bound on the 2-norm of the Gaussian noise: $\rho = 30$

- Initial conditions of the states and input:

<table>
<thead>
<tr>
<th>su(0)</th>
<th>su(1)</th>
<th>se(0)</th>
<th>se(1)</th>
<th>na(0)</th>
<th>na(1)</th>
<th>T(0)</th>
<th>T(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5212</td>
<td>0.3460</td>
<td>4.4280</td>
<td>4.3242</td>
<td>-0.3542</td>
<td>1.0817</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.3: Initial Condition of States and Input
6.7.1 Semi-Definite Programming (SDP) Result of a Patient

The optimization problem to be solved here is a mixed-integer, semi-definite convex problem. There are many ways to solve it, but in the simulations performed here, the requirement $T = T_k \in \{0, 1\}$ is relaxed to $T = T_k \in [0, 1]$ to use a general, semi-definite programming solver. The control applied is

$$T_k^{\text{applied}} = \text{round}(T_k).$$

The simulation was run for 150 time instances corresponding to a real time interval of 50 days. CVX is used with “sedumi” as semi definite programming solver [70] to the problem (6.10). The developed control law systematically improved smoking urge, since the treatment is only applied when needed. At the same time, our control law increased self-efficacy while reducing negative affect in order to reduce smoking urge. Table 6.4 represents a typical example of the results obtained.

<table>
<thead>
<tr>
<th></th>
<th>Adaptive $T$</th>
<th>Full $T$</th>
<th>No $T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean(su)</td>
<td>1.31</td>
<td>1.46</td>
<td>1.51</td>
</tr>
<tr>
<td>Mean(se)</td>
<td>3.8412</td>
<td>3.6424</td>
<td>3.4863</td>
</tr>
<tr>
<td>Mean(na)</td>
<td>0.3096</td>
<td>0.4571</td>
<td>0.5698</td>
</tr>
<tr>
<td>$|su|_2$</td>
<td>22.55</td>
<td>24.48</td>
<td>25.15</td>
</tr>
<tr>
<td>$\sum_{k=1}^{50} T_k$</td>
<td>43</td>
<td>150</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6.4: Intervention Performance of SDP

Table 6.4 shows the performance of the adaptive treatment design in terms of objective function, average values of smoking urge, self-efficacy, negative affect, and total amount of treatment that the adaptive intervention applied. The improvement in smoking urge is small, likely because this is a population of heavy smokers in [43, 44]. As a result, this improvement is extremely important at the beginning of the treatment process and leads to significant benefits over a long term period [43, 44]. Therefore, one can see that there is a significant improvement in smoking urge not only in terms of the average but also in the fact that a consistent decrease in urge is obtained. Figure 6.6 shows the smoking urge, self-efficacy and
negative-affect measure of a virtual patient under both full treatment and adaptive intervention. Additionally, asterisks indicate when treatment is provided by the adaptive treatment algorithm. Figure 6.7 depicts the exogenous sparse disturbance that is applied to this particular patient for this simulation.

As expected, the algorithm carefully “chooses” when to apply treatment. It is mainly applied when external perturbations lead to a significant increasing trend in smoking urge.

Finally, Figure 6.8 shows decrease in smoking urge. In Figure 6.8, $su_{Full\ Treatment} - su_{Adaptive\ Treatment}$ is depicted, where $su_{Full\ Treatment}$ is smoking urge measure under the full treatment and $su_{Adaptive\ Treatment}$ is smoking urge measured under the adaptive treatment. Although the full treatment performs better than the adaptive treatment in the beginning of the treatment process, in the long term the adaptive treatment works better because the algorithm applies the treatment only
when it is really needed, thus reducing treatment burden.

6.7.2 Mixed-Integer Semi-Definite Programming Result of an Individual

In this section, the results of a mixed integer, semi-defined program are presented. Instead of rounding \( \text{round}(T_k) \) for SDP, \( T_k \) is defined as binary variable. YALMIP [92] is used with “branch and bound” selected as the mixed-integer, semi-definite programming solver.

The constraint \( T_k \in \{0, 1\} \) in problem (6.9) can be replaced by defining \( T_k \) as a binary variable. Instead of the constraint on the continuous control variable \( T \), it is defined as a binary variable. Again the main model (3.2) will have randomly generated uncertainty/noise, which is the same uncertainty/noise used for SDP.
In order to compare the results, the problem is solved under the same parameters $\rho$ in B.10, $\theta$ in refeq:obj, $l$ in constraints in 6.9, exogenous input $w$, initial condition of states and noise. As a result, the semi-definite programming results and mixed-integer, semi-definite programming results are the same. (See Table 6.4 and Figure 6.6).

### 6.7.3 SDP Results of a Population

Since some of the parameters are given as random in the simulation, the algorithm is run 400 times to estimate the average improvement in smoking urge, self-efficacy, and negative-affect. The same parameters in Section 6.7, except that $l = 6$ in (6.9), are used to run the algorithm. Moreover, for the random parameters, the disturbances $\tilde{w}_k$ are generated uniformly such that $\tilde{w}_k \in \mathcal{W}_k$. The Gaussian disturbances are generated under a normal distribution such that $\tilde{e}^{su} \sim \mathcal{N}(0, 1.3)$.
Figure 6.9: Adaptive and Full Treatment Intervention Results

\( \varepsilon^{na} \sim \mathcal{N}(0, 0.78) \), and \( \varepsilon^{se} \sim \mathcal{N}(0, 0.55) \). For the initial conditions, \( su_0 \in U(3.6, 0) \), \( se_0 \in U(4, 0) \), and \( na_0 \in U(2, 0) \) are generated. Figure 6.9 depicts average improvement in smoking urge \( (su^{No \text{ Treatment}} - su^{Adaptive \ or \ Full \ Treatment}) \), self-efficacy \( (se^{Adaptive \ or \ Full \ Treatment} - se^{No \text{ Treatment}}) \), and negative-affect \( (na^{No \text{ Treatment}} - na^{Adaptive \ or \ Full \ Treatment}) \). It is shown that the adaptive treatment increases \( se \) while it decreases \( su \) and \( na \), even with less treatment. The average number of treatments provided is 54.6.

- \( su^{no}, se^{no}, na^{no} \) - Smoking urge, self-efficiency, and negative-affect without treatment
- \( su^{full}, se^{full}, na^{full} \) - Smoking urge, self-efficiency, and negative-affect under full treatment
- \( su^{adaptive}, se^{adaptive}, na^{adaptive} \) - Smoking urge, self-efficiency, and negative-affect with adaptive treatment
6.8 Conclusion

Adaptive intensive interventions are of increasing interest across a variety of health behavior change settings. They are particularly promising in settings in which behavior change is challenging and/or difficult to maintain. This is because the treatment can be adapted to individual progress, provided over longer periods of time, and strategically provided so as to prevent negative effects of participant burden. The possible applications within psychotherapy are numerous, including momentary interventions to encourage meditation practice or increase self-awareness (e.g., of stress) and using momentary GPS data to detect locations associated with greater risk (e.g., of alcohol use relapse) indicating immediate/increased need for contact from a trained professional.

Tools from engineering have the potential to transform how many behavioral interventions are designed and to maximize their effectiveness. Here we illustrated how a collaborative team of behavioral scientists and a control engineer can use ILD to design an adaptive intensive intervention. Our simulation study showed that when the dynamical model incorporates the main characteristics of the behavior, the resulting adaptive intervention performs well even in the presence of substantial uncertainties. That is, the control engineering approach is able to accommodate settings in which patient behavior and treatment response cannot be precisely modeled; rather, this approach takes into account a degree of (perhaps large) uncertainty in the design of an adaptive intensive intervention. This is often the case where control engineering currently is used [93]. One reason for the robust performance is the adaptive nature of the intervention. At each time point, the adaptive intervention algorithm uses both the individuals current status and a prediction of the individuals future status to recommend treatment. Moreover, again by looking at the current behavior of the individual, this algorithm detects when participants deviate from model predictions and adjusts recommendations accordingly.

Our simulated example is concerned about when to provide treatment, how often to provide treatment, and how to adapt these decisions to each individual as they progress; adaptive intensive interventions can also adapt the specific types of treatment for the individual. To assist in achieving the promise of adaptive inten-
sive interventions, methods that can use ILD to inform the design of high-quality, adaptive intensive interventions are needed. The control engineering approach described here provides a principled approach to designing adaptive interventions based on available ILD on the phenomenon under investigation.

From a clinical perspective, several challenges remain for implementing these engineering methods to design effective adaptive intensive interventions.

- First, dynamical models for modeling behavior change are required; work in this direction is underway e.g., [94, 95, 96]. These models must allow for discrete and semi-continuous variables, both prevalent across many areas of behavioral change.

- Second, an objective function as in equation (6.4) must be specified. There are many possible choices for the objective function; further research is needed to determine which objective functions are best suited for use in the development of each adaptive intensive intervention.

- Third, although exposure to treatment need not be randomized in the data, the approach used in this study assumes that there are no confounders (i.e., third variables) that explain an observed association between treatment and the outcome. Thus, as is always the case, researchers who wish to estimate treatment effects using observational data need to strive to collect important confounders.

- Fourth, in the example, data was collected at equidistant time instants, and the example had no missing data. However, we developed two types of algorithm in Chapters 4 and 5 to estimate missing data.

Lastly, the effectiveness of an adaptive intensive intervention developed via the use of control systems approach should be empirically tested in a randomized controlled clinical trial. Depending on the setting, a comparison group might be assigned to use decision rules developed using more traditional approaches or a comparison group might be assigned only EMA data collection or a standard care group.
Conclusion and Discussion

Designing robust adaptive treatment algorithms is a crucial task for improving the efficiency of treatments [6]. Clinicians and behavioral scientists usually attempt to alter the treatment sequence or treatment type according to the treatment response of a patient to obtain better results. However, this treatment adjustment might not occur frequently enough to address rapid behavioral changes. Furthermore, the proliferation of portable devices that collect patient information at a high frequency (intensive longitudinal data [ILD]) and enable timely treatment have opened the possibility for developing effective personalized interventions [97]. These kinds of technological tools are used as sensors or feedback to collect more patient information and to apply treatments in real time. On the other hand, clinicians do not have the necessary tools or systematic methods to use ILD and to design treatment sequences (adaptive treatment) to address behavioral problems. Therefore, in this dissertation, a robust adaptive treatment algorithm for behavioral problems is developed by utilizing the methods from control engineering and robust optimization.

In this dissertation, we argue that control engineering methods, such as feedback or adaptation and robust optimization, can provide a systematic way for designing adaptive treatments. Control engineering methods can be applied to create personalized robust behavioral interventions while increasing the effectiveness of the treatment and reducing treatment burden. Existing theory for adaptive treatment design is primarily qualitative and thus does not provide precise guidance regarding frequency, timing, and personalization of treatment. With the developed
adaptive intervention design algorithm, treatments can be adapted and re-adapted in response to an individual’s progress over a long period of time. This method holds promise for maintaining desired behavior in situations where controlling behavior is challenging due to complex dynamics.

In this thesis, we demonstrated two main steps in the development of adaptive treatments:

1-) From study/individual behavioral data, we used a “simple” dynamical model with uncertainties to model the behavioral data. In particular, several types of uncertainty with a simple model are proposed to model the behaviors. Then we show how one can use Lasso (in Chapter 3) to identify model parameters and quantify these uncertainties. Moreover the missing part of the behavioral data is not interpolated in the Lasso algorithm even though missing data is common in a behavioral context. Therefore, in this dissertation, we developed two types of identification algorithms for behavioral data that can simultaneously identify the model parameters and interpolate the incomplete parts of the data. A parsimonious system identification algorithm for single input single output (SISO) systems is proposed in Chapter 4. Then a more general identification algorithm for multi input multi output systems is developed in Chapter 5. These two algorithms can actually deal with the incomplete/noisy behavioral data from many different patients to approximate the model parameters and complete the data. The results of these algorithms are illustrated in Chapters 4 and 5.

2-) Once the model parameters and the quantification of uncertainties are identified, then we can propose a treatment algorithm based on the model and a person’s current status and context. This algorithm provides a treatment or sequence of treatments that will achieve the desired results for all (probable) values of the uncertainty. To have an effective robust adaptive treatment algorithm, we use the MPC approach in Chapter 6. In this Chapter, all the objectives of the treatment and a mathematical description of these objectives are explicitly explained. Additionally, two different methods are analyzed to solve the robust optimization problem in the controller algorithm. As a result, this new robust adaptive treatment algorithm performs better and utilizes less treatment than the regular treatment methods. Many results of the developed algorithm are also given in Chapter 6.
Mobile technology, such as the smart-phone, is central to this process of delivering interventions informed by feedback on a person’s past behavior and historical responses to different interventions. Therefore, even though we do not develop an implementation tool in this dissertation, health specialists need to use this kind of robust adaptive treatment algorithm to address real life problems. Therefore, we explain three possible implementation methods and steps in Chapter 6.

As a result, through collaborations with clinicians and behavioral scientists, engineers can provide alternative or supplemental interventions that are principled, cost-effective, and more efficacious than standard treatment, ultimately improving the mental and physical health of patients. Such work can advance clinical and behavioral interventions by, for example, reducing variability across time in symptoms by providing treatment when needed, adapting treatments based on patients previous and current states, and compensating for uncertainty that will always exist in real-world settings. These methods may lead to improved operationalization of performance criteria and then to intervention designs that target these criteria [98]. In sum, the new methods described here hold the potential to transform how behavioral interventions are conceived of, designed, and conducted so that treatments can be given intensively over time in a way that maximizes effectiveness while minimizing usage of resources and participant burden.

In terms of future work, first of all, one can use these identification and adaptive treatment design algorithms for real behavioral problems. For this reason, we will develop algorithms that use intensive longitudinal data on physical activity and sedentary behavior, to develop a cell phone-based, social-cognitive intervention aimed at increasing activity and reducing inactivity. Note that, an implementation tool such as mobile phone app is needed to implement the developed algorithms. Second, effort is being put in modeling behaviors since instead of using a simple model with uncertainties, another possible nonlinear models [93] such as time varying models [99, 100, 101] etc. might give better future prediction. Third, we will put our effort to develop more efficient numerical algorithms for solving the optimization problem (B.11). This will allow for the consideration of a larger future horizon and, hence, better performance.
Appendix A

Proof of Lemma 14

For simplicity, assume input does not have repeated poles and a system with impulse response $h(k) = \sum_{j=1}^{n} h_j p_n^k$ where $p_n \in \mathbb{D}_{\rho}$ and input $u(k) = \sum_{j=1}^{l} z_j p_u^k$ is given. Once the input persistently excites the linear system then the total response of LTI system can be expressed as a sum of zero-state/zero-input responses or natural/forced responses (A.1) [15, Chapter 2].

$$y(k) = \sum_{j=1}^{n} c_j p_n^k + \sum_{j=1}^{l} r_j p_u^k$$ (A.1)

where $p_n$ is non-repeated characteristic modes and $p_u$ is non-repeated noncharacteristic modes of the system for $k = 0, 1, 2, \ldots, N$. For repeated $p_u$, the term $\sum_{j=1}^{l} r_j p_u^k$ should be appropriately modified. Then using the information on Equation (A.1), the Equation (3.5) can be easily constructed as:

$$\mathcal{H}_y = \mathcal{H}_{y_n} + \mathcal{H}_{y_u} = \sum_{j=1}^{n} c_j \mathcal{H}_{p_n^j} + \sum_{j=1}^{l} r_j \mathcal{H}_{p_u^j}.$$ (A.2)

Additionally, it is proven in [67] that if the inputs themselves are the output of the LTI systems (inputs are the function of the some states of the system) then the rank cancellation occurs $\dim[\text{span}_{\text{row}}(X) \cap \text{span}_{\text{row}}(\mathcal{H}_u)] \neq 0$. Therefore no rank cancellation assumption 9 implies $p_n \neq p_u$. As a result, output block Hankel matrix $\mathcal{H}_y \in \mathbb{R}^{(r_m \nu) \times (N-r)}$ consists of the natural response (characteristic modes) $\mathcal{H}_{y_n} \in \mathbb{R}^{(r_m \nu) \times (N-r)}$.
\( \mathbb{R}^{(rm_0) \times (N-r)} \) and forced response (non-characteristic modes) \( \mathcal{H}_y u \in \mathbb{R}^{(rm_0) \times (N-r)} \) in (A.2).

Moreover the Hankel matrix of input is \( \mathcal{H}_u = \sum_{j=1}^l z_j \mathcal{H}_{p_{u_j}} \). It is easily seen that Hankel matrix of the forced response in Equation (A.2) is actually the scalar multiplication of input Hankel matrix. Then because of the properties of the null space, the null space of \( \mathcal{N}(\mathcal{H}_u) \) is equivalent to the null space of \( \mathcal{N}(\mathcal{H}_y u) \). As a result, \( \mathcal{H}_y \mathcal{H}_u^\perp \) in (A.2) only contains the natural response of the system:

\[
\mathcal{H}_y \mathcal{H}_u^\perp = \mathcal{H}_{y_n} \mathcal{H}_u^\perp + \mathcal{H}_{y_u} \mathcal{H}_u^\perp = \mathcal{H}_{y_n} \mathcal{H}_u^\perp
\]

This fact is also true for state matrix \( X \) since it is formed by the natural response (function of characteristic modes) and forced response (function of non-characteristic modes). This is similar to (A.1), \( X = f(p_n) + f(p_u) \) where \( \mathcal{H}_u^\perp \) eliminates the \( f(p_u) \) in \( X \) and \( X \mathcal{H}_u^\perp \) has only the natural response of the system.
Appendix B

Technical Details and Proofs of Theorems in Chapter 6

B.1 Technical Details Concerning the Objective Function and MPC

Note that in this section, we assume $w_k = 0$. In practice, three tuning parameters need to be specified (see below for discussion of how to set the values of these tuning parameters). First, often a tuning parameter is subtracted from each smoking urge; $\theta$ can be interpreted as the average level of smoking urge achievable by the particular treatment(s). Second, when there are many time points per individual, minimizing Equation (6.4) becomes computationally intractable. Usually one only considers a sum over $K$ time points,

$$\sum_{t=k+2}^{t=k+K} (su_t - \theta)^2$$

(B.1)

This introduces a second tuning parameter, $K$. Third, the total amount of treatment provided over a time period including recent past and the future time window of size $K$ is usually constrained, for instance by $T_{\text{total}}$ units; $T_{\text{total}}$ is the third tuning parameter. The rationale for adding this constraint is to reflect the concern that participants may experience treatment fatigue above the level of fatigue that can be represented by the models in Equations (6.2) and (6.3). Imposing
a constraint on the total amount of treatment allows us to accommodate long-term treatment fatigue without making these models more complicated (e.g., including terms for time lags greater than 2). Next note that for any \( t \) in Equation (B.1), \( su_t \) (see Equation (6.1)) depends on the error \( \varepsilon_k \). Moreover, it also depends on \( se, na \) and, given equations (6.2) and (6.3), this implies that \( su_t \) also depends on errors \( \varepsilon_u^{na}, \varepsilon_u^{se} \) for time points \( u < t \). These errors are unobservable. A robust approach is to first maximize Equation (B.1) over a grid of values for these normal error terms and then minimize the result to obtain the optimal level of treatment for individual \( i \) at time \( k \).

The adaptive intervention design algorithm is completely specified by Equation (B.1) and values for the tuning parameters, \( \theta, K, \) and \( T_{total} \). Values for the tuning parameters should be selected based on simulation studies completed prior to providing treatment; the values of the tuning parameters that produce the best adaptive intervention in simulation are then used in implementation. The guidelines that we used in our simulations to choose tuning parameter values were the following:

1. \( \theta \) was specified to reflect our goal of significantly reducing smoking urge. More precisely, our objective was approximately a 30 percent improvement over mean smoking urge without treatment (this translated to \( \theta = 2 \) since mean smoking urge without treatment and \( w_k \) is approximately 3; see table 6.2 for details. We did not choose \( \theta = 0 \) because this is not realistic for the population in this study. Careful analysis of the data showed that a zero smoking urge can only be achieved for short periods of time under full treatment and it is not sustainable in the long term.

2. We specified \( K=20 \) time points; the choice of \( K \) is usually limited by computational resources: larger values lead to more computationally complex optimization problems.

3. \( T_{total} = 26/3 \) (treatment approximately once per day) was chosen based on data analyses to determine the threshold for the number of consecutive treatments that led to a significant drop in effectiveness. Recall that, in the simulation study here, one has the possibility of providing treatment 3 times per day. The number \( T_{total} = 26/3 \) then means that we look at the number of
treatments provided in the last two days (6 instances of possible treatment) and at the window of future treatments that we are considering ($K = 20$) and restrict the treatment schedule in such a way as to avoid providing treatment more than one third of the time; i.e., on average, over this time interval we only provide one treatment per day.

In some cases, previous research or theoretical knowledge can guide selection of the tuning parameters, for example by specifying the maximum amount of treatment that participants are expected to tolerate. Further details of the algorithm can be found at http://dx.doi.org/10.1037/a0037736.supp.

B.2 Simulated Data for 150 Participants

The simulation models used to generate each of the simulated participants are:

\begin{equation}
su_k^i = 1.1 + .50su_{k-1}^i + .35su_{k-2}^i + .01na_{k-1}^i - .12se_{k-1}^i - .01se_{k-2}^i + \varepsilon_k^i \tag{B.2}
\end{equation}

\begin{equation}
na_k^i = .15 + .01su_{k-1}^i + .49na_{k-1}^i + .24na_{k-2}^i - .05se_{k-1}^i + \text{treat}^{na}(\hat{T}_k) + \varepsilon_{na}^i \tag{B.3}
\end{equation}

\begin{equation}
se_k^i = .52 - .01su_{k-1}^i - .02na_{k-1}^i + .49se_{k-1}^i + .36se_{k-2}^i + \text{treat}^{se}(\hat{T}_k) + \varepsilon_{se}^i \tag{B.4}
\end{equation}

where $\sigma(\varepsilon_k^i) = 1.2$, $\sigma(\varepsilon_{na}^i) = 0.6$, $\sigma(\varepsilon_{se}^i) = 0.3$ for $k = 1, \ldots, 150$ and $\{\varepsilon_1^i, \varepsilon_{na}^1, \varepsilon_{se}^1, \varepsilon_2^i, \varepsilon_{na}^2, \varepsilon_{se}^2, \ldots, \varepsilon_{150}^i, \varepsilon_{na}^{150}, \varepsilon_{se}^{150}\}$ are independent, normal, mean zero errors.

From Equation (B.2) we see that higher smoking urge, more negative affect, and lower self-efficacy at the previous time predicts higher smoking urge at time $k$. In Equation (B.3) we see that higher smoking urge, more negative affect, and lower self-efficacy at past time points predicts more negative affect at time $k$. Finally, in Equation (B.4), lower smoking urge, less negative affect, and higher self-efficacy at past time points predicts higher self-efficacy at time $k$.

The treatment terms $\text{treat}^{na}(\hat{T}_k)$ in Equation (B.3) and $\text{treat}^{se}(\hat{T}_k)$ in Equation
(B.4) are given by

\[
\text{treat}^{na}(\bar{T}_k) = -0.2T_{k-1} + 0.2\text{logistic}(\sum_{u=k-44}^{u=k-2} T_u^i)T_{k-2} \quad (\text{B.5})
\]

\[
\text{treat}^{se}(\bar{T}_k) = 0.2T_{k-1}^i - 0.2\text{logistic}(\sum_{u=k-44}^{u=k-2} T_u^i)T_{k-2}^i \quad (\text{B.6})
\]

See below for the form of the logistic function; the value of this logistic function is close to 1 if the total amount of treatment, \(\sum_{u=k-44}^{u=k-2} T_u^i\), is large; it is close to 0 if the total amount of past treatment is close to 0. The coefficients of \(T_{k-1}^i\) imply that providing treatment at time \(k-1\) reduces negative affect and increases self-efficacy. However, providing treatment many times during the last 2 weeks (e.g., \(\sum_{u=k-44}^{u=k-2} T_u^i\) is large) produces treatment fatigue, to the extent that providing treatment at time \(k-2\) greatly reduces the effect of treatment at time \(k-1\) on both negative affect and self-efficacy.

### B.2.1 Logistic Function

The function \textit{logistic} is a logistic function of the form \(\text{logistic}(x) = 1/(1+e^{-(x-17)})\), and a plot is given below.
B.3 Theorem in [3, 4, 5]

Theorem 16. [3, 4, 5] Assume the following problem is defined:

$$\phi(A, b, x) \triangleq \min_{x} \max_{\|\varepsilon\| \leq \rho} \|A(\varepsilon)x - b(\varepsilon)\|_2^2.$$  (B.7)

Given $A_0, \ldots, A_p \in \mathbb{R}^{n \times m}$, $b_0, \ldots, b_p \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}^p$, the following uncertainty structure can be defined:

$$A(\varepsilon) \triangleq A_0 + \sum_{i=1}^{p} \varepsilon_i A_i, \quad b(\varepsilon) \triangleq b_0 + \sum_{i=1}^{p} \varepsilon_i b_i.$$  (B.8)

Then for $\rho \geq 0$, define

$$F(x) \triangleq [A_1x - b_1 \ldots A_p x - b_p].$$  (B.9)

The solution of the structured robust least square problem (B.7) can be calculated by solving the following semi-definite problem

$$\min_{\tau, x, \lambda} \tau$$  (B.10)

subject to

$$\begin{bmatrix} \tau - \lambda \rho^2 & 0 & (A_0x - b_0)^T \\ 0 & \lambda I & F(x)^T \\ (A_0x - b_0) & F(x) & I \end{bmatrix} \geq 0.$$  

B.4 Proof of Theorem 14

Proof. We use the structure of $\mathcal{Y}_{k+1}$ in (6.6) to analyze the robustness of the optimization problem. Thus, $\mathcal{Y}_{k+1}$ in the objective function in (6.7) is replaced by the $\mathcal{Y}_{k+1}$ in equation (6.6) as:

$$\min_{T \in T} \max_{\|\varepsilon\| \leq \rho} \| (A_f - \theta + \tilde{A}Y_0 + \tilde{B}T_k + \tilde{D}\varepsilon_k + \tilde{E}\tilde{w}_k)X \|_2^2.$$  (B.11)
Given objective function in (B.11), we can define vectors and matrix in (B.8) as:

\[ A_0 = \tilde{B}, A_i = 0, b_0(\tilde{w}_k) = A_f - \theta + \tilde{A}Y_0 + \tilde{E}\tilde{w}_k, \text{and } b_i = \tilde{D}_{:,i} \]

Then, for fixed \( \tilde{T}_k \) and \( \tilde{w}_k \), worst case residual is defined by using the methods in [5, Section 4] as:

\[ r_s(A, b, \tilde{T}_k, \tilde{w}_k)^2 = \max_{\|\varepsilon\|_2 \leq \rho} \|\tilde{B}\tilde{T}_k - b(\varepsilon, \tilde{w}_k)\|_2, \]

and from (B.9), define:

\[ M = F^T F, \quad g(\tilde{T}_k, \tilde{w}_k) = F^T (\tilde{B}\tilde{T}_k - b_0), \quad h(\tilde{T}_k, \tilde{w}_k) = \|\tilde{B}\tilde{T}_k - b_0\|_2^2 \]

then without loss of generality assuming \( \rho = 1 \) yields,

\[ r_s(A, b, \tilde{T}_k, \tilde{w}_k)^2 = \max_{\|\varepsilon\|_2 \leq 1} \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}^T \begin{bmatrix} h(\tilde{T}_k, \tilde{w}_k) & g(\tilde{T}_k, \tilde{w}_k)^T \\ g(\tilde{T}_k, \tilde{w}_k) & M \end{bmatrix} \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix} \leq \tau, \quad \text{(B.12)} \]

Given \( \tau \geq 0 \), using the S-procedure [5, Lemma 2.1], equation (B.12) can be converted as:

\[ \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}^T \begin{bmatrix} h(\tilde{T}_k, \tilde{w}_k) & g(\tilde{T}_k, \tilde{w}_k)^T \\ g(\tilde{T}_k, \tilde{w}_k) & M \end{bmatrix} \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix} \leq 0 \quad \text{for every } \varepsilon \in \mathbb{R}^p \quad \text{and fixed } \tilde{T}_k \text{ and } \tilde{w}_k. \quad \text{(B.13)} \]

for every \( \varepsilon, \varepsilon^T \varepsilon \leq 1 \) if and only if there exist a scalar \( \lambda \geq 0 \) such that

\[ \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}^T \begin{bmatrix} \tau - \lambda - h(\tilde{T}_k, \tilde{w}_k) & -g(\tilde{T}_k, \tilde{w}_k)^T \\ -g(\tilde{T}_k, \tilde{w}_k) & \lambda I - M \end{bmatrix} \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix} \geq 0 \quad \text{for every } \varepsilon \in \mathbb{R}^p \quad \text{and fixed } \tilde{T}_k \text{ and } \tilde{w}_k. \quad \text{(B.14)} \]

Using the fact that \( \lambda \geq 0 \) is implied by \( \lambda I \geq M \), then above condition can be rewritten as:

\[ \mathcal{F}(\tau, \lambda, \tilde{T}_k, \tilde{w}_k) = \begin{bmatrix} \tau - \lambda - h(\tilde{T}_k, \tilde{w}_k) & -g(\tilde{T}_k, \tilde{w}_k)^T \\ -g(\tilde{T}_k, \tilde{w}_k) & \lambda I - M \end{bmatrix} \geq 0. \]
Then the upper bound on residual such that \( r_s(A, b, \tilde{T}_k, \tilde{w}_k)^2 \leq \tau \) can be obtained by solving following optimization problem [5, Theorem 4.1]:

\[
\tau^* = \min_{\tau, \lambda} \tau \\
\text{s.t.} \\
\begin{bmatrix}
\tau - \lambda - h(\tilde{T}_k, \tilde{w}_k) & -g(\tilde{T}_k, \tilde{w}_k)^T \\
-g(\tilde{T}_k, \tilde{w}_k) & \lambda I - M
\end{bmatrix} \geq 0 \text{ for all } \tilde{w}_k \in W_k, \quad (B.15)
\]

In addition to that, instead of checking the all \( \tilde{w}_k \in W_k \) in LMI constraint (B.15), one can use the result in [102]. This result states that all the members of a polytope of matrices are positive semidefinite if and only if all extremes of the polytope are positive semidefinite. Furthermore, the set \( W_k \) is a union of sets bounded by polytopes. Using the same reasoning as in [102], it suffices for the LMI to be satisfied at the extremes of each of the polytopes. Therefore we need to check all extremes of these polytopes. LMI in (B.15) is satisfied in the convex hull of these extremes. It is just a consequence of the fact that is an LMI is satisfied in a set of points, it is also satisfied in the convex hull of these points. Then we can have following result:

\[
\tau^* = \min_{\tau, \lambda} \tau \\
\text{s.t.} \\
\begin{bmatrix}
\tau - \lambda - h(\tilde{T}_k, \tilde{w}_k) & -g(\tilde{T}_k, \tilde{w}_k)^T \\
-g(\tilde{T}_k, \tilde{w}_k) & \lambda I - M
\end{bmatrix} \geq 0 \text{ for all } \tilde{w}_k \in W_{ext} \text{ and for fixed } \tilde{T}_k, \quad (B.16)
\]

where \( W_{ext} \) are the extremes of the polytopes whose union is \( W_k \). Hence, \( \tau^* \) is an upper bound on the robust performance for fixed \( \tilde{T}_k \).

For the optimality, assume there is only one extreme \( \tilde{w}^* \in W_{ext} \) such that one of the eigenvalues of LMI in (B.16) is equal to zero for some \( \tau^* \) and \( \lambda^* \) and LMI is positive definite for the rest of the extremes. Then if the \( \tau^* \) is not optimum, we can perturb \( \tau^* \) or decrease it for this specific \( \tilde{w}^* \in W_{ext} \) while keeping the LMI still positive definite for the rest of the extreme points. This contradicts the assumption \( \tau^* \) is optimum. As a result, this optimization problem gives the optimum \( \tau^* \) if only
one of the extremes is an active constraint.

Thus, for every fixed $T_k$ and for all $\tilde{w}_k \in \mathcal{W}_{ext}$, the results in [5, Theorem 4.2] or Theorem 8 can be used to convert problem (B.16) to (B.17) for ellipsoid uncertainty $\|\tilde{\varepsilon}_k\|_2 \leq \rho$ and $\tilde{w}_k \in \mathcal{W}_k$. Finally, problem in (B.16) is modified as follows:

\[
\begin{array}{ll}
\min_{\tau, T_k \in T, \lambda} & \tau \\
\text{subject to} & \begin{bmatrix}
\tau - \lambda \rho^2 & 0 & ((A_f - \theta + \tilde{A}Y_0 + \tilde{B}T_k + \tilde{D}\tilde{\varepsilon}_k + \tilde{w}_kX)^T \\
0 & \lambda I & (\tilde{D}X)^T \\
(A_f - \theta + \tilde{A}Y_0 + \tilde{B}T_k + \tilde{D}\tilde{\varepsilon}_k + \tilde{w}_k)X & \tilde{D}X & I
\end{bmatrix} \geq 0
\end{array}
\] for all $\tilde{w}_k \in \mathcal{W}_{ext}$ where again $\mathcal{W}_{ext}$ are the extremes of the polytopes whose union is $\mathcal{W}_k$. \hfill \Box

### B.5 Proof of Theorem 15

**Proof.** The first two cases ($\alpha \geq \gamma$ and $n_{nz}\alpha \leq \gamma$) are immediate since they correspond to the cases where either the $\ell_1$ or the $\ell_\infty$ are the only binding constraints. Hence, in this proof, we concentrate on the third case ($\gamma < n_{nz}\alpha < n_{nz}\gamma$).

Note that, given the symmetry of the problem, we can concentrate on the subset of the elements of $\mathcal{W}_k$ that are positive and satisfy

$$w_{i+1} \leq w_i \text{ for all } i = k : k + K - 1, \text{ and } w_i \in \tilde{w}_k.$$ All other extremes will be obtained from this by permutations and sign changes of the entries. Recall that in the case we are considering,

$$\gamma < n_{nz}\alpha < n_{nz}\gamma,$$
and the elements of vector $\tilde{w}_k$ are in nonincreasing order. Therefore, the extreme of the set is given by “pushing” as many of the first few elements as possible to their maximum value. Given that $\|\tilde{w}_k\|_\infty \leq \alpha$ and $\|\tilde{w}_k\|_1 \leq \gamma$, one can only have
the first

\[ n_+ = \left\lfloor \frac{\gamma}{\alpha} \right\rfloor \]

elements of vector \( \tilde{w}_k \) equal to \( \alpha \). To reach the extreme, and recalling that the entries of the vector \( w \) are in nonincreasing order, the \((n_+ + 1)\)-th element of the vector must be at its maximum value. Given that \( \|\tilde{w}_k\|_1 \leq \gamma \) and the first \( N_+ \) terms are equal to \( \alpha \),

\[ w_{n_+ + 1} \leq \gamma - n_+ \alpha. \]

Hence, for vectors with entries in decreasing order, the extreme of the set is attained by a vector of the form

\[
\begin{bmatrix}
\alpha \cdots \alpha \\
\vdots \\
\alpha \cdots \alpha \\
n_+ \\
K - n_+ - 1
\end{bmatrix}
\]
Bibliography


[70] Michael Grant and Stephen Boyd., “Matlab software for disciplined convex programming, version 2.0 beta.”


Vita
Korkut Bekiroglu

Education:

- 2010 – 2015: The Pennsylvania State Univ. (PSU), Electrical Eng., University Park, USA
  PhD, supervised by Professor Constantino Lagoa
  PhD Minor, The Pennsylvania State Univ. (PSU), Industrial and Manufacturing Eng., University Park, USA

- 2008 – 2010: Northeastern University, Electrical and Computer Eng., Boston, MA, USA
  MSc, supervised by Professors Mario Sznaier and Bahram Shafai

- 2000 – 2005: Black Sea Technical University, Electrical and Electronics Eng., Trabzon, Turkey
  BS, senior project supervised by Professor Sefa Akpinar

- 1996 – 2000: Fatih High School, Kahramanmaras, Turkey

Experience:

- September 2010 – August 2015: Research Assistant and Grader in PSU
  RA: Robust Machine Intelligence and Control Lab.
    - “Designing Robust Interventions”
    - Built a complete framework for robust adaptive behavioral treatment
    - Use robust optimization methods to design robust MPC for individualized treatment algorithms
    - Develop new system identification algorithms for behavioral data.
  Grader: EE 380 Linear Control Systems and EE 580 Linear Control Systems (Electrical Eng.)

- September 2008 – August 2010: RA in Electrical and Electronics Eng., Northeastern University
  RA: Robust System Lab.
    - “Computer Vision Based Control of Autonomous Aircraft”
    - Built an approximated nonlinear model for an unmanned air vehicle.
    - Design a controller for this air vehicle with available approximated model.