2020

Data analytics and optimization methods in biomedical systems: from microbes to humans

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

DATA ANALYTICS AND OPTIMIZATION METHODS IN BIOMEDICAL SYSTEMS: FROM MICROBES TO HUMANS

by

TAIYAO WANG

B.S., China Agricultural University, 2011
M.S., University of Chinese Academy of Sciences, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader

Ioannis Ch. Paschalidis, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

Second Reader

Christos G. Cassandras, PhD
Distinguished Professor of Engineering
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Third Reader

Daniel Segrè, PhD
Professor of Biology
Professor of Biomedical Engineering
Professor of Physics

Fourth Reader

Pirooz Vakili, PhD
Associate Professor of Mechanical Engineering
Associate Professor of Systems Engineering

Essentially, all models are wrong, but some are useful.

George E. P. Box (1919 – 2013)

Acknowledgments

First of all, I wish to express my gratitude to my advisor, Professor Ioannis Paschalidis, for his continuous support during my Ph.D. study and research, his patience, motivation, enthusiasm, and immense knowledge. Yannis is known for many contributions to science. But for me, his real impact was as a mentor to students. A Ph.D. is hard. But a good supervisor makes it much easier (White, 2018). I have been working with him for 5 years, and I learned a lot from him. He has excellent time management skills and he always shows his passion for new challenges in research. I truly appreciate his consistent support, patience, and encouragement throughout my Ph.D. study. His academic advice was essential to the completion of this dissertation.

I would like to thank the committee members, Professor Segrè, Professor Cassandras and Professor Vakili for reading the draft of my dissertation and providing valuable comments that improved the quality of this dissertation. Professor Segrè was also influential in co-advising the work on the metabolic network project that is part of the dissertation.

I would like to thank all my collaborators during my Ph.D. time. It was a great pleasure to work with all of them. I would like to sincerely thank Tingting Xu, Dr. Theodora S. Brisimi, Dr. Wuyang Dai and Professor William G. Adams for collaborations on predicting chronic disease hospitalizations from electronic health records. I would like to sincerely thank Professor George Kasotakis, Professor Dimitris Bertsimas, Michael Lingzhi Li, and Dr. Henghui Zhu for collaborations and useful discussions on 30-day hospital readmissions after general surgery. I would like to sincerely thank Dr. Kyle Hansen, Dr. Joshua Loving, Dr. Eran Simhon, and Dr. Helen van Aggelen for collaborations on predicting Antimicrobial Resistance in the Intensive Care Unit. I would like to sincerely thank Professor Daniel Segrè, Meghan Thommes, Dr. Joshua Goldford, Dr. Qi Zhao for collaborations and discussions on

community-level Flux Balance Analysis. I would like to sincerely thank Professor Lauren A. Wise, Professor Shruthi Mahalingaiah, Professor Elizabeth Hatch, and Sydney Willis for collaborations and useful discussions on predicting female infertility problems. I would like to sincerely thank Dr. Ruidi Chen for discussions on Joint Clustering and Regression. Finally, I would like to thank Dr. Jing Zhang for collaborations on Kaggle Competitions.

Help and advice from SE and CISE staff have been invaluable. Many thanks to Elizabeth Flagg, Ruth Mason, Cheryl Stewart, Gabriella McNevin, Denise Joseph, Maureen Stanton and Christina Polyzos for all their help during my Ph.D. study.

I wish to acknowledge the support received from my lab mates, SE/ECE Alumni and friends. We had many interesting and good-spirited discussions related to life and research. Last, but not the least, I would like to express special thanks to my family for their understanding, support and love during my Ph.D. study. Their encouragement was in the end what made this dissertation possible. We would like to acknowledge support by the NSF under grants DMS-1664644, CNS-1645681, IIS-1237022 and IIS-1914792, by the ONR under MURI grant N00014-16-1-2832, by the ARO under grant W911NF-12-1-0390, by the NIH under grants R01-GM089978 and 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, by the Boston University Digital Health Initiative, and by the Boston University Center for Information and Systems Engineering.

Taiyao Wang

Division of Systems Engineering

DATA ANALYTICS AND OPTIMIZATION METHODS IN BIOMEDICAL SYSTEMS: FROM MICROBES TO HUMANS

TAIYAO WANG

Boston University, College of Engineering, 2020

Major Professor: Ioannis Ch. Paschalidis
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

ABSTRACT

Data analytics and optimization theory are well-developed techniques to describe, predict and optimize real-world systems, and they have been widely used in engineering and science. This dissertation focuses on applications in biomedical systems, ranging from the scale of microbial communities to problems relating to human disease and health care.

Starting from the microbial level, the first problem considered is to design metabolic division of labor in microbial communities. Given a number of microbial species living in a community, the starting point of the analysis is a list of all metabolic reactions present in the community, expressed in terms of the metabolite proportions involved in each reaction. Leveraging tools from Flux Balance Analysis (FBA), the problem is formulated as a Mixed Integer Program (MIP) and new methods are developed to solve large scale instances. The strategies found reveal a large space of nuanced and non-intuitive metabolic division of labor opportunities, including, for example, split-

ting the Tricarboxylic Acid Cycle (TCA) cycle into two separate halves. More broadly, the landscape of possible 1-, 2-, and 3-strain solutions is systematically mapped at increasingly tight constraints on the number of allowed reactions.

The second problem addressed involves the prediction and prevention of short-term (30-day) hospital re-admissions. To develop predictive models, a variety of classification algorithms are adapted and coupled with robust (regularized) learning and heuristic feature selection approaches. Using real, large datasets, these methods are shown to reliably predict re-admissions of patients undergoing general surgery, within 30-days of discharge. Beyond predictions, a novel prescriptive method is developed that computes specific control actions with the effect of altering the outcome. This method, termed Prescriptive Support Vector Machines (PSVM), is based on an underlying SVM classifier. Applied to the hospital re-admission data, it is shown to reduce 30-day re-admissions after surgery through better control of the patient's pre-operative condition. Specifically, using the new method the patient's pre-operative hematocrit is regulated through limited blood transfusion.

In the last problem in this dissertation, a framework for parameter estimation in Regularized Mixed Linear Regression (MLR) problems is developed. In the specific MLR setting considered, training data are generated from a mixture of distinct linear models (or clusters) and the task is to identify the corresponding coefficient vectors. The problem is formulated as a Mixed Integer Program (MIP) subject to regularization constraints on the coefficient vectors. A number of results on the convergence of parameter estimates for MLR are established. In addition, experimental prediction results are presented comparing the prediction algorithm with mean absolute error regression and random forest regression, in terms of both accuracy and interpretability.

Contents

List of Tables

List of Figures

List of Abbreviations

ACS	.....	American College of Surgeons
ASA	.....	American Society of Anesthesiology
AUC	.....	Area Under the ROC Curve
BMC	.....	Boston Medical Center
BUN	.....	Blood urea nitrogen
CDC	.....	Centers for Disease Control and Prevention
CPT	.....	Current Procedural Terminology
DAG	.....	Directed acyclic graph
EHR	.....	Electronic Health Records
EMR	.....	Electronic Medical Records
FBA	.....	Flux Balance Analysis
GBM	.....	Gradient Boosting Machine
HCT	.....	Hematocrit
ICD9	.....	International Classification of Diseases-Ninth Revision
INR	.....	International Normalized Ratio
JCC	.....	Joint Clustering and Classification
LR	.....	Logistic Regression
MILP	.....	Mixed Integer Linear Program
MIP	.....	Mixed Integer Program
NN	.....	Neural networks
NSQIP	.....	National Surgical Quality Improvement Program
PATOS	.....	Present at the time of surgery
PT	.....	Prothrombin time
PTT	.....	Partial thromboplastin time
RBF	.....	Radial Basis Function
RF	.....	Random forests
ROC	.....	Receiver Operating Characteristic Curve
SGOT	.....	Serum glutamic-oxaloacetic transaminase
SLSVM	.....	Sparse Linear SVM
SSI	.....	Surgical Site Infection
SVM	.....	Support Vector Machine
WBC	.....	White blood cell count

Chapter 1

Introduction

Data analytics and optimization theory are important techniques to discover, interpret, predict and optimize real-world problems and they have been widely used in business, biology, engineering, management science and many other disciplines. In this thesis, we focus on applications in biomedical systems ranging from the smaller scale of microbial communities to the problems arising at the disease, human level and health care systems.

1.1 Context and Related Work for Problems in Microbial Communities

Each microbial cell harbors a finite number of metabolic functions, shaped by natural selection into a tradeoff between the cost of carrying and expressing each gene, and the usefulness of such genes under different environments. This tradeoff is considered to be one of the possible sources of diversity in natural microbial communities, giving rise to metabolically differentiated groups of microbes rather than individual superorganism. The emergence of metabolically differentiated subpopulations has also been documented to occur from isogenic populations in a fixed environment (Elena and Lenski, 2003; Ferea et al., 1999; van Gestel et al., 2015; Friesen et al., 2004; Vlamakis et al., 2015; Le Gac et al., 2008; Rosenthal et al., 2018; Rozen and Lenski, 2000; Rozen et al., 2005; Spencer et al., 2007; Spencer et al., 2008). The viability of populations of metabolically differentiated strains or species is often enabled by

the exchange of metabolites (Elena and Lenski, 2003; Rozen and Lenski, 2000; Rozen et al., 2005; Embree et al., 2015; Rosenzweig et al., 1994; Treves et al., 1998). For example, initially clonal populations of E. coli evolved on minimal glucose medium have been observed to give rise to a specialized subpopulation of cells that use the acetate secreted upon glucose fermentation (Rozen and Lenski, 2000; Rosenzweig et al., 1994; Treves et al., 1998). In turn, obligate metabolic interdependencies (such as mutualism) are believed to contribute to the high prevalence of unculturability among natural microbial strains (Pande et al., 2014; Zelezniak et al., 2015).

An exciting new strategy to study microbial interdependencies is the construction (or evolution) of artificial microbial consortia specifically designed to display obligate mutualism. Current approaches to building synthetic communities of interacting microbes have so far mainly relied on biochemical intuition about simple genetic perturbations that would cause organisms to engage in obligate cross-feeding, where one strain is unable to synthesize an essential metabolite (e.g. an amino acid) that is supplied via overproduction or leakage by another strain (Hoek et al., 2016; Kerner et al., 2012; Wintermute and Silver, 2010; Mee et al., 2014; Shou et al., 2007). This ensures that the two strains require each other's presence in order to grow. These metabolic interactions are believed to help maintain the diversity and stability of natural microbial communities, as well as contribute to the emergent capabilities of communities to accomplish metabolically-intensive tasks (Klitgord and Segrè, 2010b; Klitgord and Segrè, 2011; Embree et al., 2015; Tsoi et al., 2018). While interesting and valuable, these strategies explore only a small portion of the very large and complex space of possible environmental and organismal modifications: in principle, organisms may have the potential to display complex cross-feeding strategies for multiple metabolites simultaneously, in an environment-dependent manner. In fact, given the complexity of metabolism and its evolutionary history, it is pos-

sible that naturally evolved cross-feeding strategies may involve complex metabolic mutualism beyond single amino acid exchanges. In particular, loss of functions in one organism due to compensation by others has been hypothesized to be widespread (Morris et al., 2012), and may involve multiple genes and complex pathway architectures (Zomorrodi and Segrè, 2017) Furthermore, in the engineering of consortia for specific metabolic engineering tasks, exploring this larger space of possibilities may open up novel strategies for bioproduction.

Exploring the space of possible paths for metabolic differentiation leading to obligate mutualism is a combinatorially difficult problem. While future elaborations of existing methods for high throughout genetic modifications (e.g., MAGE (Vlamakis et al., 2015)) may enable a systematic exploration of this space, computational models can provide a preliminary assessment of the landscape of possible strategies and of how these strategies depend on different constraints on metabolic network complexity. Constraint-based models of metabolic networks, such as Flux Balance Analysis (FBA) (O'Brien et al., 2015) , can specifically be leveraged to ask questions that cannot be easily addressed experimentally. FBA represents metabolism as a set of biochemical reactions, which are inferred from genome annotations and literature curation, and views cellular metabolism as a resource-allocation problem. Given a set of biochemical, thermodynamic, and environmental constraints, FBA uses Linear Program to determine a metabolic network's optimal flux distribution to achieve a certain objective. Typically, the objective function is to maximize growth, so FBA determines how a cell should optimally allocate nutrients based on its environment and biochemical capabilities so that the growth rate is maximized. FBA has also been increasingly used to study metabolic interactions in microbial consortia.

1.2 A Context for Problems in Healthcare

An estimated 3 trillion dollars annually is spent on health care in the U.S. alone, a value that exceeds 17% of the country's Gross Domestic Product (GDP), which, as a fraction of GDP, exceeds by 50% the next-highest spender (France) among the 13 high-income Organization for Economic Cooperation and Development countries (Squires and Anderson, 2015). The Centers for Medicare and Medicaid Services have identified hospital readmissions, typically defined as an additional admission to address the same issue within 30 days of discharge, as an important - and potentially preventable - source of excessive resource utilization and increased cost of care (Centers for Medicare & Medicaid Services, 2018).

An analysis of 2005 Medicare claims demonstrated that about 75% of 30-day readmissions, representing about 12 billion dollars in Medicare spending, were potentially preventable (James, 2013). As a result, through the enactment of the Readmissions Reduction Program subsection of the Affordable Care Act in 2012, readmissions are increasingly deemed as a care quality metric, and their reduction is mandated for certain diseases to minimize reimbursement fines by insurers, that may be as high as 3% of all Medicare payments (Centers for Medicare & Medicaid Services, 2018). Healthcare organizations have since allocated resources towards reducing unplanned readmissions as a way of improving quality of care, without unnecessarily prolonging hospital length of stay. While the measures have so far concentrated on readmissions after hospital stays secondary to most medical conditions (such as acute myocardial infarction, congestive heart failure, pneumonia, chronic obstructive pulmonary disease exacerbations) and commonly performed orthopedic procedures (such as hip and knee arthroplasty), the list is likely to expand and potentially include several commonly performed general surgery procedures.

In light of these changes, many surgical departments around the country have

started to closely monitor their readmission rates, and are putting in place processes aimed at reducing them. Several authors have sought to determine the most common causes of readmissions after a general surgical procedure, and most appear to relate to pre-existing conditions (Gonzalez et al., 2016; Tosoian et al., 2015; Petrigliano et al., 2014; Kimbrough et al., 2014) or development of complications after surgery (Petrigliano et al., 2014; Kimbrough et al., 2014), at least in the setting of elective procedures, when comorbidities are routinely attempted to be optimized, and perioperative measures aimed at limiting surgical site infections and venous thromboembolic events among others, are commonly undertaken. However, despite the identification of risk factors that may predispose patients to readmissions, few efforts have been made to generate models that take into account several parameters towards the identification of subjects at risk for readmission, and even fewer interventions have been shown to yield satisfactory readmission rate reductions (Chakravarthy et al., 2018; McHugh et al., 2017).

Prediction is the first step towards prescription and prevention. It allows the health systems to target individuals with the highest risk and employ the limited health resources more effectively and efficiently. Prescriptive analytics uses predictive models to recommend actions aimed at improving future outcomes.

1.3 Parameter Estimates for Regularized Mixed Linear Regression Models

Motivated by the predictive modeling discussed in Section 1.2, we considered more general models that extend beyond the commonly assumed linear mapping from variables to responses. In particular, we consider "piecewise" linear models, where input data may belong to multiple subsets and in each subset there exists a linear relationship between input variables and responses.

Mixed Linear Regression (MLR) (Yi et al., 2014; Zhong et al., 2016) is also known as mixtures of linear regressions (Chaganty and Liang, 2013) or cluster-wise linear regression (Park et al., 2017). It involves the identification of two or more linear regression models from unlabeled samples generated from an unknown mixture of these models. This can be seen as a joint clustering and regression problem. The problem is related to the identification of hybrid and switched linear systems (Paoletti et al., 2007; Vidal, 2008) and has many diverse applications. There are many applications with MLR models in marketing (DeSarbo and Cron, 1988), health insurance claims (Gitman et al., 2018), rainfall prediction (Bagirov et al., 2017), and pavement condition prediction (Luo and Chou, 2006). In this study, we focus on the fundamental problem of establishing strong consistency of parameter estimates, i.e., establishing that the estimated parameters converge to their true values as the number of the training samples grows. Furthermore, we study how to apply MLR for large scale prediction problems in reality.

1.4 Our Contributions of the Thesis

1.4.1 Division Of Labor In Microbial Communities

First, we study designing metabolic division of labor in microbial communities. We develop a novel Mixed Integer Linear Program (MILP) based approach to optimally allocate the metabolic functions among organisms in a microbial ecosystem. We propose heuristic methods to speed up the branch and bound algorithm for the class of division of labor MILP problems of interest. We test the method on both a community composed of E. coli core models and a community composed of E. coli iJR904 models. In both cases, the method helps us identify the individual metabolic network topology and elucidate the interaction between species in the microbial community. For instance, an interesting outcome obtained for subnetworks under some constraints

of transport reactions and intracellular reactions, was the discovery of a metabolic strategy in which each strain performs half of the tricarboxylic acid (TCA) cycle. It provides a new platform for the rational design of organisms and communities towards future synthetic ecology applications.

1.4.2 Predictive Analytics for Hospital Readmissions

In the second part of the dissertation, we extend our scope from microbial ecosystems to the individual healthcare and study predictive analytics for 30-day hospital readmissions after general surgery discharge. We employ and develop supervised learning methods from the field of machine learning to learn efficiently and effectively from large datasets. In order to obtain interpretable models, we use feature engineering (classical feature pre-processing, missing values imputation by regression models), variable selection by statistical tests and recursive feature elimination in the context of Sparse Linear SVM (SLSVM) and Joint Clustering and Classification (JCC). Our methods are validated by using the actual clinical data from our tertiary urban academic medical center. In particular, the data come from the Boston Medical Center - the largest safety-net hospital in New England with over 24,000 admissions annually; with its 13 affiliated Community Health Centers, it provides care for about 30% of Boston residents. The data from BMC span between 02/2010 and 12/2013.

1.4.3 Prescriptive Analytics for Preventing Hospital Readmissions

In the third part of the thesis, we study prescriptive analytics for 30-day hospital readmissions after general surgery discharge based on predictive analytics. We propose a new method, Prescriptive Support Vector Machines (PSVM), and evaluate the prescriptive analytics results by different predictive machine learning methods. PSVM is a prescriptive method that is based on SLSVM and JCC. After optimal hyperplanes are generated by JCC, the prescription decisions are generated through

"moving" the readmitted patient from the positive side to the negative side or the non-readmitted side of the separating hyperplane. Our methods are validated by using the National Surgical Quality Improvement Program (NSQIP) dataset during 2014 collected by the American College of Surgeons (ACS), which included information on 722,101 surgeries.

1.4.4 Parameter Estimates for Regularized Mixed Linear Regression Models

As our fourth contribution, we introduce a general MIP formulation for MLR subject to norm-based regularization constraints. The formulation is general enough to include regularization constraints on the regression coefficients. We study the consistency conditions of parameter estimates for MLR using the MIP formulation in both a noiseless and a noisy case. We propose identifiable conditions and establish that optimal solutions of the MIP converge almost surely (rather than w.h.p.) to the true parameters in the noiseless case as the sample size increases. Subject to cluster separability assumptions, we also establish that MIP solutions can identify the proper cluster for each given sample. To the best of our knowledge, our study is the first to study strong consistency of parameter estimates for MLR under general noise conditions and general feature conditions rather than convergence with high probability. For the special case of a single cluster, we show that the MIP solution converges to the true parameter vector in the presence of noise satisfying a martingale difference assumption (Lai et al., 1982; Chow and Teicher, 2012; Chen and Guo, 2012). For multiple clusters in the presence of noise, we not only derive the convergence conditions, i.e., a stronger identifiable condition and a cluster consistency condition, but also provide a counterexample, suggesting that one can not in general recover the true parameters if the cluster consistency condition is violated. Besides the convergence results, we propose a novel method to apply MLR for large scale prediction

problems. We present experimental prediction results and compare our prediction algorithm against mean absolute error regression and Random Forest regression in terms of both accuracy and interpretability.

1.5 Bibliographic Notes

Large parts of the thesis appears in published or working research papers:(Wang and Paschalidis, 2019a; Wang and Paschalidis, 2019b; Thommes et al., 2019; Thommes et al., 2018; Zhao et al., 2016; Xu et al., 2016; Brisimi et al., 2019; Brisimi et al., 2018a).

Notational conventions: All vectors are column vectors. For economy of space, we write $\mathbf{x} = \left(x_1, \ldots, x_{\dim(\mathbf{x})}\right)$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. We use prime to denote the transpose of vectors, for which we use boldface lower case letters. Matrices are denoted using boldface upper case letters. $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. We use $\mathbf{0}$ and $\mathbf{1}$ for the vectors with all entries equal to zero and one, respectively. $diag(\mathbf{x})$ denotes a diagonal matrix with the main diagonal being the vector $\mathbf{x}$ and all other off-diagonal elements being zero. Unless otherwise specified, $\|\cdot\|$ denotes the ℓ_2 norm , $\|\cdot\|_1$ the ℓ_1 norm and $\|\mathbf{x}\|_p = (\sum_{i=1}^{\dim(\mathbf{x})} |x_i|^p)^{1/p}$ the ℓ_p norm, where $p \geq 1$. $\|\cdot\|_0$ denotes the ℓ_0 counting "norm." $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the minimum and maximum eigenvalue of a (symmetric) matrix. We use $\emptyset$ to denote the empty set, $[n]$ for the set $\{1, \ldots, n\}$, and $|S|$ for the cardinality of the set S. We write $f(n) = O(g(n))$ if there exist positive numbers n_0 and c such that $f(n) \leq cg(n), \forall n \geq n_0$. We write $f(n) = \Omega(g(n))$ if there exist positive numbers n_0 and c such that $f(n) \geq cg(n), \forall n \geq n_0$. We write $f(n) = \Theta(g(n))$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ hold. Finally, E and P denote expectation and probability, respectively.

Chapter 2

Designing Metabolic Division Of Labor In Microbial Communities

Microbes often encounter a tradeoff between metabolic independence, which requires numerous costly functions, and inter-dependence, in which some essential metabolites are provided by neighboring organisms in a network of cross-feeding. This balance of conflicting strategies is likely a key determinant of microbial community structure and dynamics, with important implications for microbiome research and synthetic ecology. A thought experiment to investigate this tradeoff would involve gradually limiting the number of metabolic reactions allowed in a given species. The expectation is that below a certain number of reactions, no individual organism would be able to grow in isolation, and cross-feeding partnerships and division of labor would emerge. We implemented this idealized experiment using in silico genome-scale models. In particular, we used Mixed Integer Linear Program (MILP) to identify tradeoff solutions in communities of Escherichia coli variants. The strategies we found are more complex than previously engineered syntrophies. For example, two E. coli variants survive in coculture by exchanging intermediates that enable each variant to perform half of the tricarboxylic acid (TCA) cycle. More broadly, we systematically mapped the landscape of possible 1-, 2-, and 3-strain solutions at increasingly tight constraints on the number of allowed reactions. This landscape displays a nonlinear boundary, indicating that the loss of an intracellular reaction is not necessarily compensated by a single imported metabolite. Different regions in this landscape are associated with

specific solutions and patterns of exchanged metabolites. Our model also predicts the existence of regions in this landscape where independent bacteria are feasible, but outcompeted by cross-feeding pairs, providing a possible mechanism for the rise of division of labor.

This part of the thesis is based on collaboration with Qi Zhao, Meghan Thommes and Daniel Segrè (Zhao et al., 2016; Thommes et al., 2018). We are grateful to Sara Collins and Joshua Goldford for discussions on division of labor in microbial communities, and to all members of the Segrè Lab for helpful discussions and feedback.

2.1 Background

We explore how metabolic differentiation emerges from an isogenic population by developing constraint-based modeling approach for identifying Division Of Labor in Metabolic Networks (DOLMN). In particular, using DOLMN, we explore the space of feasible single-strain or multi-strain metabolic networks by systematically limiting the number of intracellular and transport reactions in each metabolic model. After introducing the mathematical and Linear Program formulation of DOLMN, we illustrate its capabilities through an analysis of division of labor based on a core Escherichia coli model. We next apply DOLMN to a full E. coli model, and show that metabolically differentiated and interdependent communities are able to exist under harsher reaction constraints than a single, isolated strain, and even outcompete the single strain. Our algorithm and the outcome of our thought experiments broaden our perspective on possible metabolic interdependence between metabolically differentiated species, with applications in understanding diversity in natural microbial communities, and in designing new artificial consortia.

2.2 Methods

2.2.1 Flux Balance Analysis (FBA)

A metabolic network is used to describe the process of thousands of enzymatic reactions used to convert nutrients into metabolites and energy. Organisms have different optimal performances (e.g., maximizing growth rate, ATP generation) under a range of growth conditions (Edwards et al., 2001; Sauer et al., 1998; Papoutsakis, 1984). Flux Balance Analysis (FBA) (Orth et al., 2010b; Kauffman et al., 2003) has emerged as one of the most important methodologies to analyze the metabolic network in steady-state. FBA formulates the problem of predicting the metabolic reaction fluxes as a Linear Program problem. Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ denote the stoichiometric matrix (expressing mass balance) where m is the number of metabolites and n the number of metabolic fluxes, $\mathbf{x} \in \mathbb{R}^n$ the vector of metabolic fluxes (internal and external), and $\mathbf{x}_{lb}$, $\mathbf{x}_{ub}$ lower and upper bounds on the metabolic fluxes implied by the composition of the growth medium. External fluxes represent the transfer rates between metabolites which exist inside and outside of the cell. Internal fluxes represent the reactions rates among metabolites inside of the cell. The cellular objective is expressed as a vector of weight coefficients for each reaction $\mathbf{c}$ (e.g., biomass) and the optimal objective value is a scalar.The FBA problem is formulated as:

$$\begin{array}{rl} \max_{\mathbf{x}} & \mathbf{c}'\mathbf{x} \\ \text{s.t.} & \mathbf{S}\mathbf{x} = \mathbf{0}, \\ & \mathbf{x}_{lb} \leq \mathbf{x} \leq \mathbf{x}_{ub}, \end{array} \tag{2.1}$$

where $\mathbf{0}$ is the vector of all zeroes and primes indicates transpose.

FBA is a mathematical method for simulating metabolism in genome-scale reconstructions of metabolic networks. Compared to other modeling methods, FBA is less intensive in terms of the input data required for constructing the model and calculations performed using FBA are computationally inexpensive. The results of

FBA can be visualized using flux maps which illustrate the steady-state fluxes carried by reactions. FBA, on the other hand, only predicts steady state fluxes so it cannot predict dynamic changes in fluxes over time. FBA is also not particularly useful in predicting the concentrations of these metabolites since it only predicts fluxes. FBA can be applied to an entire cell or entity but has not been widely used for microbial communities.

2.2.2 Community-level Flux Balance Analysis

We introduce a "universal stoichiometric matrix" denoted by $\mathbf{S}$, which expresses mass balance for all possible reactions in a microbial community irrespective of the organism they belong to. Specifically, $\mathbf{S} \in \mathbb{R}^{M \times N}$ where $M = M_e + M_i$ represents the number of distinct metabolites and $N = N_e + N_t + N_i$ represents the number of distinct reactions. The M distinct metabolites consist of two types: M_e external and M_i internal metabolites. The external metabolites exist in the shared extracellular environment of all organisms and the internal metabolites are intracellular. There are 3 different types of reactions: N_e exchange reactions, N_t transport reactions and N_i intracellular reactions. The availability of nutrients (external metabolites) from the environment is encoded in the exchange reactions. With transport reactions, organisms transport these metabolites between their intracellular compartment and the extracellular environment. Intracellular reactions take place among internal metabolites.

Assuming that the community consists of K different species, we use $\mathbf{S}^k$, $k \in [K]$, to denote the (individual) stoichiometric matrix of species k. This is the matrix that expresses mass balance constraints for all reactions utilized by the metabolic network of that species. The columns in $\mathbf{S}^k$ are a subset of the columns in $\mathbf{S}$. The matrices $\mathbf{S}^k$ may have identical columns (reactions) if more than one species use the same reaction. Without loss of generality, the structure of $\mathbf{S}$ is shown in Fig. 2·1. Given the matrix

$\mathbf{S}$ and the number of species K, our goal is to design the individual stoichiometric matrices $\mathbf{S}^k$ that render each species and the community viable, that is, satisfying appropriate optimality criteria.

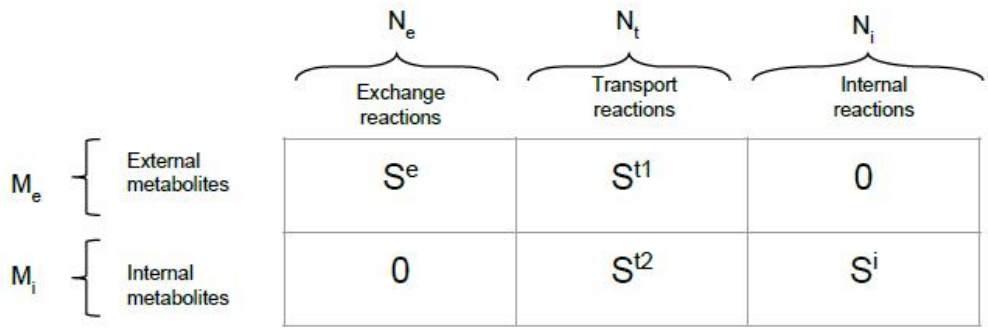

Figure 2·1: The structure of the universal stoichiometric matrix $\mathbf{S}$. Block $\mathbf{S}^e$ represents the set of exchange reactions used to absorb nutrients from the environment. Blocks $[\mathbf{S}^{t1}; \mathbf{S}^{t2}]$represent the set of transport reactions between external and internal metabolites. $\mathbf{S}^i$ represents the set of intracellular reactions among internal metabolites.

In order to formulate the design problem, we first reformulate the universal stoichiometric matrix $\mathbf{S}$ to construct putative stoichiometric matrices for each species in the community (Klitgord and Segrè, 2010a). In particular, we construct a community stoichiometric matrix $\mathbf{S}^c$ whose structure is shown in Fig. 2·2. The block matrices $\mathbf{S}^e$, $\mathbf{S}^{t1}$, and $\mathbf{S}^{t2}$ in $\mathbf{S}^c$ are consistent with those in $\mathbf{S}$. Organisms in the community share the same nutrients and exchange reactions. Because there are K organisms in the community, we replicate the block $[\mathbf{S}^{t2}, \mathbf{S}^i]$ that includes transport reactions and intracellular reactions K times and diagonally arrange them in $\mathbf{S}^c$. This arrangement represents the process according to which organisms absorb the nutrients via the same exchange reactions and then allocate them to individual organisms in the community.

After obtaining the internal metabolites via the transport reactions, intracellular reactions take place inside each organism. This construction leads to a community

stoichiometric matrix $\mathbf{S}^c \in \mathbb{R}^{M_c \times N_c}$, where $M_c = M_e + KM_i$ and $N_c = N_e + K(N_t + N_i)$. Notice that $\mathbf{S}^c$ has one block column for exchange reactions (N_e columns) and K block columns (of dimension $N_t + N_i$), one for each organism, including all transport and intracellular reactions.

To capture design choices, we introduce a binary putative vector $\mathbf{t} = (t_1, \cdots t_{N_c})$, where $t_i \in \{0, 1\}$ is a binary variable, indicating whether the i-th reaction is included or not in the corresponding organism (cf. Fig. 2·2). With $\mathbf{t}$ and $\mathbf{S}^c$ available, we can partition $\mathbf{S}^c$ to K individual matrices, $\mathbf{S}^k$, by removing columns j with $t_j = 0$.

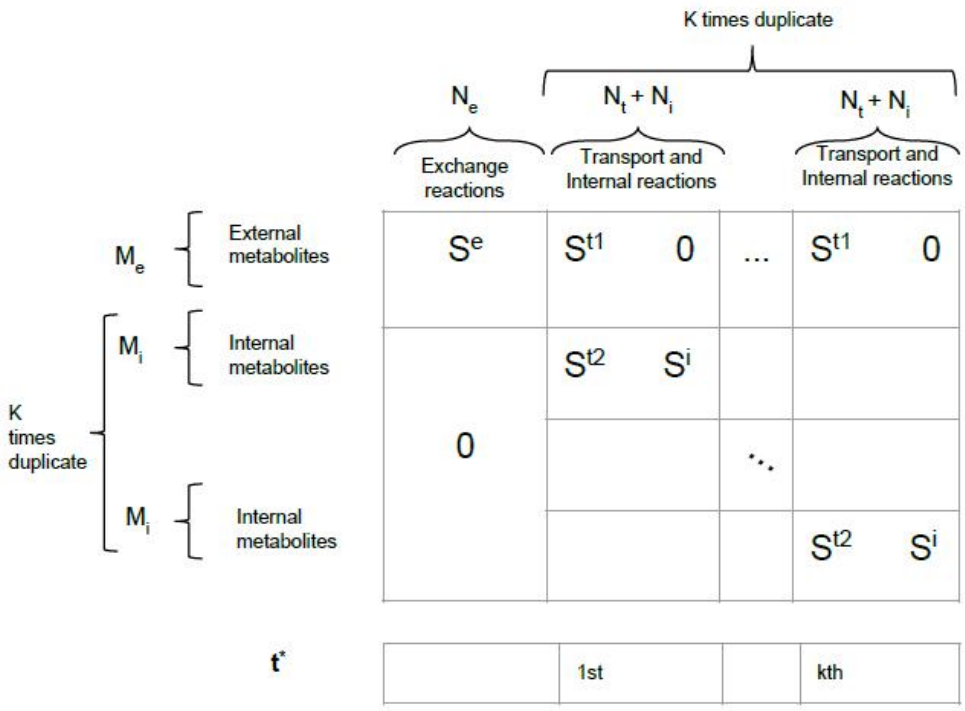

Figure 2·2: The structure of the stoichiometric matrix for the whole community $\mathbf{S}^c \in \mathbb{R}^{M_c \times N_c}$.

In practice, the putative vector $\mathbf{t}$ is unavailable. The problem of identifying $\mathbf{t}$ can

now be formulated as the following *Mixed-Integer Linear Program (MILP)* problem:

$$\begin{aligned} \max_{\mathbf{x},\mathbf{t}} \quad & \mathbf{c}'\mathbf{x} \\ \text{s.t.} \quad & \mathbf{S}^c\mathbf{x} = \mathbf{0}, \\ & diag(\mathbf{x}_{lb})\mathbf{t} \leq \mathbf{x} \leq diag(\mathbf{x}_{ub})\mathbf{t}, \\ & t_i \in \{0,1\}, \\ & \mathbf{t}_{min} \leq \mathbf{R}\mathbf{t} \leq \mathbf{t}_{max}, \end{aligned} \tag{2.2}$$

where $\mathbf{c} \in \mathbb{R}^{N_c}$ corresponds to the community objective (e.g., maximizing the total biomass growth rate, maximizing the total exchange rate of a limited nutrient, etc. (Sauer et al., 1998)) and $\mathbf{x}_{ub}$ and $\mathbf{x}_{lb}$ are upper and lower bounds on fluxes. The constraints $diag(\mathbf{x}_{lb})\mathbf{t} \leq \mathbf{x} \leq diag(\mathbf{x}_{ub})\mathbf{t}$ guarantee that the flux of a reaction that is not included in an organism is set to 0. Additional regularization constraints on $\mathbf{t}$ are necessary to obtain a biologically meaningful solution. Specifically, we impose

$$\mathbf{t}_{min} \leq \mathbf{R}\mathbf{t} \leq \mathbf{t}_{max}, \tag{2.3}$$

where $\mathbf{R} \in \mathbb{R}^{M_r \times N_c}$ is a regularization matrix. The number of rows M_r depends on the number of regularization constraints we wish to introduce. Such constraints can, for example, impose upper and lower bounds on the number of reactions active in each organism, and/or enforce upper and lower bounds on the number of repeated reactions in different organisms (hence, controlling for community diversity and robustness). The regularization constraints enable us to partition the community ($\mathbf{S}^c$) into individual species ($\mathbf{S}^k$) that are biologically meaningful. For instance, cross-feeding partnerships and division of labor would emerge or individual species may collaborate under some tight constraints on the number of reactions since no individual organism would be able to grow in isolation.

2.2.3 A General Formulation

If $\mathbf{S}^c$ and $\mathbf{t}$ are known, we can formulate the FBA problem to predict the flux distribution vector in the entire community as:

$$\begin{array}{rl} \max_{\mathbf{x}} & \mathbf{c}'\mathbf{x} \\ \text{s.t.} & \mathbf{S}^c\mathbf{x} = \mathbf{0}, \\ & diag(\mathbf{x}_{lb})\mathbf{t} \leq \mathbf{x} \leq diag(\mathbf{x}_{ub})\mathbf{t}. \end{array} \tag{2.4}$$

Problem (2.4) is a Linear Program problem and it can be solved efficiently. The optimal solution of (2.4), denoted by $\mathbf{x}^*$, represents the flux vector for the entire community.

We now write another Linear Program problem which, as we will see, reduces to the dual of (2.4) when all the non-active (with a corresponding $t_j = 0$) reactions j and their fluxes x_j are removed from (2.4).

$$\begin{array}{rl} \min_{\mathbf{p},\mathbf{q}_1,\mathbf{q}_2} & \mathbf{q}_2'\mathbf{x}_{ub} - \mathbf{q}_1'\mathbf{x}_{lb} \\ \text{s.t.} & \mathbf{S}^{c\prime}\mathbf{p} - \mathbf{q}_1 + \mathbf{q}_2 - \mathbf{c} - L(\mathbf{1} - \mathbf{t}) \leq \mathbf{0}, \\ & \mathbf{S}^{c\prime}\mathbf{p} - \mathbf{q}_1 + \mathbf{q}_2 - \mathbf{c} + L(\mathbf{1} - \mathbf{t}) \geq \mathbf{0}, \\ & \mathbf{q}_1, \mathbf{q}_2 \geq \mathbf{0}, \\ & \mathbf{q}_1, \mathbf{q}_2 \leq L\mathbf{t}, \end{array} \tag{2.5}$$

where $\mathbf{p}$, $\mathbf{q}_1$ and $\mathbf{q}_2$ are dual variables corresponding to primal constraints $\mathbf{S}^c\mathbf{x} = \mathbf{0}$, $\mathbf{x} \geq diag(\mathbf{x}_{lb})\mathbf{t}$, and $\mathbf{x} \leq diag(\mathbf{x}_{ub})\mathbf{t}$ and L is a sufficiently large constant. The role of L is to affect the dual feasibility constraints. The primal variable x_i in (2.4) corresponds to the ith dual constraint in $-L(\mathbf{1}-\mathbf{t}) \leq \mathbf{S}^{c\prime}\mathbf{p}-\mathbf{q}_1+\mathbf{q}_2-\mathbf{c} \leq L(\mathbf{1}-\mathbf{t})$. For the reactions included in the corresponding organism we have $t_i = 1$ for i in some index set $\mathcal{I}$. For these $i \in \mathcal{I}$, the corresponding dual constraint becomes $[\mathbf{S}^{c\prime}\mathbf{p}]_i - q_{1i} + q_{2i} - c_i = 0$. For those reactions j, however, which are not included in the corresponding organism ($t_j = 0$, $j \notin \mathcal{I}$), the corresponding dual constraint becomes $-L \leq [\mathbf{S}^{c\prime}\mathbf{p}]_j - q_{1j} + q_{2j} - c_j \leq L$, which is trivially satisfied for a large enough L. Similarly, the dual variables $\mathbf{q}_1$ and $\mathbf{q}_2$ correspond to the lower and upper bounds on $\mathbf{x}$ in (2.4). If x_i is non-zero ($t_i = 1$),

q_{1i} and q_{2i} are both non-negative and can take arbitrarily large values bounded by the large constant L. If, however, x_j is set to zero ($t_j = 0$), then q_{1j} and q_{2j} are set to 0 as well to avoid an unbounded objective value in (2.5).

In fact, identifying $\mathbf{t}$ is our primary goal in this chapter. Based on the analysis of the community-level FBA and the corresponding dual problem, the feasible set of $\mathbf{t}$, denoted by $\mathcal{F}$ can be represented by the following constraints:

$$\begin{aligned}
&\mathbf{S}^c\mathbf{x} = \mathbf{0},\\
&\mathbf{x} \geq diag(\mathbf{x}_{lb})\mathbf{t},\\
&\mathbf{x} \leq diag(\mathbf{x}_{ub})\mathbf{t},\\
&\mathbf{S}^{c\prime}\mathbf{p} - \mathbf{q}_1 + \mathbf{q}_2 - \mathbf{c} - L(\mathbf{1} - \mathbf{t}) \leq \mathbf{0},\\
&\mathbf{S}^{c\prime}\mathbf{p} - \mathbf{q}_1 + \mathbf{q}_2 - \mathbf{c} + L(\mathbf{1} - \mathbf{t}) \geq \mathbf{0},\\
&\mathbf{q}_1, \mathbf{q}_2 \geq \mathbf{0},\\
&\mathbf{q}_1, \mathbf{q}_2 \leq L\mathbf{t},\\
&t_i \in \{0, 1\},\\
&\mathbf{c}'\mathbf{x} = \mathbf{q}_2'\mathbf{x}_{ub} - \mathbf{q}_1'\mathbf{x}_{lb}.
\end{aligned} \tag{2.6}$$

These constraints capture primal feasibility, dual feasibility and strong duality conditions, respectively. The problem of identifying $\mathbf{t}$ can now be formulated as the following *Mixed Integer Linear Program* problem:

$$\begin{aligned}
\max_{\mathbf{x},\mathbf{p},\mathbf{q}_1,\mathbf{q}_2,\mathbf{t}} \quad & f(\mathbf{x}, \mathbf{p}, \mathbf{q}_1, \mathbf{q}_2, \mathbf{t})\\
\text{s.t.} \quad & \mathbf{x}, \mathbf{p}, \mathbf{q}_1, \mathbf{q}_2, \mathbf{t} \in \mathcal{F},\\
& \mathbf{t}_{min} \leq \mathbf{R}\mathbf{t} \leq \mathbf{t}_{max},
\end{aligned} \tag{2.7}$$

where $f(\cdot)$ is a global objective function. Relevant objective functions could include maximization of the total exchange rate of some nutrients or maximization of the total biomass growth rate. By solving problem (2.7), we can obtain the putative vector $\mathbf{t}^*$ and the flux distribution $\mathbf{x}^*$ for the entire community. Then, we can partition $\mathbf{S}^c$ into a set of $\mathbf{S}^k$ stoichiometric matrices, one for each species $k \in [K]$, based on $\mathbf{t}^*$. If $f(\mathbf{x}, \mathbf{p}, \mathbf{q}_1, \mathbf{q}_2, \mathbf{t}) = \mathbf{c}'\mathbf{x}$ for problem (2.7), e.g., maximization of the total biomass growth rate, the problem of identifying $\mathbf{t}$ can now be simplified as (2.2).

2.2.4 E. coli Core model and iJR904

The core E. coli model (Orth et al., 2010a) is a small-scale model of the central metabolism of E. coli which contains 134 genes, 95 reactions, and 72 metabolites. This model is used for educational purposes, since the results of most constraint-based calculations are easier to interpret on this smaller scale. It is also useful for testing new constraint-based analysis methods.

In 2003, the iJE660 network was updated to form iJR904 (Reed et al., 2003). This model is significantly expanded, containing 904 genes, 931 compartments, and 625 metabolites. iJR904 contains explicit gene-protein-reaction interactions, Boolean rules that define which genes are required for each reaction. Reactions were checked for proper charge balancing, and gaps in the model were identified and filled when possible (EcoliWiki, 2018).

Table 2.1: Number of reactions for different models

Model name	# of exchange reactions	# of transport reactions	# of intracellular reactions	# of all reactions
E. coli core 1	20	25	50	95
E. coli core 2	20	50	100	170
E. coli core 3	20	75	150	245
iJR904 1	143	205	727	1075
iJR904 2	143	410	1454	2007
iJR904 3	143	615	2181	2939

2.2.5 The First Optimization Problem

Suppose there are K organisms. The upper bound on the number of nonzero transport and intracellular reactions in each organism is denoted by T_{TR} and T_{IN}, respectively. We let x_{biom_k} denote the flux of the biomass reaction for each organism $k \in [K]$.

We also let TR_k and IN_k denote the index sets of the transport and intracellular reactions for each organism, respectively. The problem in (2.2) subject to these constraints takes the form

$$\begin{aligned} \max_{\mathbf{x},\mathbf{t}} \quad & \mathbf{c}'\mathbf{x} \\ \text{s.t.} \quad & \mathbf{S}^c\mathbf{x} = \mathbf{0}, \\ & \mathbf{x}_{lb} \le \mathbf{x} \le \mathbf{x}_{ub}, \\ & diag(\mathbf{x}_{lb})\mathbf{t} \le \mathbf{x} \le diag(\mathbf{x}_{ub})\mathbf{t}, \\ & x_{biom_1} = x_{biom_k} \ge 0.1, \\ & \textstyle\sum_{i\in TR_k} t_i \le T_{TR}, \\ & \textstyle\sum_{i\in IN_k} t_i \le T_{IN},\ t_i \in \{0,1\},\ k \in [K]. \end{aligned} \tag{2.8}$$

We define the optimal flux of the first stage as $\mathbf{x}^*$, and the optimal integer variable of the first stage as $\mathbf{t}^*$. This is a *Mixed Integer Linear Program* problem.

2.2.6 The Second Optimization Problem

In order to reduce redundant fluxes in transport and intracellular reactions, the second optimization problem is introduced when the integer variable is fixed as the optimal integer variable of the first stage as $\mathbf{t}^*$. $\mathbf{x}_{N_e+1:end}$ is the flux vector that corresponds to transport and intracellular reactions but excludes the extracellular reactions. Specifically, we have

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}_{N_e+1:end}\|_1 \\ \text{s.t.} \quad & \mathbf{S}^c\mathbf{x} = \mathbf{0}, \\ & \mathbf{x}_{lb} \le \mathbf{x} \le \mathbf{x}_{ub}, \\ & diag(\mathbf{x}_{lb})\mathbf{t}^* \le \mathbf{x} \le diag(\mathbf{x}_{ub})\mathbf{t}^*, \\ & x_{biom_k} = x^*_{biom_k},\ k \in [K]. \end{aligned} \tag{2.9}$$

x_{biom_k} denotes the flux of the biomass reaction for each organism $k \in [K]$ and $x^*_{biom_k}$ denotes the optimal flux of the biomass reaction for each organism $k \in [K]$ attained in the first optimization problem. This problem minimizes the ℓ_1 norm of the flux vector to induce sparsity and can be rewritten as a *Linear Program (LP)* problem.

2.2.7 Essential Intracellular Reactions

For every such reaction R_i in one organism or multiple organisms, set $t_j = 1, \forall j \neq i$, and $t_i = 0$. Further, set the lower bound on biomass to 0.1, and check the feasibility of the Linear Program that is simplified from the MILP (2.8). If the Linear Program is not feasible, the reaction is essential for bacteria biomass growth in one organism. Setting $t_i = 1$ for the essential reaction improves the speed of solving the MILP (2.8) since the number of integer variables is reduced.

2.2.8 Heuristic Solutions to speed up Branch and Bound

It is intractable to solve a large-scale MILP with hundreds or thousands of integer variables and heuristic methods are often used since MILP is an NP-complete problem.

From our experiments, the MILP with a few hundred or fewer integer variables can be solved quickly in minutes or hours for a community model of one or two E. coli core models without resorting into any heuristic method. For a community model of a single iJR904 model or more E. coli core models, the MILP can be solved quickly if the constraint upper bound is not very tight. However, it is necessary to solve the MILP under tight constraints on the number of reactions since cross-feeding partnerships and division of labor would then naturally emerge. In that case, solving the MILP becomes very slow or impossible, necessitating the use of heuristic methods. We use heuristic methods first to obtain the near-optimal feasible solution and offer it to the solver, thus, reducing the time needed by the solver to reach an optimal solution.

Similarity-Based Heuristic solutions

The upper bound on the number of nonzero transport and intracellular reactions in each organism is T_{TR} and T_{IN}, respectively (see the MILP (2.8)). Using the solution corresponding to an E. coli core model, we increase or decrease T_{TR} and T_{IN} by a small number, e.g., one or two, and find that the active reactions (with a corresponding $t_i = 1$) are similar except relatively few reactions. The following display shows heuristic methods based on constraints of intracellular reactions and it also applies to transport reactions.

Algorithm 1 Similarity-Based Heuristic

Given a feasible solution under the constraint $(T_{IN} + m),\ m \in [M]$,
repeat
  for each $m \in [M]$ **do**
    if no a feasible solution under the constraint T_{IN} is found **then**
      set $OPTION = \{$the active reactions found under the constraint $(T_{IN}+m)\}$
    else
      set $OPTION = \{$the union of active reactions found under the constraint $(T_{IN} + m)$ and $T_{IN}\}$
      use $(\mathbf{x}^*, \mathbf{t}^*)$ as the start solution
    end if
    set $OPTION = OPTION \cup R$, where $R = \{$some reactions randomly chosen from the inactive reactions$\}$
    set $t_i = 0,\ \forall i \notin OPTION$
    solve the MILP under the constraint T_{IN}
    if a feasible solution under the constraint T_{IN} is found **then**
      save the feasible solution $(\mathbf{x}^*, \mathbf{t}^*)$
    end if
  end for
until the objective function value does not change over several iterations or iterations > max-iterations

An example is listed to show how the heuristic method works. Suppose we obtain an optimal solution for $T_{TR} = 25$ and $T_{IN} = 50$ and seek a near-optimal feasible solution for $T_{TR} = 25$ and $T_{IN} = 49$.

- Define an OPTION set as the set containing all reactions for which a decision

(active or inactive) is to be made by the MILP and initialize it as the empty set. Put the active reactions found with $T_{IN} = 50$ into the OPTION set and also randomly choose some reactions from the inactive reactions to put into the OPTION set.

- Force all reactions out of the OPTION set to be inactive (i.e., $t_i = 0$), and solve the MILP under the constraint $T_{IN} = 49$. This typically can be done fast as there are relatively few integer variables.

- If a feasible solution is not found, repeat the previous step many times until a feasible solution is achieved. If a feasible solution is not eventually found, we claim no feasible solution exists under the more restrictive constraint and exit.

- If the feasible solution is achieved, put the active reactions corresponding to $T_{IN} = 50$ and $T_{IN} = 49$ into the OPTION set and randomly choose some reactions from remaining inactive reactions to put into the OPTION set.

- Force all reactions out of OPTION set to be inactive, and solve the MILP with a few hundred or fewer integer variables quickly.

- Repeat the previous step until the objective function value does not change over several iterations.

Once the feasible solution is achieved, it can be proved that the objective function value must be monotone nondecreasing with iterations, i.e., the new feasible solution achieved must have the better or the same objective function value.

Game Theory-Based Heuristic solutions

Suppose we can solve a community model of just one organism quickly and can also obtain a feasible solution for a community model of K organisms. We want to seek

a near-optimal feasible solution (or local optimum) for a community model of K organisms.

Algorithm 2 Game Theory-Based Heuristic

Given a feasible solution $(\mathbf{x}^*, \mathbf{t}^*)$ for a community model of K organisms,
repeat
  for each $k \in [K]$ **do**
    set $FIXEDSET = \{$all reactions except those corresponding to kth organism$\}$
    set $t_i = t_i^*,\ \forall i \in FIXEDSET$
    solve the MILP under the constraint with only integer variables corresponding to the community of the kth organism and use $(\mathbf{x}^*, \mathbf{t}^*)$ as the start solution
    save the feasible solution
  end for
until the objective function value does not change over several iterations or iterations > max-iterations

- Split K organisms into two communities of $K-1$ organisms and 1 organism. We may fix integer variables corresponding to reactions of the $K-1$ organisms and solve the MILP with only integer variables corresponding to the community of the single organism.

- Repeat the previous step until the objective function value does not change over several iterations.

The intuition is that one organism wants to optimize its own utility given others' growth in the game. From such game-theory based solutions, organisms eventually reach a Nash Equilibrium, i.e., one organism will have no benefit from changing reaction strategies assuming other organisms remain unchanged in their reaction strategies.

There may be multiple Game Theory-Based heuristic solutions. If the initial solutions are different, different heuristic solutions may appear in the end. Similarly, we can also split the organism into different parts, e.g., exchange, transport, intracellular, so that the MIP scale is small and optimize one part while fixing others.

2.2.9 Computer Specifications and Software

MILPs were solved by GUROBI 7.0 or 7.5 (http://www.gurobi.com) and MATLAB 2017a on Shared Computing Cluster (SCC) Compute Nodes with 36 (2 eighteen-core 2.4 GHz Intel Xeon E7-8867v4) or 28 cores (2 fourteen-core 2.6 GHz Intel Gold 6132). SCC is the Boston University high performance computing resource located in Holyoke, MA at the Massachusetts Green High Performance Computing Center (MGHPCC), a collaboration between 5 major universities and the Commonwealth of Massachusetts. Some examples of highly optimized integer optimization solvers include CPLEX, GLPK, MOSEK and GUROBI based on branch-and-bound methods.

2.3 Results and Discussion

2.3.1 Results for *E.coli* Core Model

As a first test and illustrative example of DOLMN, we investigated how E. coli core carbon metabolism (Orth et al., 2010a) on minimal glucose medium would be partitioned between two strains (i.e., two trimmed versions of the E. coli core network) for a given limit on the number of allowed reactions. An interesting outcome of this analysis, obtained for subnetworks under some constraints of transport reactions and intracellular reactions, was the discovery of a metabolic strategy in which each strain performs half of the tricarboxylic acid (TCA) cycle. None of the strains, in this case, were able to perform all needed metabolic functions without the inflow of specific metabolites produced by the partner. In particular, exchange of 2-oxoglutarate and pyruvate was necessary for survival of this 2-species consortium. This example illustrates how, even for a relatively small network, DOLMN can provide predicted division of labor strategies that could not be easily designed manually. Furthermore, DOLMN could be applied to other core metabolic models (Edirisinghe et al., 2016), which have been generated for a large number of organisms. Figure 2·3 (a) shows

the metabolic network of the core E. coli model, containing 95 reactions and 72 metabolites. This model contains 20 extracellular metabolites and 52 intracellular metabolites, as well as 20 exchange reactions, 25 transport reactions, 49 intracellular metabolic reactions, and 1 biomass reaction. The biomass reaction is not shown. Pathways and extracellular metabolites are labeled. Key intracellular reactions are labeled using the legend. Figure 2·3 (b) shows the solution of 2-strain communities when 11 transport reactions ($T_{TR} = 11$) and 26 intracellular reactions ($T_{IN} = 26$) are allowed. Reactions that the algorithm identifies as excluded or that have zero flux are indicated as a hollow arrow. The tricarboxylic acid (TCA) cycle is split between the two strains. Both strains consume oxygen, glucose, phosphate, and ammonium, and secrete carbon dioxide and water. The strains exchange the TCA intermediate 2-oxoglutarate and the glycolytic intermediate pyruvate (bolded).

2.3.2 Results for *E.coli* Full Model

We next applied DOLMN to a much larger global network, namely genome-scale E. coli metabolism (Reed et al., 2003). In this case, individual strains found by the algorithm would represent E. coli variants with a reduced set of functionalities. We systematically mapped the landscapes of possible 1-, 2-, and 3-strain simulations to display how the growth rates vary as a function of T_{IN} and T_{TR}. One first observation, consistent with expectations, is that as T_{IN} decreases (for unconstrained number of transport reactions) individual strains reach a limit beyond which they cannot sustain growth, whereas consortia of two and three strains are still viable. For the example analyzed in Figure 2·4, a 1-strain subnetwork needs at least 254 intracellular reactions to grow, whereas 2-strain subnetworks only require 215 intracellular reactions each, and 3-strain subnetworks require 203 intracellular reactions each.

We use heuristic methods to speed up the solution process when the constraint upper bound is very tight, e.g., when $T_{TR} = 22$ and $T_{IN} \leq 268$ for a single iJR904

model. Furthermore, we can verify using the solver that the heuristic solutions for a single iJR904 model are optimal when $T_{TR} = 22$ and $T_{IN} \leq 268$. To that end, we first obtain heuristic solutions. Second, we provide these heuristic solutions as initial solutions to the solver and solve MILP exactly with the solver. We observe that the solver can verify the optimality of the heuristic solutions relatively faster with our heuristic solutions than without any initial solutions. The reason is that the branch and bound search method used by the solver can use the heuristic solution value as a bound and thus eliminate many nodes of the corresponding branch and bound tree.

The observed landscapes display a fundamental nonlinear trade-off between minimizing T_{IN} (intracellular complexity) and minimizing T_{TR} (metabolic exchange). This nonlinearity implies that removing the same number of transport reactions at different points along the frontier of the feasible region can be compensated by adding different numbers of intracellular reactions. For example, decreasing T_{TR} by 2 at large T_{TR} can be compensated by adding a single intracellular reaction (increasing T_{IN} by 1), while removing the same number of transport reactions at small T_{TR} will require a much larger compensation with intracellular reactions.

It is important to note that decreasing T_{TR} negatively influences growth because it restricts not only each strain's ability to take up metabolites, but also its ability to secrete metabolites. If an organism cannot secrete metabolites, it accumulates waste (which results in an infeasible FBA solution). Irrespective of the number of strains in the community, it looks like the E. coli strain subnetworks require at least 9 transport reactions in order to support growth.

Further analysis of the landscapes for 1-, 2-, and 3-strain communities also reveals the existence of regions in which division of labor potentially provides a competitive advantage. Given that multiple strains co-existing in a consortium have to share available resources, they will tend to grow slower than individually growing strains

(Figure 2·4 a,b). One notable exception is a thin strip at the boundary in which an individual strain can grow. At this frontier for a single strain, we observe that 2-strain communities can grow more rapidly than 1-strain communities (Figure 2·4 c). A biologically important implication of this result is the fact that the 2-strain communities would in principle have the chance to collectively outcompete the 1-strain ones. Similarly, 3-strain communities grow faster than 2-strain communities along the boundary in which 2-strain communities can grow. These results suggest that the number of strains that achieve the highest growth rates under a given set of circumstances may naturally increase as environmental constraints tighten. This situation could rise if the burden of protein cost in the cell were to increase, or if selection processes were to gradually favor streamlined strains, e.g., as previously observed experimentally by (Ferea et al., 1999; Friesen et al., 2004; Elena and Lenski, 2003; van Gestel et al., 2015; Le Gac et al., 2008; Rosenthal et al., 2018; Rozen and Lenski, 2000; Rozen et al., 2005; Spencer et al., 2007; Spencer et al., 2008).

2.4 Conclusions

Motivated by the need to interpret how organisms interact in a microbial community, we have developed an optimization framework to identify the individual metabolic network of each organism in the community and to predict the resulting flux distributions. We formulated the problem of allocating reactions to organisms as an MILP problem. We tested the method on both the E. coli core model and iJR904. In both cases, the method helped us identify the individual metabolic network topologies and elucidate the interaction between species in the microbial community.

The proposed method provides a meaningful way to analyze and simulate the combinatorial complexity of metabolite-mediated interactions between multiple organisms in a microbial community. It also offers a new platform for the rational

design of organisms and communities towards future synthetic ecology applications.

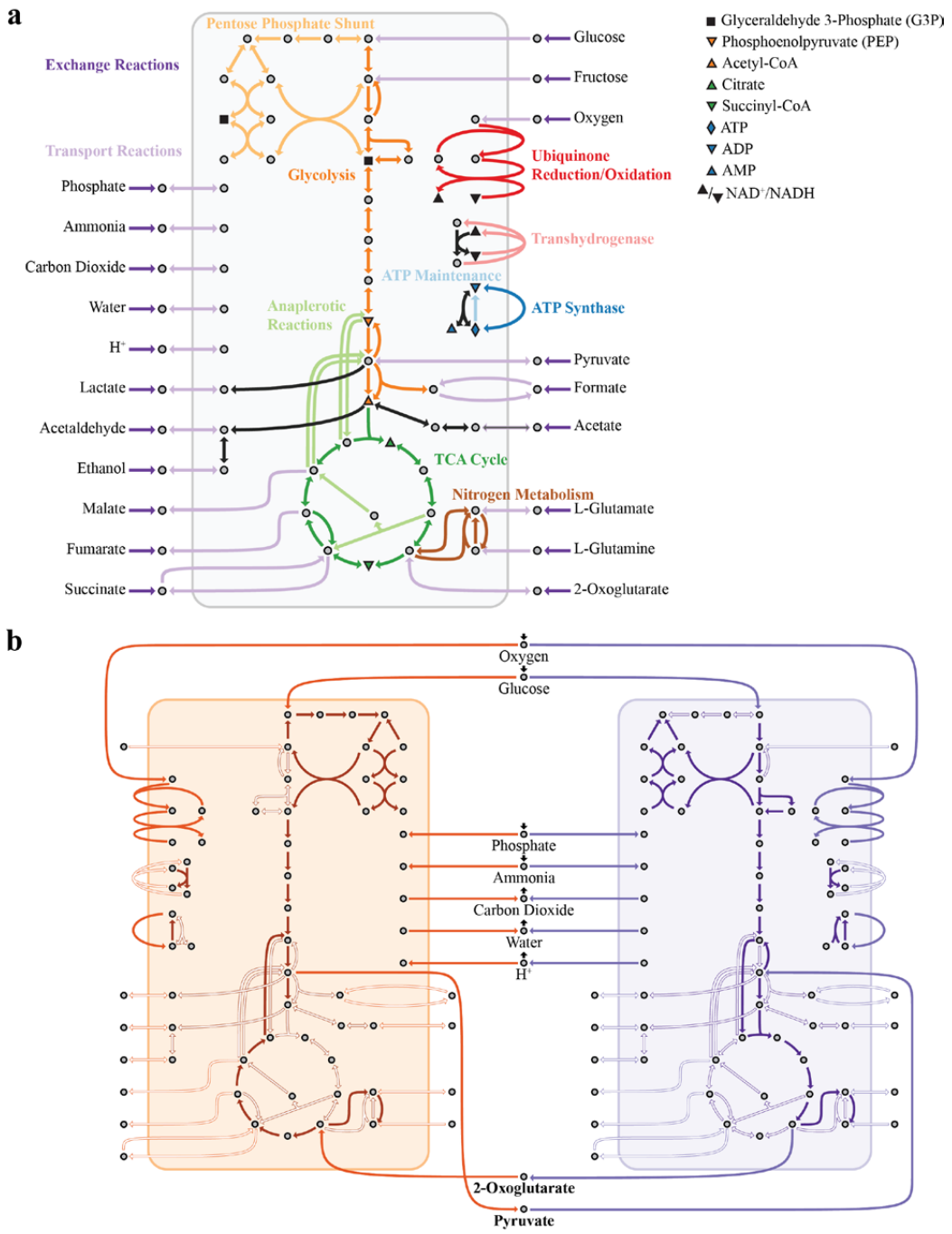

Figure 2·3: A DOLMN flux solution of E. coli core carbon metabolism.

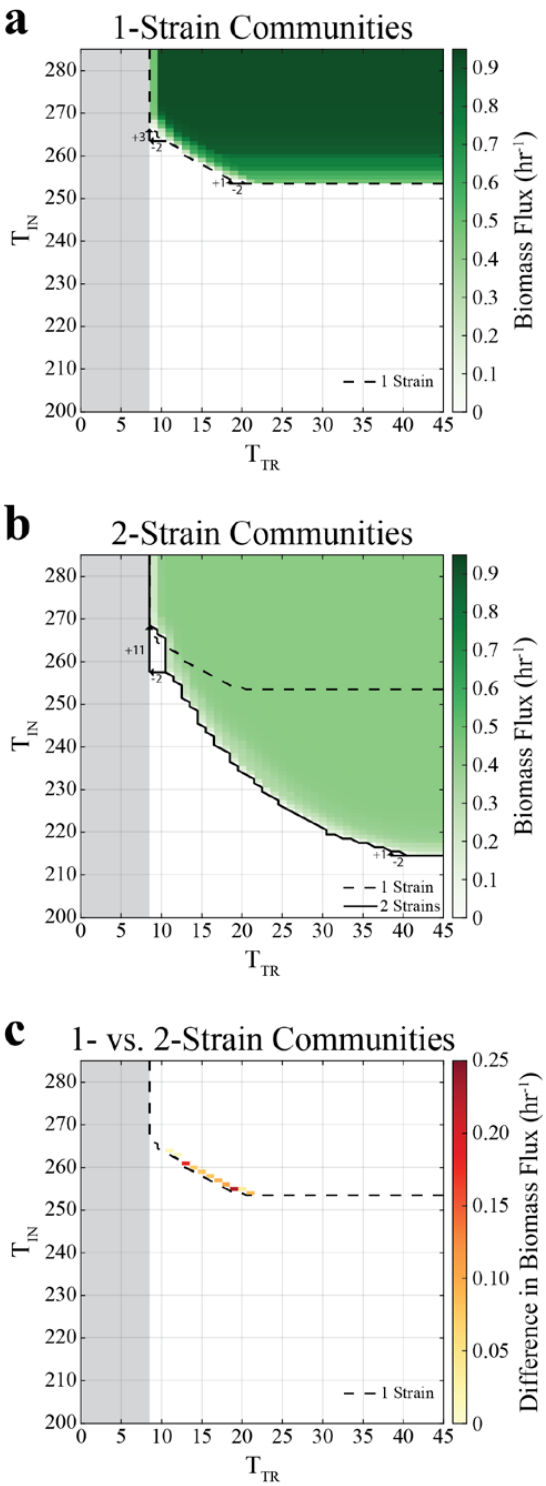

Figure 2·4: (T_{TR}, T_{IN}) growth landscapes of 1- and 2-strain communities.

Chapter 3

Predictive Analytics for 30-day Hospital Readmissions following General Surgery

Hospital readmissions represent a metric of care quality. With increasing data availability, an opportunity arises to learn and develop models that can predict readmissions and guide interventions at the physician, practice and institution level to help minimize them. The objective of this study is to develop supervised learning algorithms that can reliably predict readmissions in patients undergoing surgery within 30-days from discharge.

3.1 Background

In 2005, the American College of Surgeons (ACS) established the National Surgical Quality Improvement Program (NSQIP), which collects detailed demographic, laboratory, clinical, procedure and postoperative occurrence data in several surgical subspecialties. The aim of this effort has been to provide participating institutions with reliable, risk-adjusted data and benchmarking on surgery-related outcomes, including as of 2010 readmissions. With this project, we aim to use our tertiary safety-net academic medical center SQIP data, in addition to socioeconomic variables collected ad-hoc to the routinely collected SQIP features, to determine risk factors for readmission; and develop comprehensive models and analytics using machine learning to identify patients at risk for readmissions after general surgery procedures.

We employ supervised learning methods from the field of machine learning that,

unlike common statistical methods, do not require restrictive statistical assumptions, and can learn efficiently and effectively from large datasets. These methods improve over time, the more they are being utilized and the larger the amount of data they use for training (Brisimi et al., 2018a; Xu et al., 2016).

3.2 Methods

Various supervised classification methods, described in more detail below, were employed. Many of these are well validated and standardized, while Sparse Linear SVM (SLSVM) and Joint Clustering and Classification (JCC) are developed for the purposes of this study.

3.2.1 Standard Classification Methods

The standard methods are random forests (RF), Support Vector Machines (SVMs), and logistic regression (Friedman et al., 2001).

- A random forest (Breiman, 2001) is a large collection of decision trees and it classifies by averaging the decisions of each tree.

- SVMs, which are also used in the two methods we developed, are very efficient two-class classifiers (Cortes and Vapnik, 1995). Linear SVM (Lin. SVM) finds a separating hyperplane in the variable space so that the data points from the two different classes reside on different sides of that hyperplane. Kernel functions are used to elevate the variables into a higher dimensional space where data can be linearly separable. A Radial Basis Function (RBF) SVM (Scholkopf et al., 1997) is used as the kernel function in some of our experiments.

- Another method we implement is logistic regression, which is a widely used as a base for comparison in medical machine learning studies (Friedman et al.,

2001). For our purposes, we implemented logistic regression with an additional regularization term: either an ℓ_1-norm term to induce sparsity, similarly to the SLSVM method, or an ℓ_2-norm term used in ridge regression (Friedman et al., 2001).

- A baseline method we consider is based on a popular readmission predictive index called LACE (van Walraven et al., 2010). In several studies, LACE achieves a prediction accuracy on the order of 65-70% (Area Under the ROC Curve or C-statistic) (van Walraven et al., 2010; Low et al., 2015; Wang et al., 2014; Tan et al., 2013). LACE was derived using logistic regression to identify predictive variables and then translated into an index. The index uses four variables: Length of stay (L), whether the admission was acute/emergent (A), the Charlson comorbidity index (C), and the number of visits to the Emergency Room (E) during the 6 months that precede the admission. We refer the reader to (van Walraven et al., 2010) for details on how the LACE index is computed. In our case, we do not have access to the last variable, hence we compute a LAC index using only the first three variables in exactly the same way as in (van Walraven et al., 2010).

All methods are implemented in Matlab (MathWorks, Natick, MA) except the RF, for which we used the statistical package R (R, Auckland, New Zealand). For RF, the number of trees grown was 500. No normalization method and cross-validation are required to tune parameters of the method and the same was true for the LAC index.

For the remaining methods, we used 3-fold cross-validation (with only training data) for parameter tuning. Before we trained the various classifiers, we scaled the vectors of variables as follows: for each variable, we subtracted the mode of the empirical distribution and normalized by the standard deviation. We used the mode

since the dataset was sparse, containing several missing entries across patients for many variables, and the mean was not necessarily representative of the true mean for that variable. In addition, variables containing information for only a small number of patients (no more than 10) are completely removed.

A random subset of 80% of the entire patient set was used to form the training and validation set, while the remainder constituted the test set. The training process was repeated 10 times for each of the described predictive methods. Each method had its accuracy and predictive ability assessed with the plotting of a Receiver Operating Characteristic (ROC) curve, and calculation of the Area Under the ROC Curve (AUC). An ideal prediction model has an AUC close to 1, whereas a random prediction would yield an AUC of 0.5. Anything with an AUC greater than 0.75 would be considered a good predictive model. The mean (Avg.) and standard deviation (Std.) of AUC with each method is reported.

3.2.2 SLSVM: Sparse Linear SVM

Following (Brisimi et al., 2018a; Xu et al., 2016) and our interest in interpretable classifiers we formulate a sparse version of linear SVM (SLSVM) as follows. We are given training data $\mathbf{x}_i \in \mathbb{R}^D$ and labels $y_i \in \{-1, 1\}$, $i \in [n]$, where $\mathbf{x}_i$ is the vector of features for the ith patient and $y_i = 1$ (resp., $y_i = -1$) indicates that the patient will (resp., not) be readmitted. We seek to find the classifier coefficients $(\boldsymbol{\beta}, \beta_0)$, $\boldsymbol{\beta} \in \mathbb{R}^D, \beta_0 \in \mathbb{R}$ by solving:

$$\begin{aligned} \min_{\boldsymbol{\beta},\beta_0,\xi_i} \quad & \tfrac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \quad \forall i, \\ & y_i(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad \forall i, \\ & \|\boldsymbol{\beta}\|_1 \leq K. \end{aligned} \tag{3.1}$$

where C is a tunable parameter and ξ_i is a misclassification penalty. The $\|\boldsymbol{\beta}\|_1$ constraint imposes sparsity in the feature vector $\boldsymbol{\beta}$, thus allowing only a sparse subset of features to be selected for the classification decision. ρ is a tunable parameter controlling the sparsity constraint, and when $K = \infty$, it degenerates to a standard SVM model. For SLSVM involving ℓ_1 norm or absolute values, we need to do reformulations (Appendix A) in order to use optimization solvers, e.g., GUROBI.

3.2.3 JCC: Joint Clustering and Classification

(Brisimi et al., 2018a; Xu et al., 2016) propose a Joint Clustering and Classification (JCC) method based on the Sparse Linear Support Vector Machine (SLSVM) framework. The SLSVM method we discussed in the previous part can in fact be seen as a special case of JCC where only one cluster is being used. The classification problem under consideration satisfies the following assumptions:

- The negative class samples are assumed to be i.i.d. and drawn from a single cluster with distribution P_0.
- The positive class samples belong to L clusters, with distributions $P_1^1, \cdots, P_1^L$.
- Different positive clusters have different features that separate them from the negative samples (cf. Fig. 3·1).

Let $\mathbf{x}_i^+$ and $\mathbf{x}_j^-$ be the D dimensional positive and negative samples, y_i^+, y_j^- be the corresponding labels, where $i \in [N^+]$ and $j \in [N^-]$ and $y_i^+ = 1,\ \forall i$ and $y_j^- = -1,\ \forall j$. Let T be a parameter controlling the sparsity of the classifiers. Assuming L hidden clusters in the positive class, we try to discover: (a) the L hidden clusters (denoted by a mapping function $l(i) = l,\ l \in [L]$) and (b) L classifiers as the solution of the following JCC problem:

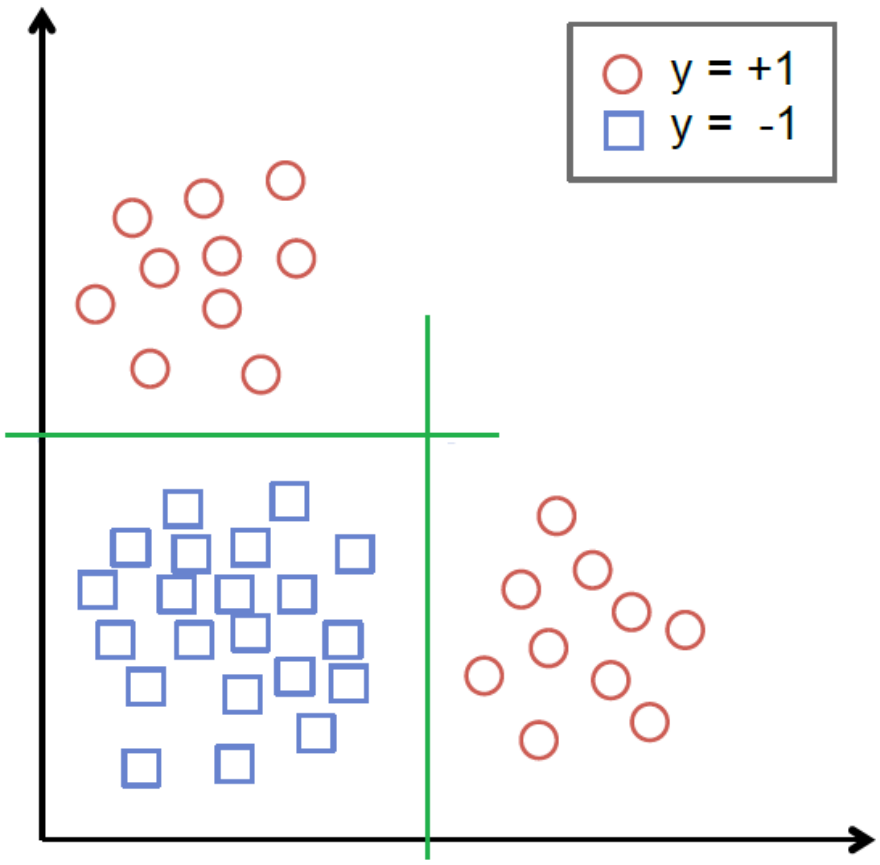

Figure 3·1: The positive class contains two clusters and each cluster is linearly separable from the negative class.

$$
\begin{aligned}
\min_{\boldsymbol{\beta}^l,\beta_0^l,l(i)} \quad & \sum_{l=1}^{L}\left(\tfrac{1}{2}||\boldsymbol{\beta}^l||^2+\lambda^+\sum_{i:l(i)=l}\xi_i^{l(i)}+\lambda^-\sum_{j=1}^{N^-}\zeta_j^l\right)\\
\text{s.t.} \quad & \sum_{d=1}^{D}|\beta_d^l|\le T^l, \qquad \forall l\in[L],\\
& \xi_i^{l(i)}\ge 1-y_i^+\beta_0^{l(i)}-\sum_{d=1}^{D}y_i^+\beta_d^{l(i)}x_{i,d}^+, \quad \forall i\in[N^+],\\
& \zeta_j^l\ge 1-y_j^-\beta_0^l-\sum_{d=1}^{D}y_j^-\beta_d^l x_{j,d}^-, \quad \forall j\in[N^-],\forall l\in[L],\\
& \xi_i^{l(i)},\ \zeta_j^l\ge 0, \qquad \forall i\in[N^+],\forall j\in[N^-],\forall l\in[L],
\end{aligned}
\tag{3.2}
$$

where $y_i^+=1,\ \forall i$ and $y_j^-=-1,\ \forall j$.

In formulation (3.2), the empirical costs of the negative samples are counted L

times, since, because they are drawn from a single distribution, they are not clustered but simply copied into each cluster. Parameters λ^- and λ^+ control the relative weight of costs from negative samples compared to that of the positive samples. The constraint $\sum_{d=1}^{D} |\beta_d^l| \leq T$ is an ℓ_1-relaxation of the sparsity requirement to the local classifiers, which is essential to align the formulation with the problem assumptions and to estimate more robust local classifiers.

Two different approaches has been proposed to solve formulation (3.2) (Brisimi et al., 2018a; Xu et al., 2016). The first approach is to transform the joint problem into Mixed Integer Program problems, which are NP-hard. Thus, this approach suffers from scaling limitations and applies to small-scale problems. The second approach is the alternating optimization approach, i.e., alternately train a classification model and then re-cluster the positive samples, which applies to large-scale problems and also leads to theoretical performance guarantees.

3.3 Data Description and Pre-processing

BMC NSQIP data included

- baseline demographics (gender, race, height, weight);
- pre-existing comorbidity information (such as preoperative functional status, smoking, diabetes, hypertension, congestive heart failure, other pre-existing diagnoses, ASA classification);
- preoperative variables (such as preoperative laboratories);
- index admission-related diagnosis and procedure information (such as emergency room admission, wound classification, anesthesia/operation start/end times, procedure CPT codes, level of resident involved); postoperative events and complications (such as surgical site infections, sepsis, pneumonia, urinary

tract infections, bacteremia, thromboembolic events, unplanned intubation, unplanned return to the operating room, wound dehiscence, hospital length of stay) (Rowell et al., 2007).

- Additional socioeconomic variables (including zip code of main residence, median income for zip code, employment status, profession per the ISCO-08 classification (International Labour Office, 2008), and insurance status) were also collected ad-hoc, and for the purpose of our analysis.

Data were pre-processed as follows:

- Patients who died within 30 days of discharge or had a postoperative length of stay greater than 30 days were excluded.
- Categorical variables (such as race, discharge destination, insurance type) were numerically encoded and units homogenized;
- missing values were replaced by the mode or median;
- all variables were normalized by subtracting the mode and dividing by the standard deviation.

Some important lab variables (highest potassium, lowest potassium, lowest hematocrit, highest sodium, and lowest sodium) miss a substantial number of entries. The reason may be the clinicians discontinue to report after one year so the missing values may not be randomly missing. For these variables, we replace the missing values using recursive regression models. Specifically, we perform a linear regression where the output is the variable of interest, and all other variables (excluding readmission information) are used as predictors. We then use the predicted value from this regression to replace missing values. The benefit is that we can produce sparse models with few important variables rather than deleting them. If we treat them as randomly missing

and replace them with the median, we can observe that the AUC would increase if we do not include them as variables even though they are critical predictive variables. Furthermore, they are highly correlated with readmissions when they are not missing.

3.4 Results

3.4.1 Sample Characteristics

Our sample included a total of 5,741 patients who underwent general surgery procedures at BMC between 02/2010 and 12/2013. These patients do not include the ones who either died within 30 days of discharge (51 patients) or had a postoperative hospital length-of-stay exceeding 30 days (29 patients). Table 3.1 summarizes the baseline demographic and clinical characteristics of these 5,741 patients, 374 of which were readmitted within 30 days of discharge, resulting in a readmission rate of 6.51%. Readmitted patients tend to be older and with larger percentages of white and/or male but smaller percentages of Hispanic or other races than the non-readmitted. Insurance status is different in the readmitted population, having a smaller percentage of uninsured and a higher percentage of patients covered by Medicare. For each variable in Table 3.1, we computed a p-value using Welch's t-test; the p-value is the probability that the difference in the corresponding variable between readmitted and non-readmitted patients is not statistically significant and due to chance.

3.4.2 Model Performance

Table 3.2 summarizes the performance of the various predictive models we described in session (3.2) in terms of AUC. We train each model by using 80% of the patients for training model parameters and evaluate the model's performance on the remaining 20%. We repeat this process 10 times and compute the mean (Avg.) and standard deviation (Std.) of AUC. In Table 3.2, Lin. and RBF SVM are Support

Table 3.1: Demographic and insurance profile of readmitted and non-readmitted patients.

Variable names	All patients	Readmitted	Non-Readmitted	p-value
Female	0.59	0.51	0.59	1.04E-03
Age (years), mean (std)	49.28 (16.17)	53.63 (16.69)	48.97 (16.09)	2.70E-07
White	0.38	0.46	0.38	3.89E-03
Black or African American	0.28	0.34	0.28	2.16E-02
Hispanic	0.26	0.17	0.27	5.90E-06
Asian, American Indian, Alaska Native or Others	0.07	0.03	0.07	2.92E-05
Unknown	0.01	0.01	0.01	3.61E-01
Uninsured or Freecare	0.12	0.05	0.13	1.00E-10
Private	0.29	0.31	0.29	4.40E-01
Medicare	0.21	0.32	0.20	1.88E-06
Medicaid or MassHealth	0.37	0.31	0.38	9.10E-03

Vector Machine classifiers using a linear and an RBF kernel, respectively. JCC is our Joint Clustering and Classification method and L denotes the number of clusters of readmitted patients formed. If one cluster is formed, the method is equivalent to a Sparse Lin. SVM (SLSVM). ℓ_2 and ℓ_1 logistic regression are logistic regression models obtained with an ℓ_2-norm and ℓ_1-norm regularization, respectively. LAC index is a logistic regression model using only Length of Stay, whether an admission is Acute or not, and the Charlson comorbidity index.

For each of the models in Table 3.2, the AUC reported has an associated p-value less than 0.001, where the null hypothesis is obtained by randomly perturbing the labels as in Test 1 of (Ojala and Garriga, 2010). In this case, p-value is the probability that the observed AUC in the test set was obtained by chance, only because the model identified a pattern that happened to be random.

Table 3.2: Performance of the various prediction models.*

Settings	Avg. AUC	Std. AUC
RF	83.63%	1.07%
Lin. SVM	79.25%	2.14%
RBF SVM	79.63%	1.78%
SLSVM	83.85%	1.44%
JCC (L=2)	83.58%	1.55%
JCC (L=3)	83.63%	1.51%
ℓ_2 logistic regression	79.45%	1.94%
ℓ_1 logistic regression	78.44%	2.18%
LAC index	66.76%	2.50%

*p-value less than 0.001.

SLSVM (which is JCC with L=1 cluster) performs the best and, due to sparsity, provides a simple prediction model using a small number of variables (discussed in Section 3.4.3). As such, the model is easy to implement in a clinical setting (does not require tracking many variables) and yields interpretable predictions since it identifies the variables that led to a readmission prediction. JCC with more clusters (L=2 and L=3) is close behind in terms of accuracy and has the potentially useful feature of assigning patients predicted to be readmitted into clusters with distinct characteristics, thus, identifying cohorts of patients who are readmitted for similar reasons. RF perform slightly worse than JCC (with L=1, or SLSVM) and produce a very complex classifier combining predictions from many decision trees (500 or more). As a result, it becomes hard to interpret a prediction. We also note that the LAC index produces significantly less accurate predictions. This implies that using the entirety of the SQIP variables, and not a few that may appear the most intuitive, provides significant benefits.

3.4.3 Predictive Variables

Next, we examine variables used by our models to make a readmission prediction. For methods that use SVM-type classification (including SLSVM and JCC), it is particularly easy to identify important discriminative variables. In SLSVM, let $(\boldsymbol{\beta}, \beta_0)$ be the vector orthogonal to the hyperplane separating the two classes. Non-zero elements of the vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_D)$ identify discriminative variables and, given the sparsity constraint, these variables are not that many. In particular, the value of β_i can be interpreted as the weight of variable i in the prediction. Positive values contribute to a positive readmission prediction whereas negative values reduce the chance of a readmission prediction.

In Table 3.3, we report results from 10 independent runs of SLSVM. In each SLSVM run, we rank variables in terms of the absolute value of the corresponding element in $\boldsymbol{\beta}$ (from largest to smallest). Notice that the signs of the coefficients are consistent with medical intuition. Recall that variables have been normalized by subtracting the mode and dividing by the standard deviation. For example, more complications and infections (than the mode of the distribution) result into a higher chance of a readmission. On the other hand, lowest hematocrit value has a negative coefficient, which implies that the higher the lowest hematocrit value, the less likely it is for a patient to be readmitted.

To compute the entries in Table 3.3, we compute the median rank for each variable over the 10 runs and select the top 12 variables. We then run SVM with only these 12 variables to re-optimize $(\boldsymbol{\beta}, \beta_0)$. Table 3.3 reports the 12 variables, their type, the weights β_i found by the 12-variable SVM, and the correlation coefficient of the variable with a readmission. We find that this 12-variable SVM can achieve better results than a Linear SVM using all variables (see Table 3.2). Sparsity, that is, helps as it avoids overfitting and improves the generalization ability of the classifier learned

from the training data.

Table 3.3: Discriminative variables by SLSVM.

Variable	Weight β_i	Corr. with y	Type
any complication	0.53	0.29	Binary
lowest potassium value	-0.49	-0.2	Continuous
postoperative infectious complications	0.46	0.29	Integer
highest sodium value	0.44	0.17	Continuous
highest potassium value	0.35	0.21	Continuous
lowest sodium value	-0.27	-0.17	Continuous
lowest hematocrit value	-0.26	-0.2	Continuous
organ or space surgical site infection	0.22	0.22	Binary
ASA 3	0.25	0.11	Binary
highest level of resident is 6	0.2	0.05	Binary
postoperative superficial incisional SSI	0.19	0.14	Binary
postoperative other occurrences	0.19	0.09	Binary

Most of the variables listed in Table 3.3 are self-explanatory. Any complication and organ or space Surgical Site Infection (SSI) are indicators of whether there was a complication during or after surgery and whether there was an organ or surgical site infection that appears to be related to the surgery. Lowest/highest values of lab results are simply the lowest/highest values recorded during the patient's hospital stay. Postoperative infectious complications is a count of such complications. ASA 3 is an indicator variable of the patient being classified in the American Society of Anesthesiology (ASA) class 3, which implies a patient with severe systemic disease. Highest level of resident indicates the seniority of the resident present in the operation. Postoperative superficial incisional SSI indicates superficial SSI associated with the incision. Postoperative other occurrences indicates any other occurrences after surgery, seeing as complicating the condition of the patient.

We note that the variables listed in Table 3.3 are consensus variables since they were obtained from multiple runs; in each run the classifier was trained with a different

training set of patients. We can state that an SVM model using these variables will be highly predictive of readmissions; this does not necessarily imply that other variables not listed are not informative. As an illustration, an SVM model using the variables in Table 3.3 achieves out-of-sample (i.e., in a test set of patients not seen in training) a false positive rate of 24.67% and a true positive rate of 87.34%. This corresponds to a specific point on the ROC curve. Other combinations of the false/true positive rates are possible; depending on each hospital's tolerance for false positives.

3.4.4 Clustering of Readmitted Patients

As we discussed earlier, the JCC method identifies patient clusters in the readmitted population and, thus, can provide through clustering additional insight for a readmission prediction. In a specific run, the method found two clusters and split the training set such that 69% of the patients are assigned to Cluster 1 and 31% to Cluster 2. Figure 1 depicts how these two clusters separate in terms of the significant variables. As the figure suggests, the patients in Cluster 2 have more complications, lower hematocrit, and are more likely to have postoperative infections and sepsis compared to patients in Cluster 1. In Table 3.4, we list the significant variables identified by JCC in each of the two clusters; the weights have the same role as the $\boldsymbol{\beta}$ in Table 3.3, i.e., the higher the weight the larger the influence of that variable in the prediction.

3.5 Discussion

We have applied different supervised learning methods to predict which postoperative patients are likely to be readmitted within 30-days after discharge. Although few aspects of this exercise may prove to not be easily actionable, it may still be useful to target these individuals, by means of optimizing baseline health conditions, fully addressing postoperative complications prior to discharge, or ensuring close, frequent

Table 3.4: Discriminative variables by JCC.

Variables for the classifier in Cluster 1	Weight	Variables for the classifier in Cluster 2	Weight
functional status prior to surgery	0.09	lowest potassium value	-0.04
highest sodium value	0.12	lowest hematocrit value	-0.05
lowest sodium value	-0.11	postoperative superficial incisional SSI	0.1
highest potassium value	0.15	postoperative deep incisional SSI	0.07
lowest potassium value	-0.15	organ or space SSI	0.1
lowest hematocrit value	-0.15	postoperative urinary tract infection	0.05
No. of comorbidities	0.13	postoperative sepsis	0.09
ASA 1 (normal healthy patient)	-0.11	No. of complications	0.11
ASA 3 (patients with severe systemic disease)	0.1	any complication	0.17
highest level of resident is 6	0.08	postoperative infectious complications	0.2

postoperative follow-up on an outpatient basis. The fact that we can accurately predict which patients are likely to be readmitted, allows physicians to direct more attention to these individuals.

Our methods achieve prediction accuracies exceeding 83.5% (AUC). This substantially outperforms earlier work in the literature which resulted in models with accuracies not exceeding 70% (van Walraven et al., 2010; Low et al., 2015; Wang et al., 2014; Tan et al., 2013; Hartney et al., 2014).

Among the methods examined in this project, special purpose methods we developed such as SLSVM and JCC perform the best (in terms of out-of-sample AUC) and provide interpretable predictions. Using these methods, and given that sparse classifiers are being obtained, each readmission prediction comes with a small list of variables that led to the prediction. The models we derive are easy to implement in software and use in practice; specifically, our SLSVM model uses just 12 variables. In addition, JCC splits patients to be readmitted into two clusters: in one cluster, patients suffer more postoperative complications and infections and have lower hematocrit, making them more likely to be readmitted.

3.6 Conclusions

We conclude by emphasizing the main premise of our analysis. The availability of high-quality and detailed data from BMC NSQIP efforts opens the door to rigorous predictive analytics, which offer ways to positively impact quality of care and reduce hospital costs. Specifically, our methods can predict 30-day readmissions with an accuracy (AUC) of almost 84% by properly weighing a small set of predictive variables that include: complications, lab values (potassium, sodium, hematocrit), surgical site infections, and ASA classification.

Chapter 4

Prescriptive Analytics for 30-day Hospital Readmissions following General Surgery

New financial incentives (such as reduced Medicare reimbursements) have led hospitals to closely monitor their readmission rates and initiate efforts aimed at reducing them. In this context, many surgical departments participate in the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP), which collects detailed demographic, laboratory, clinical, procedure and perioperative occurrence data. The availability of such data enables the development of data science methods which predict readmissions and, as done in This part of the thesis, offer specific recommendations aimed at preventing readmissions. The goal of this project is to explore and develop prescriptive models offering real-time, personalized treatment recommendations for surgical patients during their hospital stay, aimed at reducing the risk of a 30-day readmission using NSQIP Data during 2014, which included information on 722,101 surgeries.

This chapter is based on collaboration with Dimitris Bertsimas, and Michael Lingzhi Li. We would like to thank Dr. George Velmahos at the Massachusetts General Hospital and Dr. George Kasotakis at the Boston Medical Center for useful discussions and for providing access to the NSQIP data.

4.1 Background and Significance

The United States spends $3 trillion annually on healthcare, corresponding to more than 17% of the U.S. GDP and far exceeding the next-highest spender among high-income countries (Squires and Anderson, 2015). While many factors contribute to higher spending, hospital readmissions, defined as an additional admission to address the same issue within 30 days after discharge, are an important and potentially preventable source of excessive resource utilization (James, 2013; Centers for Medicare & Medicaid Services, 2018).

In an effort to reduce unnecessary costs, the Affordable Care Act of 2012 introduced financial penalties for hospitals with readmission rates above the national average. While these measures have so far concentrated on medical conditions (e.g., acute myocardial infarction, congestive heart failure, pneumonia) and common orthopedic procedures (e.g., hip and knee arthroplasty (Li et al., 2019)), the list could expand to include common general surgical procedures.

In anticipation of these changes, Surgical Departments have started to closely monitor their readmission rates, and establishing processes aimed at reducing them. Several authors have sought to determine common causes of readmission after general surgical procedures, and most appear to relate to pre-existing conditions (Gonzalez et al., 2016; Tosoian et al., 2015; Kimbrough et al., 2014; Escobar et al., 2019; Cai et al., 2016), frailty (Min and Hoffman, 2019; Hoffman et al., 2019), and complications after surgery (Kimbrough et al., 2014). Other works have sought to define alternative hospital quality metrics instead of unplanned readmissions (Chhabra et al., 2019; Graham et al., 2019).

In 2005, the American College of Surgeons (ACS) established the National Surgical Quality Improvement Program (NSQIP), which collects detailed demographic, laboratory, clinical, procedure and perioperative occurrence data, currently for Gen-

eral Surgery, and eventually in several subspecialties. The availability of such data, enables the development of data analytics methods relevant to the readmission reduction efforts.

4.2 Objective

While earlier work has primarily focused on readmission predictive methods, there has only been limited attention given to specific interventions with the potential to reduce readmissions; and that has focused mostly on post-discharge care (Chakravarthy et al., 2018; Lasater and McHugh, 2016; Vest et al., 2015). Earlier work on predictive methods for hospitalizations have been successful but focused on specific diseases (Dai et al., 2015; Brisimi et al., 2018a; Brisimi et al., 2019).

The objective of this work is to develop more direct prescriptive methods that offer specific treatment recommendations during the patients' hospital stay with the potential to reduce readmission risk. Our recommended interventions are driven by data; essentially, for each patient, we learn from data what has been effective in preventing a readmission for other "similar" patients. While the methodologies we develop are general and can be applicable to any sort of interventions, we focus on in-hospital treatment because the NSQIP data we leverage contain only such variables. We further focus on the patients' pre-operative hematocrit because it is commonly measured, important for assessing readmission risk, and easily modulated through blood transfusion.

4.3 Prescriptive Analytics

4.3.1 SVM Based Prescriptive Analytics

Prescriptive Support Vector Machines (PSVM) is a prescriptive method we introduce in this chapter that builds on top of SLSVM and JCC.

Suppose we have generated the per-cluster optimal predictive hyperplanes using the JCC approach we described in Section 3.2.3. Let $\mathcal{C}$ be an index set of variables for each patient we can control/modulate by applying certain interventions/therapies. For each patient i in the positive class, with variable vector $\mathbf{x}_i$, we are interested in optimizing the value of the controllable variables $x_{i,d}$, for $d \in \mathcal{C}$, so that the patient is predicted to belong to the negative class.

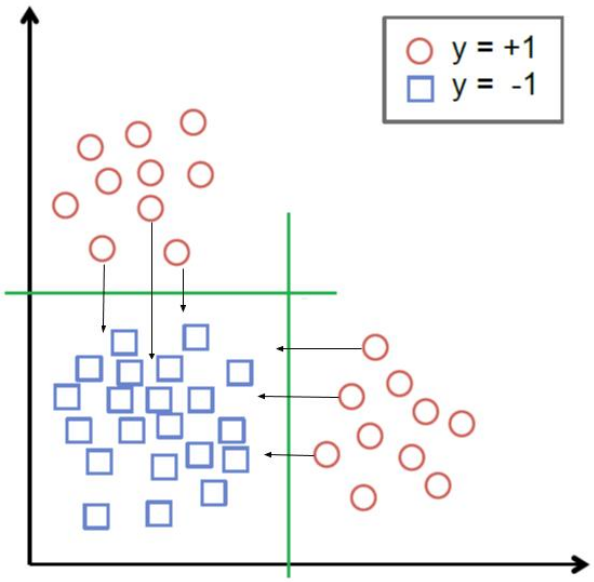

Figure 4·1: The readmitted patients moved from the readmitted side of the prediction hyperplane to the non-readmitted side.

There is, however, a cost for large changes to the value of the controllable variables, which introduces a trade-off between "flipping" the patient to the negative class and implementing interventions that lead to large changes in the controllable variables (see Fig. 4·1). The following formulation optimizes a linear combination of the corresponding two terms in the objective. Specifically, consider a patient i in

cluster l, where $i \in [N^+]$, $\mathbf{x}_i$ is vector of variables characterizing the patient, and $\mathbf{y}_i$ is the patient's variables after applying the prescription/intervention. Let $(\boldsymbol{\beta}^l, \beta_0^l)$ be the coefficients associated with the predictive hyperplane discovered by JCC in the l-th cluster. To determine $\mathbf{y}_i$ we solve the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{y}_i, \xi_i} \quad & \lambda \xi_i + \|\mathbf{y}_i - \mathbf{x}_i\|_p^p \\ \text{s.t.} \quad & \beta_0^l + (\boldsymbol{\beta}^l)' \mathbf{y}_i \leq \xi_i - 1, \\ & y_{i,d} = x_{i,d}, \ \forall d \notin \mathcal{C}, \\ & \xi_i \geq 0, \\ & L_{i,d} \leq y_{i,d} \leq U_{i,d}, \ \forall d \in \mathcal{C}, \end{aligned} \tag{4.1}$$

where $L_{i,d}$ and $U_{i,d}$ are bounds on the controllable variables for each patient i. The parameter λ trades-off the failure to flip the patient to the negative side of the hyperplane with the required change in the patient's characteristics measured by the term $\|\mathbf{y}_i - \mathbf{x}_i\|_p^p$. The higher the value of λ, the more attention is given to the goal of preventing a readmission. To select an appropriate value for λ we can use cross-validation, based on some cost function that accounts for the cost of reducing readmissions and the cost of prescriptions.

Notice that problem (4.1) can be solved independently for each patient who is predicted to belong to the positive class (readmitted). Thus, it is naturally distributed and can obtain a prescription for each at-risk patient with only local computations. The form of the problem (4.1) depends on the selection of the ℓ_p norm; for instance, when $p = 2$, we have a Quadratic Program problem and when $p = 1$, we have a Linear Program problem.

4.3.2 Tree Based Prescriptive Analytics

Optimal Prescriptive Trees (OPT) is a prescriptive method based on Optimal Classification Trees (OCT). OCTs (Bertsimas and Dunn, 2017), use Integer Program to build a decision tree that optimizes the accuracy of predictions over the training set. Such a tree, assigns each patient to a leaf node of the tree and makes a prediction for the patient by a majority vote of other patients assigned to the same leaf. OPTs similarly builds an optimal decision tree but with a modified objective, a linear combination of prediction accuracy and the readmission rate.

4.4 NSQIP Dataset Description and Pre-processing

4.4.1 NSQIP Dataset Description

The ACS-NSQIP was created to improve surgical techniques and outcomes and catalogs over 300 variables on comorbidities, intraoperative events, and 30-day outcomes using prospective random sampling (Ingraham et al., 2010). This study is exempt from the institutional review board approval, as the ACS-NSQIP Participant-Use Data Files contain no protected health information. No 30-day hospital readmissions information are available before 2011 so readmissions information of this version of NSQIP data are available from 2011 to 2014.

The NSQIP dataset contains variables such as:

- Baseline demographic and health care status characteristics (e.g., age, gender, race, body mass index, smoking, diabetes, hypertension requiring medication, admittance from the emergency room).

- Procedure information (e.g., procedure CPT codes, ICD9 codes, the American Society of Anesthesiologists (ASA) classification, wound classification).

- Preoperative, intraoperative, and postoperative variables, including hospital

length of stay information, Surgical Site Infections (SSI, superficial/deep/organ space) and complications (e.g., pneumonia, infections, bleeding, thromboembolic events).

- Laboratories, including pre-operative and post-operative values.

4.4.2 NSQIP Dataset Pre-processing

The NSQIP dataset at our disposal included more than 2.2 million surgeries during 2011-2014. While the NSQIP program provided high-quality manually curated data obtained from trained data abstractors, the variable definitions change over time. Specifically, the definitions of the occurrences listed in Table 4.1 (e.g., sepsis, pneumonia, SSIs) have changed multiple times. The sample size increased and readmission rate decreased over the years in NSQIP dataset. To avoid comparisons among variables with a different meaning, we selected only surgeries that took place during 2014. We included only variables that were continuously monitored and used throughout this period; resulting in a total of 231/187 patient variables for post-op/pre-op. There were a total of 722,101 remaining patients, 39,641 of whom were readmitted within 30 days of discharge, resulting in a readmission rate of 5.49%.

Data pre-processing steps were as follows:

- Only surgeries that took place during 2014 were selected.

- Patients who died within 30 days from discharge were excluded , as these events compete with readmission.

- Categorical variables (e.g., race, discharge, destination, insurance type, CPT codes, ICD9 codes) were numerically encoded by One Hot Encoding to generate one boolean column for each category, which transforms categorical features to a format that works better with classification and regression algorithms.

- For certain pre-operative lab variables, more than 80% of the entries were missing, and they were excluded from the study. For other variables which had missing data, we used a statistical method that uses k-nearest neighbors and clustering to find the most likely value for a missing value (Bertsimas et al., 2017).

- Features with small standard deviation (< 0.005) were removed.

- One of two features which were highly linearly correlated (absolute value of correlation > 0.8) was removed.

- Feature scaling was applied for all features to bring all values into the range $[0, 1]$, i.e., all variables were normalized by subtracting the minimum and dividing by the range.

The variables were further separated into two classes: preoperative variables and postoperative variables. Preoperative variables were those that could be reliably known before or at the main surgical procedure while postoperative variables (including complications) could only be known after the surgery was finished. The reason behind the two classes of variables was that some postoperative variables may be affected by the controllable variables which may be updated according to the prescriptive analytics. For prescriptive analytics results on NSQIP, we excluded postoperative variables in case they might be influenced by the changed controllable variable pre-operative hematocrit (HCT), the volume percentage of red blood cells in the blood. However, we can achieve more accurate prediction on our prescriptive analytics results if the postoperative variables can be used and updated with the changed controllable variable.

4.4.3 Sample Characteristics

For each patient, a total of 231 variables were extracted. Table 4.1 summarizes the baseline demographic and clinical characteristics of the 722,101 patients included in the study. We report the (unnormalized) mean values of the variables over all patients, readmitted patients, and non-readmitted patients, respectively, and only list 60 variables for which the difference between readmitted and non-readmitted patients was the most statistically significant. Specifically, for each variable we computed a two-tailed p-value using Welch's t-test, where the null hypothesis was that the two cohorts (readmitted and non-readmitted patients) have equal means. Hence, the smaller the p-value, the less likely it becomes that the variable means listed in Table 4.1 occurred by chance under the null hypothesis. We note that for indicator variables, the means reported correspond to the fraction of patients satisfying the condition.

Table 4.1: Most statistically significant differences in readmitted and non-readmitted patients.

Variable	All patients	Readmitted	Non-Readmitted	p-value
Estimated Probability of Morbidity	0.06	0.11	0.06	<1E-06
Pre-operative hematocrit	39.67	37.85	39.78	<1E-06
The American Society of Anesthesiology (ASA) Physical Status Classification	2.43	2.78	2.4	<1E-06
Estimated Probability of Mortality	0.01	0.02	0.01	<1E-06
Total operation time in minutes	111.31	148.79	109.14	<1E-06
Return to OR (binary)	0.03	0.24	0.02	<1E-06
Number of Superficial Wound Occurrences	0.02	0.08	0.01	<1E-06
Number of Deep Incisional SSI Occurrences	0.01	0.06	0	<1E-06
Number of Organ/Space SSI Occurrences	0.01	0.11	0.01	<1E-06
Number of Urinary Tract infection Occurrences	0.01	0.06	0.01	<1E-06
Number of Bleeding Transfusions Occurrences	0.06	0.13	0.05	<1E-06
Number of Sepsis Occurrences	0.02	0.1	0.01	<1E-06
Days from Operation to Discharge	2.77	4.51	2.67	<1E-06
OUTPATIENT (if surgical procedure was performed in an outpatient setting)	0.4	0.18	0.42	<1E-06
CPT_Muscl_29x: Casts and endoscopy/arthroscopy	0.03	0.01	0.03	<1E-06
Indicator for any morbidity/complications	0.12	0.49	0.1	<1E-06
no diagnosis of diabetes or diabetes controlled by diet alone.	0.85	0.77	0.85	<1E-06

Discharge Destination: Home	0.9	0.82	0.91	<1E-06
Pre-operative alkaline phosphatase	69.37	82.75	68.59	<1E-06
Pre-operative serum albumin	3.95	3.78	3.96	<1E-06
ICD9 550: Inguinal hernia	0.04	0.01	0.04	<1E-06
Work Relative Value Unit (a metric of surgical complexity)	16.35	19.75	16.16	<1E-06
Age	56.41	60.42	56.17	<1E-06
Hypertension requiring medication	0.45	0.57	0.44	<1E-06
Elective Surgery (binary)	0.8	0.69	0.81	<1E-06
Number of Pneumonia Occurrences	0.01	0.05	0.01	<1E-06
Bleeding disorders	0.04	0.09	0.04	<1E-06
Open wound/wound infection	0.03	0.07	0.03	<1E-06
Number of DVT/Thrombophlebitis Occurrences	0.01	0.04	0	<1E-06
CPT_Muscl_23x-25x: Shoulder arm wrist hand	0.03	0.01	0.03	<1E-06
CPT_CAT_2x: Musculoskeletal system	0.22	0.16	0.23	<1E-06
Discharge Destination: Skilled Care Not Home	0.06	0.11	0.05	<1E-06
Pre-operative serum creatinine	0.99	1.18	0.97	<1E-06
Pre-operative BUN	16.32	18.2	16.21	<1E-06
CPT_CAT_33x-37x: Cardiovascular system	0.07	0.12	0.06	<1E-06
Number of Pulmonary Embolism Occurences	0	0.03	0	<1E-06
History of severe COPD	0.04	0.09	0.04	<1E-06
Disseminated cancer	0.02	0.06	0.02	<1E-06
Number of Wound Disruption Occurrences	0	0.03	0	<1E-06
Functional health status Prior to Surgery	0.03	0.07	0.03	<1E-06
Pre-operative serum sodium	138.82	138.46	138.87	<1E-06
Steroid use for chronic condition	0.04	0.07	0.03	<1E-06
Currently on dialysis (pre-op)	0.01	0.04	0.01	<1E-06
TRANST_Not transferred (admitted from home)	0.96	0.93	0.96	<1E-06
CPT_Digestive_441x: Intestines - excision	0.02	0.05	0.02	<1E-06
Number of Septic Shock Occurrences	0.01	0.03	0	<1E-06
Wound classification 4: Dirty/Infected	0.05	0.08	0.05	<1E-06
Surgical Specialty: Gynecology	0.07	0.05	0.08	<1E-06
CPT_Cardio_35x: Repairs bypasses etc.	0.04	0.07	0.03	<1E-06
Organ/Space SSI PATOS (Present at the Time of Surgery)	0	0.02	0	<1E-06
CPT_Digestive_48x: Pancreas	0.01	0.03	0.01	<1E-06
CPT_Digestive_49x: Abdomen Peritoneum and Omentum	0.11	0.08	0.12	<1E-06
No dyspnea	0.05	0.08	0.05	<1E-06
Pre-operative International Normalized Ratio (INR) of PT (Prothrombin Time) values	1.07	1.1	1.07	<1E-06
CPT_CAT_60x: Endocrine system	0.03	0.02	0.03	<1E-06
Number of Progressive Renal Insufficiency Occurrences	0	0.02	0	<1E-06

Number of Myocardial Infarction Occurences	0	0.02	0	<1E-06
CPT_CAT_4x: Digestive system	0.41	0.46	0.4	<1E-06
Number of Unplanned Intubation Occurrences	0.01	0.02	0.01	<1E-06
Discharge Destination: Rehab	0.03	0.05	0.03	<1E-06

4.4.4 Controllable Variables

We consider three types of controllable variables on which to intervene using prescriptive analytics:

- Pre-operative lab tests: sodium, Blood Urea Nitrogen (BUN), serum creatinine, serum albumin, bilirubin, SGOT (Serum Glutamic-Oxaloacetic Transaminase), alkaline phosphatase, White Blood Cell count (WBC), hematocrit (HCT), platelet count, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), and International Normalized Ratio (INR) of PT values.

- Length of stay at the hospital: total length of stay, days from admission to operation, days from operation to discharge.

- SSI (Surgical Site Infection) or Infection: occurrences of deep incisional SSI, occurrences of organ space SSI, and post-operative occurrences of Urinary Tract Infection (UTI).

Pre-operative lab values could be altered through appropriate medications and treatment before the operation to bring them closer to levels not associated with readmission. The length of stay at the hospital could be to shortened, or lengthened as appropriate. Recommendations can also target the tightening of infection control measures that affect the variables described in the third item above. In the work we report in this chapter we focus on the pre-operative hematocrit (HCT), as it is a variable that can be directly impacted (increased) through blood transfusion. The

predictive models also suggest that pre-operative hematocrit (HCT) is one of the most important controllable variables.

4.4.5 Second Order Effects of Transfusion

There is evidence in the literature to suggest that perioperative transfusion (packed red blood cells on the day of surgery or one day after) or postoperative transfusion could potentially lead to adverse outcomes (Hollis et al., 2016; Goel et al., 2018; Whitlock et al., 2015). To account for the second order effects of transfusion on other variables, we modified the other characteristics of each test patient under transfusion using the regression models constructed against HCT. Specifically, we assumed the administration of blood would have a secondary effect on the other controllable variables that are highly correlated with HCT (absolute value$>$ 0.1). There included pre-op creatinine, international normalized ratio, prothrombin time, albumin, mortality probability, and morbidity probability.

We used generalized linear regression models to predict the effect on MORTPROB and MORBPROB since they have bounded values, i.e., [0,1]. In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. We applied the logit transformation, i.e., the logit of the probability is the logarithm of the odds,

$$logit(p) = \log(\frac{p}{1-p}),$$

to values of MORTPROB and MORBPROB.

Table 4.2 lists the variables highly correlated with HCT, the correlation with HCT

(corr. with HCT) and the coefficient of determination (R2) in the linear regression models.

Table 4.2: Variables highly correlated with HCT and the coefficient of determination.

	corr. with HCT	R2
PRALBUM	0.41	0.40
PRPT	-0.23	0.20
PRINR	-0.15	0.12
PRCREAT	-0.11	0.66
logit(MORTPROB)	-0.13	0.95
logit(MORBPROB)	-0.10	0.92

4.5 Performance Evaluation and Experimental Results

4.5.1 Prediction Accuracy

For the predictive task, we evaluated the methods across three distinct splits of the data into a training and a test dataset. Each split, randomly selects 80% of the data to form the training set and keeps the remaining 20% as the test set, on which model performance is evaluated. We used 3-fold cross-validation (with only training data) for parameter tuning. The mean (Avg.) and standard deviation (Std.) of AUC for each predictive method is reported in Table 4.3 and Table 4.4; the top table considers predictions using only pre-operative (PRE-op) variables, while the bottom table evaluates models using pre-operative and post-operative variables (POST-op).

The random forest (Breiman, 2001) is a large collection of decision trees and it classifies by averaging the decisions of each tree. Another method we implemented was logistic regression, which is a widely used as a base for comparison in medical machine learning studies. In this study, a logistic regression model was fitted with an additional regularization term: an ℓ_2-norm term (similar to ridge regression) (Friedman et al., 2001). Finally, we implemented two additional methods. One method is a class of

feed-forward artificial neural networks (NN), called a multilayer perception, which consists of at least three layers of nodes. Except for the input nodes, each node uses a nonlinear activation function. It can distinguish data that is not linearly separable due to its multiple layers and non-linear activation. Another method is Gradient boosting machine (GBM), also referred to as gradient-boosted decision trees, is a popular machine-learning algorithm used for regression and classification tasks.

Methods were implemented in Python (Python Software Foundation, https://www.python.org/)(Pedregosa et al., 2011) and Matlab (MathWorks, Natick, MA). For RF, the number of trees grown was 500. Cross-validation was used to tune parameters of all the method, e.g., the number of variables randomly sampled as candidates at each split for RF, regularization strength for SLSVM and ℓ_2LR.

Table 4.3: Performance of predictive models, PRE-op.

PRE-op					
Methods	Split I	Split II	Split III	Avg.	Std.
ℓ_2LR	72.55%	72.61%	72.97%	72.71%	0.23%
SLSVM	72.51%	72.58%	72.91%	72.67%	0.21%
RF	73.39%	73.24%	73.59%	73.41%	0.18%
GBM	73.49%	73.51%	73.78%	73.59%	0.16%
NN	72.50%	72.74%	73.18%	72.81%	0.34%

The methods with postoperative variables were better than preoperative variables since more useful features were added to the dataset. AUCs of the methods were similar since the NSQIP dataset contained large amount data and sufficient features of high quality and data pre-processing steps were well done. Low standard deviations across different splits for all methods, imply that the predictive power is not greatly

Table 4.4: Performance of predictive models, POST-op.

POST-op					
Methods	Split I	Split II	Split III	Avg.	Std.
ℓ_2LR	84.20%	84.36%	84.64%	84.40%	0.22%
SLSVM	84.25%	84.38%	84.68%	84.44%	0.22%
RF	85.24%	85.34%	85.67%	85.41%	0.22%
GBM	87.06%	87.32%	87.80%	87.39%	0.38%
NN	83.03%	83.06%	84.00%	83.36%	0.55%

impacted by the choice of the training data subset.

4.5.2 Important Variables

For each variable we computed a two-tailed p-value using Welch's t-test, where the null hypothesis was that the two cohorts (readmitted and non-readmitted patients) have equal means. Howerver, there were 160 variables whose p-value $< 1E-06$ since the dataset was extremely large.

An analysis of the most statistically significant differences in readmitted vs. non-readmitted patients, reveals (cf. Table 4.1) that the former tend to be patients who underwent vascular surgery, or surgeries involving the pancreas. In contrast, surgeries involving the endocrine system, or Abdomen, Peritoneum, and Omentum are less likely to lead to a readmission. Furthermore, readmitted patients tend to have more complications (e.g., septic shock, bleeding, pneumonia, organ/space/deep incisional SSIs, renal insufficiency) and show higher incidence of return to OR and unplanned intubation.

4.5.3 Causal Inference and Feature Selection

The task of structure learning for Bayesian networks refers to learning the structure of a directed acyclic graph (DAG) from data. Learning the structure of Bayesian networks can be complicated for two main reasons: difficulties in inferring causality and the super-exponential number of directed edges that could exist in a dataset. The implementations in python were done with the package pomegranate (Schreiber, 2017) that is implemented in cython for speed. The package pomegranate currently supports exact Bayesian Network Structure Learning through a shortest path Dynamic Program algorithm, an A* algorithm, a greedy algorithm, the Chow-Liu tree algorithm, and a constraint-graph based algorithm which can significantly speed up structure learning given any prior knowledge of the interactions between variables. The default is a greedy algorithm that greedily chooses a topological ordering of the variables, but optimally identifies the best parents for each variable given this ordering (Schreiber, 2017). Currently, pomegranate only supports discrete Bayesian networks, meaning that the values must be categories.

A total of eight discrete variables were selected and created from NSQIP according to doctors' advice. Variables that were identified to be predictive of readmissions included: any Surgical Site Infection (SSI), including organ/space SSI, superficial/deep Incisional SSI, organ/space SSI PATOS superficial/deep Incisional SSI PATOS; prsodm01 (pre-operative sodium$<$ 132); asacla01 (ASA classification $\geq$ 3); dischdest01 (discharge destination is not home); age01 (age $\geq$ 65); hxcopd (history of severe COPD); returnor (return to OR); and ascites.

The relationship among discrete variables was inferred by different methods and assumptions. We have prior information about how groups of nodes are connected to each other and want to exploit that. This can take the form of a global ordering, where variables can be ordered in such a manner that variables are a part of these layers

and can only have parents in another layer. These constraints can dramatically speed up structure learning through the use of loose general prior knowledge. From our experiments, the best interpretable result is the exact learning using prior information and constraint graphs as follows. We assume 4 layers sorted by time order as in Figure 4·2, i.e., demographics (age01), pre-op, post-op, readmission.

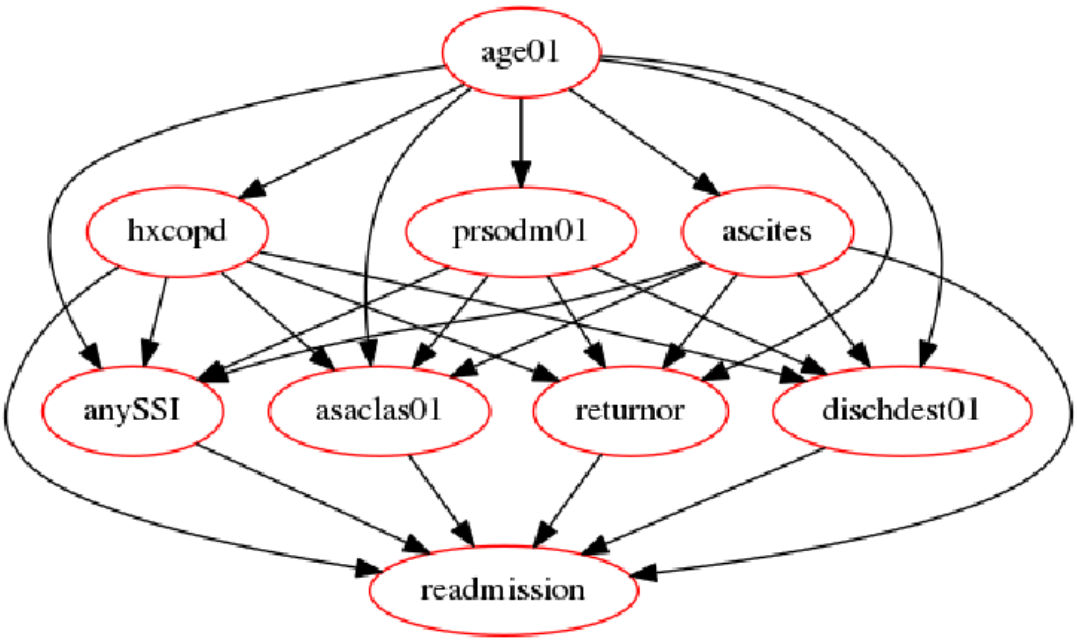

Figure 4·2: Bayesian Network structure learning and and feature selection for readmissions.

From the Bayesian network structure learning, we can deduce that age01 and prsodm01 are not directly related with readmission. It follows that removing age01 and prsodm01 should not affect prediction. As a result, we removed them and attempt to predict using only 6 features. The out-of-sample AUCs obtained with these 6 features are almost the same when using 8 features (see Table 4.5), which verifies the feature selection power of Bayesian network. We randomly chose 80% and 10% of the patients in the 2011-2014 dataset to form the training and validation set and retained the remaining 10% of the patients as a test set. We used class weight

inversely proportional to class frequencies in the data to deal with imbalanced data. No normalization method was used since they were all binary. In Table 4.5, the columns AUC_8 and AUC_6 correspond to AUC with 8 or 6 features evaluated by the various methods.

Table 4.5: Prediction performance with 8 or 6 features evaluated by the various classification methods.

Settings	AUC_8	AUC_6
RF	76.24%	75.84%
ℓ_2LR	76.17%	75.79%
ℓ_1LR	76.17%	75.79%
ℓ_1SVM	76.16%	75.77%
ℓ_2SVM	75.05%	75.41%

4.5.4 Prescriptive Results

Besides the prediction of 30-day hospital readmissions, prescriptive analytics is the key to preventing or reducing 30-day hospital readmissions after general surgical procedures. We consider the problem of using the blood transfusion to change the hematocrit (HCT). Operationally, this is achieved through blood transfusion before the surgery. Consistent with medical practice, the maximum change in the hematocrit level is limited to 9%, corresponding, roughly, to 3 standard (300cc) bags of blood, which can be considered as a safe upper limit for blood transfusion. Since any such intervention has to be applied before the surgery, the methods we develop will only use pre-operative variables (a total of 187 such variables in the dataset after one-hot encoding).

We need to establish a baseline for the actionable variable under all treatments, used by OPT to learn an effective treatment. In the NSQIP data, we utilize the

TRANSFUS variable which indicates whether a pre-operative blood transfusion took place. However, there is no information on the amount of blood transfused. We formed our baseline treatment with the assumption that everyone who has a hematocrit value over 30 had at most 1 bag of blood transfused, as the common operative transfusion threshold is 30 (Wang and Klein, 2010). Then, we add additional bags of blood with decreasing hematocrit levels to bring the patient's hematocrit level above 30. The full table of assumed baseline treatment is shown in Table 4.6.

Table 4.6: Assumed baseline treatment.

Assumed Transfusion Factuals	Condition in Data
No blood transfusion	TRANSFUS=0
1 bag of blood	HCT>30 and TRANSFUS=1
2 bags of blood	27<HCT<30 and TRANSFUS=1
3 bags of blood	HCT<27 and TRANSFUS=1

For the prescriptive task as well, we evaluated the PSVM and OPT methods across the 3 distinct splits of the data. For each split, we trained PSVM and OPT in the training set and then apply the method to obtain a recommended number of bags of blood to be transfused for each patient whose HCT is less than 30 (HCT$<$ 30) in the test set. We evaluated the outcome for each test patient using four different predictive methods: ℓ_2LR, RF, GBM, and NN. For each predictive model, we chose a threshold so that the predicted readmission rate equals the ground truth readmission rate in the training dataset.

We reported in Table 4.7 the percentage of readmissions prevented in the test set, defined as the ratio (in %) of (i) the number of patients with HCT$<$ 30 originally predicted to be readmitted (assuming no treatment) and now predicted not to be readmitted (after treatment), over (ii) the number of patients with HCT$<$ 30 predicted to be readmitted (assuming no treatment). We also reported the average

number of bags of blood per patient under the recommended treatment.

The first column of Table 4.7 lists the predictive models used to evaluate the effect of treatment, the 2nd and 4rd columns show the percentage of readmissions prevented using the OPT and PSVM prescriptions, the 3th and 5th columns show the average number of bags per patient when using OPT and PSVM prescriptions, and the last column reports a baseline percentage of readmissions prevented, assuming any patient (with HCT$<$ 30) in the test set gets 1 bag of blood.

Across a wide variety of different ground truths, two separate prescriptive methods (OPT and PSVM) are able to prescribe blood transfusion treatments that reduce predicted readmissions for patients with HCT$<$ 30, with the decrease ranging from 4.81% to 19.51% for OPT and 3.78% to 19.08% for PSVM, and with transfusions in the range of 300cc of blood per patient on average. To put the achievable readmission reductions into context, if one could reduce by the mean percentage we achieved (12%) all 30-day readmissions of patients with HCT$<$ 30 across the U.S. (over 10,000 per year), the cost savings would amount to $20.3 million on an annual basis (Bailey et al., 2019).

4.6 Conclusions

A variety of classification methods were applied to predict 30-day readmissions. Two prescriptive methods were developed to recommend pre-operative blood transfusions to increase the patient's hematocrit with the objective of preventing readmissions. The effect of these interventions was evaluated using several predictive models.

Predictions of 30-day readmissions based on the entire collection of NSQIP variables achieve an out-of-sample accuracy of 87% (Area Under the Curve—AUC). Predictions based only on pre-operative variables have an accuracy of 74% AUC, out-of-sample. Personalized interventions, in the form of pre-operative blood transfusions

Table 4.7: Prescriptive analytics performance evaluated by the various classification methods.

Split I	OPT	Average bags for patients (HCT<30)	PSVM	Average bags for patients (HCT<30)	decrease (1bag)
LR	9.27%	0.97	9.49%	1.06	6.78%
RF	14.50%	0.97	14.03%	1.06	9.03%
GBM	4.81%	0.97	3.78%	1.06	3.19%
NN	16.37%	0.97	16.50%	1.06	8.03%
Split II	OPT	Average bags for patients (HCT<30)	PSVM	Average bags for patients (HCT<30)	decrease (1bag)
LR	9.27%	0.95	9.63%	1.08	5.97%
RF	13.08%	0.95	11.30%	1.08	8.24%
GBM	5.34%	0.95	4.20%	1.08	2.84%
NN	18.96%	0.95	19.08%	1.08	9.22%
Split III	OPT	Average bags for patients (HCT<30)	PSVM	Average bags for patients (HCT<30)	decrease (1bag)
LR	10.25%	0.98	10.65%	1.07	6.94%
RF	15.84%	0.98	14.60%	1.07	8.89%
GBM	8.66%	0.98	5.72%	1.07	4.41%
NN	19.51%	0.98	18.37%	1.07	8.46%

identified by the prescriptive methods, reduce readmissions by 12%, on average, for patients considered as candidates for pre-operative transfusion (pre-operative hematoctic < 30).

This study is among the first to develop a methodology for making specific, data-driven, personalized treatment recommendations to optimize desired metrics, such as the 30-day readmission rate. These recommendations come in a form that could be interpreted by physicians. The reported predicted reduction in readmissions is significant enough, with the potential to lead to more than $20 million in savings in the U.S. annually, which could motivate a clinical trial to test the proposed methods.

Chapter 5

Parameter Estimates for Regularized Mixed Linear Regression Models

We consider *Mixed Linear Regression (MLR)*, where training data have been generated from a mixture of distinct linear models (or clusters) and we seek to identify the corresponding coefficient vectors.

5.1 Introduction

Mixed Linear Regression (MLR) (Yi et al., 2014; Zhong et al., 2016) is also known as mixtures of linear regressions (Chaganty and Liang, 2013) or cluster-wise linear regression (Park et al., 2017). It involves the identification of two or more linear regression models from unlabeled samples generated from an unknown mixture of these models. This can be seen as a joint clustering and regression problem. The problem is related to the identification of hybrid and switched linear systems (Paoletti et al., 2007; Vidal, 2008) and has many diverse applications. There are many applications with MLR models in marketing (DeSarbo and Cron, 1988), health insurance claims (Gitman et al., 2018), rainfall prediction (Bagirov et al., 2017), and pavement condition prediction (Luo and Chou, 2006). In this study, we focus on the fundamental problem of establishing strong consistency of parameter estimates, i.e., establishing that the estimated parameters converge to their true values as the number of the training samples grows.

MLR is typically solved by local search such as *Expectation Maximization (EM)*

or alternating minimization, where one alternates between clustering and regression. It has recently been shown that EM converges to the true parameters if it starts from a small enough neighborhood around them (Yi et al., 2014; Yi and Caramanis, 2015; Balakrishnan et al., 2017).

Much effort has focused on the case where training samples are not perturbed by noise. (Yi et al., 2016) proposed a combination of tensor decomposition and alternating minimization, showing that initialization by the tensor method allows alternating minimization to converge to the global optimum at a linear rate with high probability (w.h.p.). (Zhong et al., 2016) proposed a non-convex objective function with a tensor method for initialization so that the initial coefficients are in the neighborhood of the true coefficients w.h.p. (Hand and Joshi, 2018) proposed a second-order cone program, showing that it recovers all mixture components in the noiseless setting under conditions that include a well-separation assumption and a balanced measurement assumption on the data. (Li and Liang, 2018) considers data generated from a mixture of Gaussians, showing global convergence of the proposed algorithm with nearly optimal sample complexity.

Establishing convergence (w.h.p.) and consistency results for MLR in the presence of noise is much harder. (Chaganty and Liang, 2013) develops a provably consistent estimator for MLR that relies on a low-rank linear regression to recover a symmetric tensor, which can be factorized into the parameters using a tensor power method. (Chen et al., 2014) provides a convex optimization formulation for MLR with two components and upper bounds on the recovery errors under subgaussian noise assumptions. (Yen et al., 2018) studies a *MixLasso* approach but convergence results are limited to the objective function and not the solution. (Yin et al., 2018) develops an algorithm based on ideas from sparse graph codes; the convergence results however are asymptotic under Gaussian noise and for MLR with only two components.

From 1991 to 2015, algorithmic advances in integer optimization combined with hardware improvements have resulted in an astonishing 450 billion factor speedup in solving Mixed Integer Program (MIP) problems (Bertsimas and King, 2015). Recently, an increasing number of machine learning and statistics problems have been tackled using MIP methods (Bertsimas and King, 2015; Bertsimas et al., 2016; Xu et al., 2016; Brisimi et al., 2018b; Brisimi et al., 2019). (Bertsimas and Shioda, 2007) proposed an MIP formulation for MLR and a pre-clustering heuristic approach to solve large-scale problems. (Park et al., 2017) proposed a Mixed-Integer Quadratic Program (MIQP) formulation for MLR. (Angün and Altınoy, 2019) proposed a way to determine the number of clusters for MLR based on solving a series of Mixed-Integer Linear Programs.

At the same time, *regularization* in learning problems has become widespread, following the early success of LASSO (Tibshirani, 1996; Chatterjee, 2013). A recent study has obtained regularization as a consequence of solving a robust learning problem (see, (Chen and Paschalidis, 2018) and references therein).

Our convergence analysis leverages techniques for proving strong consistency of least-squares estimates for linear regression, but under weaker assumptions. The related literature is substantial. A breakthrough paper (Lai et al., 1982) established strong consistency of least-squares estimates for stochastic regression models. Asymptotic properties and strong consistency of least-squares parameter estimates have been studied in many areas including system identification and adaptive control (Chen and Guo, 2012), econometric theory (Nielsen, 2005) and time series analysis (Wei et al., 1987).

Identifiability of the parameters is a necessary condition for the existence of consistent estimators. In (Hennig, 2000), the identifiability of the parameters of MLR with Gaussian errors is investigated and such models cause other identifiability prob-

lems than do simple Gaussian mixtures. However, no researcher above in machine learning area has thoroughly considered the identifiability problems for MLR for the noisy case. In linear regression, we can perform standard normalization to get rid of the constant term. However, in MLR, this is difficult since the clusters are unknown. Therefore, the assumption that the variables $\mathbf{x}$ follow the zero-mean normal distribution in (Yi et al., 2014; Zhong et al., 2016; Yi et al., 2016; Li and Liang, 2018; Yin et al., 2018) or zero-mean sub-Gaussian distribution in (Chen et al., 2014) may not be appropriate.

Our contributions can be summarized as follows:

- We introduce a general MIP formulation for MLR subject to norm-based regularization constraints. The formulation is general enough to include regularization constraints on the regression coefficients.

- We propose an identifiable condition and establish that optimal solutions of the MIP converge almost surely (rather than w.h.p.) to the true parameters in the noiseless case as the sample size increases. Subject to cluster separability assumptions, we also establish that MIP solutions can identify the proper cluster for each given sample. To the best of our knowledge, our study is the first to study strong consistency of parameter estimates for MLR under general noise conditions and general feature conditions rather than convergence with high probability.

- For the special case of a single cluster, we show that the MIP solution converges to the true parameter vector in the presence of noise satisfying a martingale difference assumption (Lai et al., 1982; Chow and Teicher, 2012; Chen and Guo, 2012). For multiple clusters in the presence of noise, we not only derive the convergence conditions, i.e., stronger identifiable condition and cluster consistency condition but also provide a counterexample, suggesting that one can not

in general recover the true parameters if the cluster consistency condition is violated.

- Besides the convergence results, we propose a novel method to apply MLR for practical large scale prediction problems. We present experimental prediction results and compare our prediction algorithm with mean absolute error regression and Random Forest regression in terms of both accuracy and interpretability.

The rest of the chapter is organized as follows. Section 5.2 presents the formulation of the problem. The convergence results and identifiable condition are given in Section 5.3. Section 5.4 presents how to apply MLR for large scale prediction problems using real data. Section 5.5 presents numerical results. Section 5.6 presents the experimental prediction results for two datasets. Section 5.7 draws conclusions.

5.2 Problem Formulation

Consider the MLR model where data $(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1},\ i \in [n]$ are generated by

$$y_i = \mathbf{x}_i' \boldsymbol{\beta}_k + \epsilon_i,\ \text{for some } k \in [K],$$

where $\{\boldsymbol{\beta}_k,\ \forall k \in [K]\}$ are the ground truth coefficient vectors. The problem is to estimate these parameters from data. Given a training dataset $\{(\mathbf{x}_i, y_i),\ i \in [n]\}$, we formulate the problem as the following MIP:

$$
\begin{aligned}
\min_{\boldsymbol{\beta}_k, t_i, c_{ki}} \ & \frac{1}{n} \sum_{i \in [n]} t_i^p && (5.1)\\
\text{s.t. } & t_i - (y_i - \mathbf{x}_i' \boldsymbol{\beta}_k) + M(1 - c_{ki}) \geq 0, \ i \in [n], k \in [K], \\
& t_i + (y_i - \mathbf{x}_i' \boldsymbol{\beta}_k) + M(1 - c_{ki}) \geq 0, \ i \in [n], k \in [K], \\
& \sum_{k \in [K]} c_{ki} = 1, \ i \in [n], \\
& c_{ki} \in \{0, 1\}, \ i \in [n], k \in [K], \\
& t_i \geq 0, \ i \in [n], \\
& \|\boldsymbol{\beta}_k\|_q \leq d_{k,q}, \ k \in [K],
\end{aligned}
$$

where M is a large constant (big-M). Notice that when $c_{ki} = 1$, the first two constraints imply $t_i \geq |(y_i - \mathbf{x}_i' \boldsymbol{\beta}_k)|$, whereas $c_{ki} = 0$ implies that no constraint is imposed on t_i since M is a large positive constant. At optimality, Formulation (5.1) minimizes a p-norm loss function for the regression problem and assigns each data point to the cluster achieving minimal loss. The last constraint in Formulation (5.1) imposes a q-norm regularization constraint to each coefficient vector $\boldsymbol{\beta}_k$ and $d_{k,q}$ are given constants where $q \in \{0, 1, 2\}$. In statistics and machine learning, regularization is widely used to help prevent models from overfitting the training data (Chen and Paschalidis, 2018). A Bayesian understanding of regularization is that regularized least squares are equivalent to priors on the solution to the least squares problem. For $p, q = 1$, Formulation (5.1) is a linear MIP, while if either p or q (or both) are equal to 2 it is a quadratic MIP. Both can be solved by modern MIP solvers.

Formulation 5.1 is equivalent to the following MIP formulation:

$$\begin{aligned}
\min_{\boldsymbol{\beta}_k, \mathbf{z}_i, c_{ki}} \quad & \frac{1}{n} \sum_{i \in [n]} |y_i - \mathbf{x}_i^{'} \mathbf{z}_i|^p \\
\text{s.t.} \quad & \mathbf{z}_i = \sum_{k \in [K]} c_{ki} \boldsymbol{\beta}_k, \quad \forall i \in [n], \\
& \sum_{k \in [K]} c_{ki} = 1, \quad \forall i \in [n], \\
& c_{ki} \in \{0, 1\}, \quad \forall i \in [n], \forall k \in [K], \\
& \|\boldsymbol{\beta}_k\|_q \leq d_{k,q}, \quad \forall k \in [K].
\end{aligned} \tag{5.2}$$

Without the regularization constraint, another equivalent unconstrained version of Formulation (5.1) is

$$\min_{\boldsymbol{\beta}_k} \frac{1}{n} \sum_{i \in [n]} \min_{k \in [K]} |y_i - \mathbf{x}_i' \boldsymbol{\beta}_k|^p. \tag{5.3}$$

Let $\{\boldsymbol{\beta}_k^n, t_i^n, c_{ki}^n,\ k \in [K],\ i \in [n]\}$ denote an optimal solution to Formulation (5.1); we use a superscript n to explicitly denote dependence on the training set. Define $\Omega_k^n = \{i \in [n] :\ y_i = \mathbf{x}_i' \boldsymbol{\beta}_k + \epsilon_i\}, \forall k \in [K]$ and $\hat{\Omega}_k^n = \{i \in [n] :\ c_{ki}^n = 1\}, \forall k \in [K]$, which form the true and the estimated partition of the training set into the K clusters, respectively. We can show the following lemma.

Lemma 1 *For some fixed $q \in \{0, 1, 2\}$, if $\max_{k \in [K]} \|\boldsymbol{\beta}_k^n\|_q \leq \min_{k \in [K]} d_{k,q}$, then $\forall k \in [K]$,*

$$\begin{aligned}
\hat{\Omega}_k^n &\subset \{i \in [n] :\ |y_i - \mathbf{x}_i' \boldsymbol{\beta}_k^n| = \min_{m \in [K]} |y_i - \mathbf{x}_i' \boldsymbol{\beta}_m^n|\}, \\
\hat{\Omega}_k^n &\supset \{i \in [n] :\ |y_i - \mathbf{x}_i' \boldsymbol{\beta}_k^n| < \min_{m \neq k \in [K]} |y_i - \mathbf{x}_i' \boldsymbol{\beta}_m^n|\}
\end{aligned}$$

Proof Given the assumption, i.e., regularization constraint is satisfied, the equivalent formulation is (5.3). For all $i \in \hat{\Omega}_k^n$, we will prove $i \in \{i \in [n] :\ |y_i - \mathbf{x}_i' \boldsymbol{\beta}_k^n| =$

$\min_{m\in[K]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|\}$ by contradiction. If

$$|y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n| \neq \min_{m\in[K]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|,$$

then there exists $m \neq k \in [K]$, such that,

$$|y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n| > |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|.$$

If we move the sample i from cluster $\hat{\Omega}_k^n$ to cluster $\hat{\Omega}_m^n$, then the objective function value will decrease by $\frac{1}{n}(|y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n|^p - |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|^p) > 0$, which contradicts the optimality of the optimal objective function value and the cluster partition $\hat{\Omega}_k^n, \forall k \in [K]$.

Similarly,

$$\forall i \in \{i \in [n] : \ |y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n| < \min_{m\neq k\in[K]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|\},$$

we prove $i \in \hat{\Omega}_k^n$ by contradiction. Assume $i \notin \hat{\Omega}_k^n$, then $i \in \hat{\Omega}_m^n$ for some $m \in [K]$. If we move the sample i from cluster $\hat{\Omega}_m^n$ to cluster $\hat{\Omega}_k^n$, then the objective function value will decrease by $\frac{1}{n}(|y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|^p - |y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n|^p) > 0$, which contradicts the optimality of the optimal objective function value and the cluster partition $\hat{\Omega}_k^n, \forall k \in [K]$. ■

Lemma 2 *If $p = 2$, for all $k \in [K]$, the least-squares estimates of model parameters from Formulation (5.1) without imposing the last regularization constraint can also be written as:*

$$\begin{aligned}\boldsymbol{\beta}_k^n &= (\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j y_j,\\ &= (\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j(\mathbf{x}_j^{'}\boldsymbol{\beta}_{l_j}+\epsilon_j)\\ &= (\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\mathbf{x}_j^{'}\boldsymbol{\beta}_{l_j}+(\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in\hat{\Omega}_k^n}\mathbf{x}_j\epsilon_j.\end{aligned}$$

where l_j is the ground truth cluster label for every sample j.

Proof First, the ground truth cluster label is l_j for every sample j, i.e., $y_j = \mathbf{x}_j'\boldsymbol{\beta}_{l_j}+\epsilon_j$. Then, in (5.1), the objective function can be decomposed as

$$\begin{aligned}\sum_{j\in[n]}|y_j-\mathbf{x}_j^{'}\mathbf{z}_j|^2 &= \sum_{l\in[K]}\sum_{j\in\hat{\Omega}_l^n}|y_j-\mathbf{x}_j^{'}\mathbf{z}_j|^2\\ &= \sum_{j\in\hat{\Omega}_k^n}|y_j-\mathbf{x}_j^{'}\mathbf{z}_j|^2+\sum_{l\in[K],l\neq k}\sum_{j\in\hat{\Omega}_l^n}|y_j-\mathbf{x}_j^{'}\mathbf{z}_j|^2\end{aligned}$$

For fixed $k\in[K]$, if we fix the second term of the above, i.e., all $\mathbf{z}_j=\boldsymbol{\beta}_l^n, j\in\hat{\Omega}_l^n, l\neq k$ and only allow $\mathbf{z}_j, j\in\hat{\Omega}_k^n$ in the first term to change in the feasible region, we have $\boldsymbol{\beta}_k^n = \arg\min_{\mathbf{z}\in\mathbb{R}^d}\sum_{j\in\hat{\Omega}_k^n}|y_j-\mathbf{x}_j^{'}\mathbf{z}|^2$. By taking the derivative with respect to $\mathbf{z}$, we obtain the result. ■

The least-squares estimates of MLR model parameters have a similar form with the linear regression case. If the hidden partitions are known, i.e., $\hat{\Omega}_k^n=\Omega_{\pi(k)}^n$, where π is a permutation on the cluster indices, the strong consistency can be solved by K separate models and the techniques in the linear regression case (Lai et al., 1982; Chen and Guo, 2012). But the difficulty is that the sets $\hat{\Omega}_k^n$ are coupled with $\boldsymbol{\beta}_k^n, \forall k\in[K]$ and the noise sequences $\{\epsilon_j\}$ when $\Omega_{\pi(k)}^n$ are unknown.

In order to analyze the strong consistency of parameter estimates obtained by the MIP formulation, we introduce the following assumptions.

Assumption 1 (A1) *The clusters are different and not degenerate, i.e.,* $\boldsymbol{\beta}_k \neq \boldsymbol{\beta}_h$, $\forall k \neq h$, *and* $|\Omega_k^n| = \Theta(n)$, $\forall k \in [K]$.

Assumption 2 (A2) *The noise sequence* $\{\epsilon_i, \mathscr{F}_i\}$ *is a martingale difference sequence, where* $\{\mathscr{F}_i\}$ *is a sequence of increasing* σ*-fields, and* $\sup_i \mathsf{E}[|\epsilon_i|^2|\mathscr{F}_{i-1}] < \infty$, *a.s.*

Assumption 3 (A3) *The noise sequence* $\{\epsilon_i, \mathscr{F}_i\}$ *is a martingale difference sequence, where* $\{\mathscr{F}_i\}$ *is a sequence of increasing* σ*-fields, and* $\sup_i \mathsf{E}[|\epsilon_i|^\alpha|\mathscr{F}_{i-1}] < \infty$, *a.s., for some* $\alpha > 2$.

Assumption 4 (A4) $\|\boldsymbol{\beta}_k\|_q \leq d_{k,q},\ \forall k \in [K],\ \forall q \in \{0, 1, 2\}$.

The noise assumptions imposed above are relatively mild. For instance, (A2) holds if $\{\epsilon_i\}$ are *i.i.d.* random variables with zero mean and variance, including but not limited to Gaussian white noise.

The following lemmas are important in proving the strong consistency of least squares estimates in stochastic linear regression models.

Lemma 3 *(Lai et al., 1982) Suppose the noise sequence* $\{\epsilon_i\}$ *satisfies Assumption (A2). Let* $\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$*-measurable for every* i. *Define* $N = \inf\{n : \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' \text{ is nonsingular}\}$. *Assume that* $N < \infty$ *a.s., and for* $n \geq N$, *define*

$$Q_n = \left(\sum_{i=1}^n \mathbf{x}_i\epsilon_i\right)'\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i\epsilon_i\right).$$

Let $\lambda_{max}(n)$ *the maximum eigenvalue of* $\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$. *Then* $\lambda_{max}(n)$ *is non-decreasing in* n *and*

(i) On $(\lim_{n\to\infty} \log \lambda_{max}(n) < \infty)$, $Q_n = O(1)$ *a.s.*

(ii) On $(\lim_{n\to\infty} \log \lambda_{max}(n) = \infty)$, *we have that for* $\forall \delta > 0$,

$$Q_n = O((\log \lambda_{max}(n))^{1+\delta}) \text{ a.s.}$$

(iii) On $(\lim_{n\to\infty} \log \lambda_{max}(n) = \infty)$, *the previous result can be strengthened to*

$$Q_n = O(\log \lambda_{max}(n)) \text{ a.s.},$$

if the Assumption (A2) is replaced by (A3).

The following estimates for the weighted sums of martingale difference sequences are based on the Kronecker lemma, the local convergence theorem and the strong law for martingales (Chow and Teicher, 2012; Chow, 1965).

Lemma 4 *(Lai et al., 1982; Chen and Guo, 2012) Suppose the noise sequence* $\{\epsilon_i\}$ *satisfies the Assumption (A2). Let* u_i *be* $\mathscr{F}_{i-1}$*-measurable for every* i *and* $s_n = (\sum_{j\in[n]} u_j^2)^{\frac{1}{2}}$*. Then,*

(i) $\sum_{j\in[n]} u_j\epsilon_j$ *converges a.s. on* $s_n^2 < \infty$.

(ii) $\sum_{j\in[n]} u_j\epsilon_j = o(s_n[log(s_n^2)]^{\frac{1}{2}+\delta})$ *a.s.* $\forall\delta > 0$. *on* $\sum_{j\in[n]} u_j^2 = \infty$.

(iii) $\sum_{j\in[n]} u_j\epsilon_j = O(s_n[log(s_n^2)]^{\frac{1}{2}})$ *a.s. on* $\sum_{j\in[n]} u_j^2 = \infty$, *if the Assumption (A2) is replaced by (A3).*

The estimates given by these results are not as sharp as those given by the law of the iterative logarithm (Chow and Teicher, 2012) but the assumptions needed here are much more general.

5.3 Main Results

5.3.1 Noiseless Case

In this subsection we explore the strong consistency of MLR in the noiseless case. Suppose $\{\boldsymbol{\beta}_k^n,\ \forall k \in [K]\}$ are optimal solutions to Formulation (5.1) for some $p \in \{1, 2\}$ and $q \in \{0, 1, 2\}$.

Theorem 1 *Suppose Assumptions (A1) and (A4) hold in the MLR model, and*

$$\liminf_{|\mathcal{S}|\to\infty} \lambda_{min}\left(\sum_{i\in\mathcal{S}\subset\Omega_k^n} \mathbf{x}_i\mathbf{x}_i'\right) > 0,\ \ a.s.,\ \forall k \in [K]. \tag{5.4}$$

Then we have the strong consistency, i.e.,

$$\exists N \text{ such that } \boldsymbol{\beta}_k^n = \boldsymbol{\beta}_{\pi(k)},\ \ a.s., \forall n > N, k \in [K],$$

where π is a permutation of the K clusters.

Furthermore, if the optimal solution of Formulation (5.1) is unique, or there are no ties for assigning samples to clusters, i.e., $\forall i \in [n], \forall k \neq h \in [K]$, $|\mathbf{x}_i'(\beta_k - \beta_h)| > 0$, it follows

$$\exists N \text{ such that } \hat{\Omega}_k^n = \Omega_{\pi(k)}^n, \text{ a.s., } \forall n > N. \tag{5.5}$$

Proof For all $k \in [K]$, there exists $m(k) \in [K]$ such that

$$|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq |\Omega_k^n \cap \hat{\Omega}_h^n|, \ \forall h \in [K].$$

Consequently, summing the above over all $h \in [K]$, we have

$$|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq \sum_{\forall h \in [K]} |\Omega_k^n \cap \hat{\Omega}_h^n|/K = |\Omega_k^n|/K.$$

From Assumption (A1), $|\Omega_k^n|/K = \Theta(n) \to \infty$, when $n \to \infty$. Further, we have $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \to \infty$, when $n \to \infty$, i.e., for any given N_0, we have $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| > N_0$ if n is large enough. From the identifiability condition (5.4), there exists N such that $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| > N$,

$$\lambda_{min}\Big(\sum_{i \in \Omega_k^n \cap \hat{\Omega}_{m(k)}^n} \mathbf{x}_i \mathbf{x}_i' \Big) > 0, \text{ a.s.} \tag{5.6}$$

Since the true $\{\boldsymbol{\beta}_k\}$ are a feasible solution of Formulation (5.1) and there is no noise, the optimal objective value must be 0, i.e., $y_i = \mathbf{x}_i'\boldsymbol{\beta}_k$ for some k and all i. It follows

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_k = \mathbf{x}_i'\boldsymbol{\beta}_{m(k)}^n, \quad \forall i \in \Omega_k^n \cap \hat{\Omega}_{m(k)}^n.$$

As a result,

$$\mathbf{x}_i'(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{m(k)}^n) = 0,$$

and

$$\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} |\mathbf{x}_i'(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{m(k)}^n)|^2 = 0. \tag{5.7}$$

From the definition of the smallest eigenvalue, we have

$$\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} |\mathbf{x}_i'(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{m(k)}^n)|^2 \geq \lambda_{min}\Big(\sum_{i\in\Omega_k^n\cap\hat{\Omega}_{m(k)}^n} \mathbf{x}_i\mathbf{x}_i'\Big)\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{m(k)}^n\|^2. \tag{5.8}$$

Accordingly, when $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| > N$, it follows from equations (5.6), (5.7) and (5.8) that $\boldsymbol{\beta}_k - \boldsymbol{\beta}_{m(k)}^n = 0$.

Next, we show the mapping $m(\cdot)$ is a bijection by contradiction. Assume there exists $m(k) = m(h)$ for $k \neq h$. Then, $\boldsymbol{\beta}_k = \boldsymbol{\beta}_{m(k)}^n = \boldsymbol{\beta}_{m(h)}^n = \boldsymbol{\beta}_h$ when n is large enough, which contradicts the assumption that the clusters are different. Thus, $\pi = m^{-1}$ is a permutation. Finally, we can prove the cluster set equality (5.5) by contradiction. ■

Remark 1 *1. The identifiability condition, i.e., Equation (5.4) on $\mathbf{x}_i$ is not only sufficient but also necessary. $\mathcal{S}$ in (5.4) is the subset of Ω_k^n. For instance, if $\mathbf{x}_i = (1,1)'$, $\forall i \in \Omega_k^n$, and $\lambda_{min}(\sum_{i\in\Omega_k^n} \mathbf{x}_i\mathbf{x}_i') = 0$, which violates Equation (5.4), the model parameters are not identifiable no matter which method is being used.*

2. Thm. 1 holds for either $p = 1$ or $p = 2$. From a computational complexity aspect, a linear MIP ($p = 1, q = 1$) can be solved faster than a quadratic MIP ($p = 2$).

3. Even if we have the strong consistency for the model parameters, i.e., $\boldsymbol{\beta}_k^n \to \boldsymbol{\beta}_{\pi(k)}$, a.s.,$\forall k \in [K]$, we may not have $\hat{\Omega}_k^n \to \Omega_{\pi(k)}^n$, a.s., when $n \to \infty$. For instance, if $\mathbf{x}_i = \mathbf{0}$, for some i, then the sample i may be assigned to any cluster.

Thm. 1 holds for any $p > 0$ and $q \geq 0$. Assumption (5.4) is equivalent to requiring that every cluster has at least d linearly independent measurements. The exact convergence can also be achieved by other methods, e.g., the second-order cone program in (Hand and Joshi, 2018). However, to the best of our knowledge, our assump-

tions are the weakest since we do not need a "sufficient-separation" and a balanced measurement assumption on the data as in (Hand and Joshi, 2018).

5.3.2 Noisy Case with a Single Cluster

In this subsection we explore the strong consistency of linear regularized regression subject to martingale difference noise. Linear regularized regression can be seen as a specific case of MLR with a single cluster. In this section, we assume $K = 1$ and consider the presence of noise. We establish strong consistency for the model parameters under general covariate and noise distributions. Consider the regularized regression problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i\in[n]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|^2 \\ &\text{s.t. } \|\boldsymbol{\beta}\|_q \leq d_q, \quad \forall k \in [K], \end{aligned} \tag{5.9}$$

where $q \in \{0, 1, 2\}$, corresponding to ridge regression, LASSO and regression with a subset selection constraint, respectively. For $q = 0$, the problem can be formulated as a MIP problem (Bertsimas et al., 2016) as follows.

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i\in[n]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|^2 \\ &\text{s.t. } -c_j M \leq \beta_j \leq c_j M, \quad \forall j \in [d], \\ &\quad c_j \in \{0, 1\}, \quad \forall j \in [d], \\ &\quad \sum_{j\in[d]} c_j \leq d_0. \end{aligned}$$

Let $\boldsymbol{\beta}^n$ denote an optimal solution of (5.9). Let $\lambda_{max}(n) = \lambda_{max}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')$ and $\lambda_{min}(n) = \lambda_{min}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i')$.

Theorem 2 *Suppose that Assumptions (A1), (A2) and (A4) hold for (5.9), and let*

$\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$*-measurable for every* i*. If*

$$\lambda_{min}(n) \to \infty, \ \textit{a.s.},$$

then for all $\delta > 0$*,*

$$\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| \leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}(n)), \ \textit{a.s.}$$

Proof Since $\boldsymbol{\beta}$ is a feasible solution of (5.9) (cf. Ass. (A4)), we have

$$\sum_{i\in[n]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}^n|^2 \leq \sum_{i\in[n]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|^2 = \sum_{i\in[n]} |\epsilon_i|^2.$$

Substituting $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$ in the l.h.s. of the above, we obtain

$$\sum_{i\in[n]} |\epsilon_i - \mathbf{x}_i'(\boldsymbol{\beta}^n - \boldsymbol{\beta})|^2 \leq \sum_{i\in[n]} |\epsilon_i|^2,$$

and by expanding the l.h.s. of the above we obtain

$$\sum_{i\in[n]} |\mathbf{x}_i'(\boldsymbol{\beta}^n - \boldsymbol{\beta})|^2 \leq 2\sum_{i\in[n]} \epsilon_i\mathbf{x}_i'(\boldsymbol{\beta}^n - \boldsymbol{\beta}). \tag{5.10}$$

From the definition of the minimum eigenvalue, we have

$$\sum_{i\in[n]} |\mathbf{x}_i'(\boldsymbol{\beta}^n - \boldsymbol{\beta})|^2 \geq \lambda_{min}(n)\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|^2. \tag{5.11}$$

Next, we study the r.h.s. of (5.10) in order to bound the convergence rate of $\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|$. From Lemma 4(ii), we have that for any $j \in [d]$,

$$\sum_{i\in[n]} \epsilon_i x_{ij} = o(s_{nj}[log(s_{nj}^2)]^{\frac{1}{2}+\delta}) \text{ a.s., where } s_{nj} = \left(\sum_{i\in[n]} x_{ij}^2\right)^{\frac{1}{2}}.$$

For any $j \in [d]$,

$$s_{nj} = \Big(\sum_{i\in[n]} x_{ij}^2 \Big)^{\frac{1}{2}} \leq \Big(\sum_{i\in[n]} \sum_{j\in[d]} x_{ij}^2 \Big)^{\frac{1}{2}}$$
$$\leq \bigg(\mathrm{Tr}\Big(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \Big) \bigg)^{\frac{1}{2}} = \Theta\Big(\lambda_{max}^{\frac{1}{2}}(n) \Big),$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. By combining the above two equations, we have for any $j \in [d]$,

$$\sum_{i\in[n]} \epsilon_i x_{ij} = o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \text{ a.s.},$$

and thus,

$$\left\| \sum_{i\in[n]} \epsilon_i \mathbf{x}_i \right\| = o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \text{ a.s.}$$

We bound the r.h.s. of (5.10) as

$$2 \sum_{i\in[n]} \epsilon_i \mathbf{x}_i'(\boldsymbol{\beta}^n - \boldsymbol{\beta}) \leq 2 \left\| \sum_{i\in[n]} \epsilon_i \mathbf{x}_i \right\| \|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|$$
$$\leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|. \tag{5.12}$$

Combining (5.10), (5.11) and (5.12), we obtain

$$\lambda_{min}(n) \|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|^2 \leq o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}) \|\boldsymbol{\beta}^n - \boldsymbol{\beta}\|.$$

■

If the Assumption (A2) is replaced by (A3), we have a stronger version of the previous theorem (proof omitted).

Theorem 3 *Suppose that Assumptions (A1), (A3) and (A4) hold for (5.9), and let*

$\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$*-measurable for every* i*. If*

$$\lambda_{min}(n) \to \infty, \ \textit{a.s.},$$

then we have

$$\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| \leq O(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}}/\lambda_{min}(n)) \ \textit{a.s.}$$

As a point of comparison, the classical convergence rate for least square estimates of unregularized linear regression subject to martingale difference noise is given by:

$$\begin{aligned} \|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| &= o([log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}^{\frac{1}{2}}(n)) \text{ a.s. for } \alpha = 2, \\ &= O([log\lambda_{max}(n)]^{\frac{1}{2}}/\lambda_{min}^{\frac{1}{2}}(n)) \text{ a.s. for } \alpha > 2. \end{aligned} \tag{5.13}$$

The convergence rate in Theorem 2 and Theorem 3 is slower than the unregularized linear regression case in general. However, it can be seen that for well-conditioned data, i.e., $\lambda_{max}(n) \sim \lambda_{min}(n)$, we have the same convergence rate with the unregularized case. The following Corollary outlines sufficient conditions to that effect.

Corollary 1 *Suppose* $\{\mathbf{x}_i\}$ *are i.i.d. random vectors with* $\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']$ *being a positive definite matrix. Suppose* $\{\epsilon_i\}$ *are i.i.d. random variables, independent of the* $\mathbf{x}_i$*, with zero mean and variance* $\sigma^2 > 0$*. Then, we have*

$$\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| = o(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}+\delta}), \ \textit{a.s.} \ \forall \delta > 0.$$

Furthermore, if $\mathsf{E}[|\epsilon_i|^\alpha] < \infty, \ \textit{a.s. for some } \alpha > 2$*, we have*

$$\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| = O(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}}), \ \textit{a.s.} \tag{5.14}$$

Proof From the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}(\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i') = \mathsf{E}[\mathbf{x}_i\mathbf{x}_i'], \quad \text{a.s.}$$

It follows

$$\lambda_{max}(n) = \Theta(n), \qquad \lambda_{min}(n) = \Theta(n).$$

Thus, from Thm. 2 we have

$$\begin{aligned}\|\boldsymbol{\beta}^n - \boldsymbol{\beta}\| \leq & o(\lambda_{max}^{\frac{1}{2}}(n)[log\lambda_{max}(n)]^{\frac{1}{2}+\delta}/\lambda_{min}(n)) \\ = & o(n^{-\frac{1}{2}}[log(n)]^{\frac{1}{2}+\delta}), \text{ a.s. } \forall \delta > 0.\end{aligned}$$

Similarly, using Thm. 3 we can prove (5.14). ■

We point out that our setting is more general, including ridge regression, LASSO and regression with subset selection. The convergence rate is sharper in terms of the sample size than the rate achieved for LASSO in (Chatterjee, 2013). In addition, our noise assumptions are more general and weaker than the Gaussian noise used in (Chatterjee, 2013).

The next theorem considers the residual sum of squares as the consistent estimator of the noise variance under the assumption of asymptotic homoscedasticity.

Theorem 4 *Suppose that (A1), (A3) and (A4) hold in the model (5.9), and let* $\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$*-measurable for every* i*. Suppose* $\lim_{n\to\infty} \mathsf{E}[|\epsilon_i|^2|\mathscr{F}_{i-1}] = \sigma^2$*, a.s. for some constant* $\sigma > 0$*. Assume that* $N = \inf\{n \geq d : \sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^{'}$ *is of rank* $d\} < \infty$ *a.s. If*

$$\log \lambda_{max}(n) = o(n), \tag{5.15}$$

then

$$\hat{\sigma^2}(n) \triangleq \frac{1}{n}\sum_{i\in[n]} |y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}^n|^2 \to \sigma^2 \text{ a.s.} \tag{5.16}$$

Proof It follows from (A3) that (Lai et al., 1982)

$$\sum_{i\in[n]}(\epsilon_i^2 - \mathsf{E}[\epsilon_i^2|\mathscr{F}_{i-1}]) = o(n), \text{ a.s.}$$

Consequently

$$\sum_{i\in[n]}|\epsilon_i|^2 = \sum_{i\in[n]}\mathsf{E}[\epsilon_i^2|\mathscr{F}_{i-1}] + o(n) \sim n\sigma^2, \text{ a.s.} \tag{5.17}$$

First, since $\boldsymbol{\beta}$ must be feasible solution for the constrained optimization problem, (A4) implies,

$$\sum_{i\in[n]}|y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}^n|^2 \leq \sum_{i\in[n]}|y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}|^2 = \sum_{i\in[n]}|\epsilon_i|^2. \tag{5.18}$$

Second, $\boldsymbol{\beta}^n$ is the optimal solution for the constrained optimization problem (5.9) and also a feasible solution for the unconstrained linear regression problem $\sum_{i\in[n]}|y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}_{LS}^n|^2$. Assume that $\boldsymbol{\beta}_{LS}^n$ is the optimal solution for unconstrained linear regression problem. Then,

$$\sum_{i\in[n]}|y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}^n|^2 \geq \sum_{i\in[n]}|y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}_{LS}^n|^2. \tag{5.19}$$

By using the closed form solution for least squares estimates as follows,

$$\begin{aligned}\boldsymbol{\beta}_{LS}^n &= (\sum_{j\in[n]}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in[n]}\mathbf{x}_j y_j,\\ &= \boldsymbol{\beta} + (\sum_{j\in[n]}\mathbf{x}_j\mathbf{x}_j^{'})^{-1}\sum_{j\in[n]}\mathbf{x}_j\epsilon_j,\end{aligned}$$

we have

$$\begin{aligned}
&\sum_{i\in[n]} |y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}_{LS}^n|^2 \\
=&\sum_{i\in[n]} |\epsilon_i|^2 - (\sum_{i=1}^n \mathbf{x}_i\epsilon_i)^{'}(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{'})^{-1}(\sum_{i=1}^n \mathbf{x}_i\epsilon_i) \\
=&\sum_{i\in[n]} |\epsilon_i|^2 - O(\log \lambda_{max}(n)).
\end{aligned}$$

From the Assumption (5.15) on the maximum eigenvalue and the above equation, we have

$$\frac{1}{n}\sum_{i\in[n]} |y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}_{LS}^n|^2 = \frac{1}{n}\sum_{i\in[n]} |\epsilon_i|^2 - o(1), \text{ a.s.} \tag{5.20}$$

By combing the above equations (5.17), (5.18), (5.19) and (5.20), we have

$$\frac{1}{n}\sum_{i\in[n]} |y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}^n|^2 \to \sigma^2 \text{ a.s.}$$

■

Based on the last theorem, we have the convergence rate for the noise variance estimate under common assumptions used in the machine learning literature.

Corollary 2 *Suppose* $\{\mathbf{x}_i\}$ *are i.i.d. random variables with* $\mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{'}]$ *is a positive definite matrix. Suppose* $\{\epsilon_i\}$ *are i.i.d. random variables with zero mean and* $\mathsf{E}|\epsilon_i|^\alpha < \infty$, *a.s. for some* $\alpha > 2$. *The two families of random variables are independent. Thus we have*

$$|\hat{\sigma^2}(n) - \sigma^2| = O(n^{-1}[\log(n)]), \text{ a.s.},$$

where $\hat{\sigma^2}(n)$ *has been defined in (5.16).*

Proof From the strong law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n}(\sum_{i\in[n]} \mathbf{x}_i\mathbf{x}_i^{'}) = \mathsf{E}[\mathbf{x}_i\mathbf{x}_i^{'}],\ \text{a.s.}$$

Thus

$$\lambda_{max}(n) = \Theta(n),\ \text{a.s.}.$$

From Theorem 4, we have

$$\begin{aligned} &|\hat{\sigma^2}(n) - \sigma^2| \\ =&O(n^{-1}\log\lambda_{max}(n)) \\ =&O(n^{-1}[\log(n)]),\ \text{a.s.} \end{aligned}$$

■

In summary, strong consistency is established for the model parameters of linear regression with different penalty constraints under the martingale difference noise case. If the K clusters are separated far away from each other, the strong consistency can be solved by K separate models and the techniques here for the linear regression case.

5.3.3 Noisy Case with Multiple Clusters

In this subsection, we derive the convergence conditions for MLR with multiple clusters under the martingale difference noise case. The theorem below coincides with the intuition that if the K clusters are well-separated, the convergence of the estimates becomes almost equivalent to having K separate linear regression models.

Following the analysis of the noiseless case in Theorem 1, $\forall k \in [K],\ \exists m(k) \in [K]$

such that $|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq |\Omega_k^n \cap \hat{\Omega}_h^n|,\ \forall h \in [K]$. From Assumption (A1),

$$|\Omega_k^n \cap \hat{\Omega}_{m(k)}^n| \geq |\Omega_k^n|/K = \Theta(n) \to \infty, \text{ when } n \to \infty.$$

Theorem 5 *Suppose that Assumptions (A1), and either (A2) or (A3) hold for Formulation (5.1), let* $\mathbf{x}_i$ *be* $\mathscr{F}_{i-1}$*-measurable for every* i*. Further assume the stronger identifiability condition* $\lambda_{max}(\sum_{i\in\mathcal{S}\subset\Omega_k^n} \mathbf{x}_i\mathbf{x}_i') = \Theta(|\mathcal{S}|)$*, and* $\lambda_{min}(\sum_{i\in\mathcal{S}\subset\Omega_k^n} \mathbf{x}_i\mathbf{x}_i') = \Theta(|\mathcal{S}|)$*,* $\forall k \in [K]$*. Also assume the cluster consistency condition* $|\Omega_k^n \setminus \hat{\Omega}_{m(k)}^n| = o(n)$*,* $|\hat{\Omega}_{m(k)}^n \setminus \Omega_k^n| = o(n), \forall k \in [K]$*. Then, we have the strong consistency for the MLR parameters estimates from Formulation (5.1) when* $p = 2$ *and in the absence of the last regularization constraint.*

Proof From Lemma 2,

$$\begin{aligned}\boldsymbol{\beta}_{m(k)}^n &= \Big(\sum_{j\in\hat{\Omega}_{m(k)}^n} \mathbf{x}_j\mathbf{x}_j^{'}\Big)^{-1} \sum_{j\in\hat{\Omega}_{m(k)}^n} \mathbf{x}_j y_j, \quad \forall k \in [K].\\ &= \Big(\sum_{j\in\Omega_k^n} \mathbf{x}_j\mathbf{x}_j^{'} - \sum_{j\in\Omega_k^n\setminus\hat{\Omega}_{m(k)}^n} \mathbf{x}_j\mathbf{x}_j^{'} + \sum_{j\in\hat{\Omega}_{m(k)}^n\setminus\Omega_k^n} \mathbf{x}_j\mathbf{x}_j^{'}\Big)^{-1}\\ &\qquad\qquad \Big(\sum_{j\in\Omega_k^n} \mathbf{x}_j y_j - \sum_{j\in\Omega_k^n\setminus\hat{\Omega}_{m(k)}^n} \mathbf{x}_j y_j + \sum_{j\in\hat{\Omega}_{m(k)}^n\setminus\Omega_k^n} \mathbf{x}_j y_j\Big).\end{aligned}$$

From Assumptions (A1), the stronger identifiability condition and the cluster consistency condition, we have

$$\Big(\sum_{j\in\Omega_k^n} \mathbf{x}_j\mathbf{x}_j^{'} - \sum_{j\in\Omega_k^n\setminus\hat{\Omega}_{m(k)}^n} \mathbf{x}_j\mathbf{x}_j^{'} + \sum_{j\in\hat{\Omega}_{m(k)}^n\setminus\Omega_k^n} \mathbf{x}_j\mathbf{x}_j^{'}\Big) = |\Omega_k^n|\Big(\frac{1}{|\Omega_k^n|}\sum_{j\in\Omega_k^n} \mathbf{x}_j\mathbf{x}_j^{'} + o(1)\Big).$$

From Assumptions (A1), the stronger identifiability condition, either (A2) or (A3), Lemma 4 and the cluster consistency condition, we have

$$\Big(\sum_{j\in\Omega_k^n} \mathbf{x}_j y_j - \sum_{j\in\Omega_k^n\setminus\hat{\Omega}_{m(k)}^n} \mathbf{x}_j y_j + \sum_{j\in\hat{\Omega}_{m(k)}^n\setminus\Omega_k^n} \mathbf{x}_j y_j\Big) = |\Omega_k^n|\Big(\frac{1}{|\Omega_k^n|}\sum_{j\in\Omega_k^n} \mathbf{x}_j y_j + o(1)\Big).$$

Consequently,

$$\begin{aligned}
\boldsymbol{\beta}^n_{m(k)} &= \left(\frac{1}{|\Omega^n_k|}\sum_{j\in\Omega^n_k}\mathbf{x}_j\mathbf{x}_j^{'}+o(1)\right)^{-1}\left(\frac{1}{|\Omega^n_k|}\sum_{j\in\Omega^n_k}\mathbf{x}_j y_j+o(1)\right)\\
&= \left(\frac{1}{|\Omega^n_k|}\sum_{j\in\Omega^n_k}\mathbf{x}_j\mathbf{x}_j^{'}+o(1)\right)^{-1}\left(\frac{1}{|\Omega^n_k|}\sum_{j\in\Omega^n_k}\mathbf{x}_j(\mathbf{x}_j'\boldsymbol{\beta}_k+\epsilon_j)+o(1)\right)\\
&= \boldsymbol{\beta}_k+\left(\sum_{j\in\Omega^n_k}\mathbf{x}_j\mathbf{x}_j^{'}\right)^{-1}\sum_{j\in\Omega^n_k}\mathbf{x}_j\epsilon_j+o(1)\\
&= \begin{cases}\boldsymbol{\beta}_k+o(n^{-\frac{1}{2}}[\log(n)]^{\frac{1}{2}+\delta}),\ \text{a.s. } \forall\delta>0.\ \text{if } \alpha=2\\ \boldsymbol{\beta}_k+O(n^{-\frac{1}{2}}[\log(n)]^{\frac{1}{2}}),\ \text{a.s. if } \alpha>2.\end{cases}
\end{aligned}$$

■

If $\{\mathbf{x}_i\}$ are i.i.d. random vectors with $\mathsf{E}[\mathbf{x}_i\mathbf{x}_i']$ being a positive definite matrix, stronger identifiability condition will hold. From this theorem, even if the clusters converge with $o(n)$ inconsistency, we still have the strong consistency for the MLR parameters estimates. In the next subsection, we will see the cluster inconsistency bound is tight, i.e., the strong consistency may fail if the clusters diverge with $\Theta(n)$ inconsistency.

5.3.4 A Counterexample for the Noisy Case with Two Clusters

In this subsection, we provide a counterexample indicating that the presence of noise may prevent convergence to the true coefficients when we have more than a single cluster under the martingale difference noise case.

Consider the 2-cluster MLR model where the data $(x_i, y_i) \in \mathbb{R}^2$ are generated as follows:

$$\begin{aligned} x_i =& 1, \forall i \in [n], \\ \epsilon_i \in& \{1, -1\} \text{ with probability } 0.5, i \in [n], \\ \beta_k \in& \{\delta, 0\} \text{ with probability } 0.5, k \in [2], \\ y_i =& x_i \beta_k + \sigma \epsilon_i \in \{\sigma, -\sigma, \delta + \sigma, \delta - \sigma\}, i \in [n], \text{ for some } k \in [2]. \end{aligned}$$

We next describe the data generation process. Each training sample has a probability 0.5 of being drawn from the Cluster 1 distribution, and a probability 0.5 of being drawn from the Cluster 2 distribution. Given the true regression coefficients $\beta_k, k \in [2]$, we generate the training data as follows: $y_i = x_i \beta_k + \sigma \epsilon_i$, where σ is a constant and $\{\epsilon_i\}$ are i.i.d. random variables with zero mean and positive variance.

If $\delta < \sigma$, the optimal solution of Formulation (5.1) will be different from the ground truth parameters. In particular, the optimal objective function value of Formulation (5.1), is smaller than the objective function value of the ground truth parameters, and depend on δ rather than σ. In this example, the strong consistency fails since the clusters diverge with $\Theta(n)$ inconsistency.

Not only Formulation (5.1), but also other methods may fail to recover the ground truth parameters since they are "inherently unidentifiable", i.e., two sets of parameters can generate the same distribution. This counterexample shows strong consistency may fail to hold under the weakest assumptions used in our earlier analysis for the noiseless case and the noisy case with a single cluster.

5.4 Prediction Algorithm

In this section, we study how to apply MLR for large scale practical prediction problems. For simplicity of notation, we focus on MLR without the regularization con-

straint. Most of MLR studies in the machine learning area focus on the convergence of parameter estimates rather than how to use it in prediction applications. Recently (Gitman et al., 2018) reviewed and studied how to use regression models for prediction on the unseen test points. There are several challenges to that end.

First, it is computationally expensive to solve MLR in Formulation (5.1) for large scale prediction problems. We propose an Alternating Clustering and Regression (ACR) approach with MLR initialization to speed up computations. Second, MLR or ACR can not be used for prediction directly since the predicted target values depend on the clustering, which, in turn, depends on the $(y_i, \mathbf{x}_i)$. However, the target values (y_i) in the test data can not be known in advance. We tackle this challenge by training a random forest classifier or another multiclass classifier using $\mathbf{x}_i$ as the training input samples. Specifically, we learn to predict cluster labels l_i^n from MLR or ACR and use such a predictive method to predict cluster labels for the test data. Finally, we predict the target values using the MLR or ACR and predicted cluster labels.

Algorithm 3 MLR

Input: Training samples $\mathbf{x}_i$, training targets y_i, downsampling size n_down.
Diverse Downsampling: Apply clustering methods, e.g., k-means, with training samples to obtain n_down clusters. Select one sample from every cluster to obtain a subset of training samples.
Training: Solve the MIP in Formulation (5.1) with n_down selected training samples.
Output: Cluster labels l_i^n for samples in the training data, Regression coefficients $\boldsymbol{\beta}_k^n$.

When the cluster labels l_i^n are fixed or known, we just need to solve K separate regressions, i.e.,

$$\min_{\boldsymbol{\beta}_k} \frac{1}{n} \sum_{i \in [n], l_i^n = k} |y_i - \mathbf{x}_i' \boldsymbol{\beta}_k|^p, k \in [K]. \tag{5.21}$$

If $p = 2$, this amounts to a linear regression for every fixed cluster. If $p = 1$, the

problem is a mean absolute error (MAE) regression. As mentioned earlier, solving the MLR problem in Formulation (5.1) using all training data for large scale prediction problems. Using Algorithm 3, we can obtain a good feasible solution within a reasonable period of time. We apply an alternating optimization approach, Alternating Clustering and Regression (ACR) in Algorithm 4, to solve large scale problems within a given period of time. It has recently been shown that parameter estimates of ACR or alternating minimization converge to the true parameters if one starts from a small enough neighborhood around them (Yi et al., 2014; Yi and Caramanis, 2015; Balakrishnan et al., 2017). The alternating minimization approach has been widely used in other machine learning tasks, e.g., joint clustering and classification (Xu et al., 2016; Brisimi et al., 2018b; Brisimi et al., 2019). The idea behind ACR is to alternately train regression models and then re-cluster the samples, yielding an algorithm that scales well. The regression coefficients $\boldsymbol{\beta}_k^n$ from ACR are still feasible solutions of MLR but may not be the optimal solutions. The performance of ACR may be affected by the choice of the initial cluster. That is why we solve MLR with selected training samples to obtain a good initial solution.

Clearly, MLR or ACR can not be used for prediction directly since the clustering depends on the $(y_i, \mathbf{x}_i)$ and (y_i) in the test data, which are not known in advance. In order to apply MLR or ACR for prediction using real data, we must predict the hidden cluster labels l_i^n for samples in the test data. To that end, we use the classifcation approach in Algorithm 5. The performance of prediction may be affected by the choice of the multiclass classifier.

After we solve the MLR problems and predict the hidden cluster labels l_i^n for samples in the test data, we apply the MLR predictive model to each test data point (Algorithm 6).

Algorithm 4 ACR Training

Input: Training samples $\mathbf{x}_i$, Training targets y_i.

Initialization:

if Given a initial feasible solution $(\boldsymbol{\beta}_k^n)$ from Algorithm 3 or other methods **then**

assign sample i to cluster $l_i^n = k$ such that $|y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n| = \min_{m\in[K]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|$;

else

Randomly (or use clustering methods, e.g., k-means) assign sample i to cluster l_i^n, for $i \in [n]$ and $l_i^n \in [K]$.

end if

repeat

Regression Step:

Train a linear regression or mean absolute error (MAE) regression (5.21) for each cluster of samples. Each regression is a linear or quadratic optimization problem and provides a hyperplane perpendicular to $\boldsymbol{\beta}_k^n$ and a corresponding optimal objective value O^k.

Re-clustering Step:

Re-cluster the samples based on the regressions, i.e., assign sample i to cluster $l_i^n = k$ such that $|y_i - \mathbf{x}_i'\boldsymbol{\beta}_k^n| = \min_{m\in[K]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_m^n|$.

until no l_i^n is changed or $\sum_k O^k$ is not decreasing.

Output: Cluster labels l_i^n for samples in the training data, Regression coefficients $\boldsymbol{\beta}_k^n$.

Algorithm 5 MLR Classification

Input: Training and test samples $\mathbf{x}_i$, Training cluster labels l_i^n, i.e., cluster assignment which assigns sample i to cluster l_i^n from Algorithm 4.
Training: Train a random forest classifier or another multiclass classifier using $\mathbf{x}_i$ as the training input samples and l_i^n as the target values. Tune the hyperparameters with K-Folds cross-validation and grid search techniques.
Test: Predict the hidden cluster labels l_i^n for samples in the test data using the multiclass classifier trained above.
Output: Predicted cluster labels l_i^n for samples in the test data.

Algorithm 6 MLR Prediction

Input: Test samples $\mathbf{x}_i$, Regression coefficients $\boldsymbol{\beta}_k^n$ from Algorithm 4, Predicted test cluster labels l_i^n from Algorithm 5.
Output: Predicted targets $\mathbf{x}_i^{'}\boldsymbol{\beta}_{l_i^n}^n$ for samples in the test data

5.5 Numerical Results on the Convergence

In this section we provide numerical results on the convergence of parameters estimates in MLR obtained by Formulation (5.1). Computations were performed with GUROBI 8.0 and python 3.6.5 using 16 CPUs on the Boston University Shared Computing Cluster. The results below coincide with the intuition that if the K clusters are well-separated, the convergence of the estimates becomes almost equivalent to having K separate linear regression models.

5.5.1 MLR under Gaussian Noise

In this subsection we explore the convergence performance of MLR under Gaussian noise using a synthetic data. Consider a model where the data $(\mathbf{x}_i, y_i) \in \mathbb{R}^3$ are generated independently by

$$
\begin{aligned}
x_{i,1} \sim & UNIFORM(0,1), \quad x_{i,2} = 1,\ i \in [n], \\
\boldsymbol{\beta}_1 = & (-0.93, 0.1), \quad \boldsymbol{\beta}_2 = (0,0), \\
\epsilon_i \sim & N(0,1),\ i \in [n], \\
y_i = & \mathbf{x}_i' \boldsymbol{\beta}_k + 0.01\epsilon_i,\ i \in [n],\ \text{for some } k \in [2].
\end{aligned}
$$

The first element of of $\mathbf{x}_i$ consists of random samples from the uniform distribution over $[0, 1)$. The second element of $\mathbf{x}_i$ is constant. Each training sample has a probability 0.5 of being drawn from the Cluster 1 distribution, and a probability 0.5 of being drawn from the Cluster 2 distribution. Given the true regression coefficients $\beta_k, k \in [2]$, we generate the training data as follows: $y_i = x_i\beta_k + \sigma\epsilon_i$, where σ is a constant 0.01 and $\{\epsilon_i\}$ are i.i.d. random variables with zero mean and positive variance from the standard normal distribution.

Figure 5·1 plots the evolution of $\|\boldsymbol{\beta}_k^n - \boldsymbol{\beta}_k\|$ as a function of the number of samples n used in solving Formulation (5.1), where we used $p = 1$, $M = 10$, did not include a regularization constraint, and set the time limit for each model in GUROBI to 2 hours.

5.5.2 MLR under Uniform Noise

In this subsection we explore the convergence performance of MLR under uniform noise using a synthetic data. Consider now a model where the data samples $(\mathbf{x}_i, y_i) \in \mathbb{R}^3$ are generated independently by

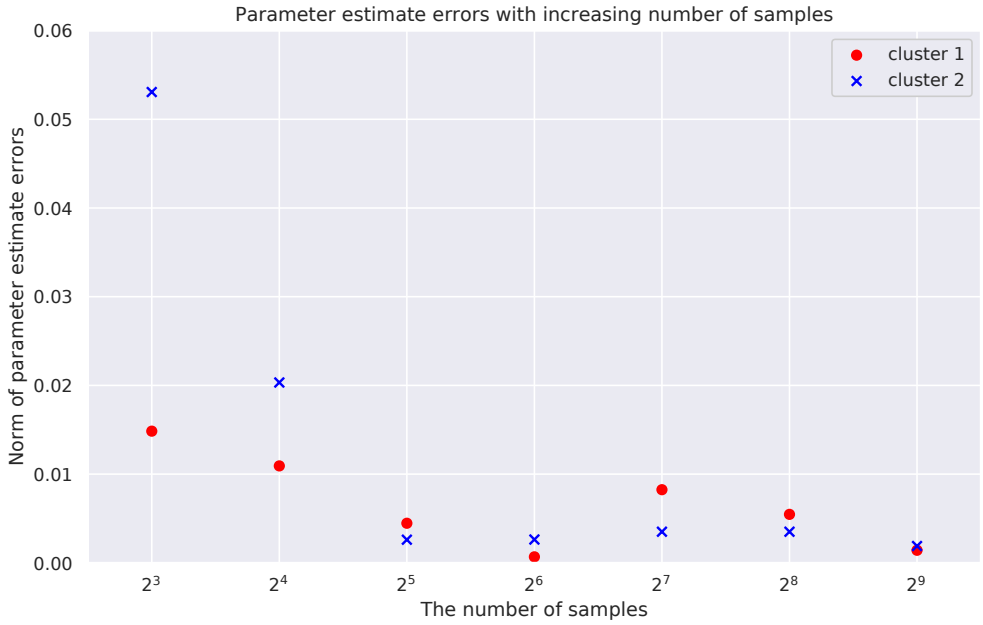

Figure 5·1: Gaussian noise case.

$$
\begin{aligned}
x_{i,1} \sim & UNIFORM(0,1), \quad x_{i,2} = 1, \ i \in [n], \\
\boldsymbol{\beta}_1 = & (-1.61, 1.25), \quad \boldsymbol{\beta}_2 = (0,0), \\
\epsilon_i \sim & UNIFORM(-1,1), \ i \in [n], \\
y_i = & \mathbf{x}_i' \boldsymbol{\beta}_k + 0.01\epsilon_i, \ i \in [n], \ \text{for some } k \in [2].
\end{aligned}
$$

The first element of $\mathbf{x}_i$ consists of random samples from the uniform distribution over $[0, 1)$, and the second element is the constant 1. Each training sample has a probability 0.5 of being drawn from the Cluster 1 distribution, and a probability 0.5 of being drawn from the Cluster 2 distribution. Given the true regression coefficients $\beta_k, k \in [2]$, we generate the training data as follows: $y_i = x_i\beta_k + \sigma\epsilon_i$ where σ is a constant 0.01 and $\{\epsilon_i\}$ are i.i.d. random variables with zero mean and positive variance from the uniform distribution over $[-1, 1)$.

Figure 5·2 plots $\|\boldsymbol{\beta}_k^n - \boldsymbol{\beta}_k\|$ as a function of the number of samples used in Formulation (5.1), where we used $p = 1$, $M = 10$, did not include a regularization constraint, and set the time limit for each model in GUROBI to 2 hours.

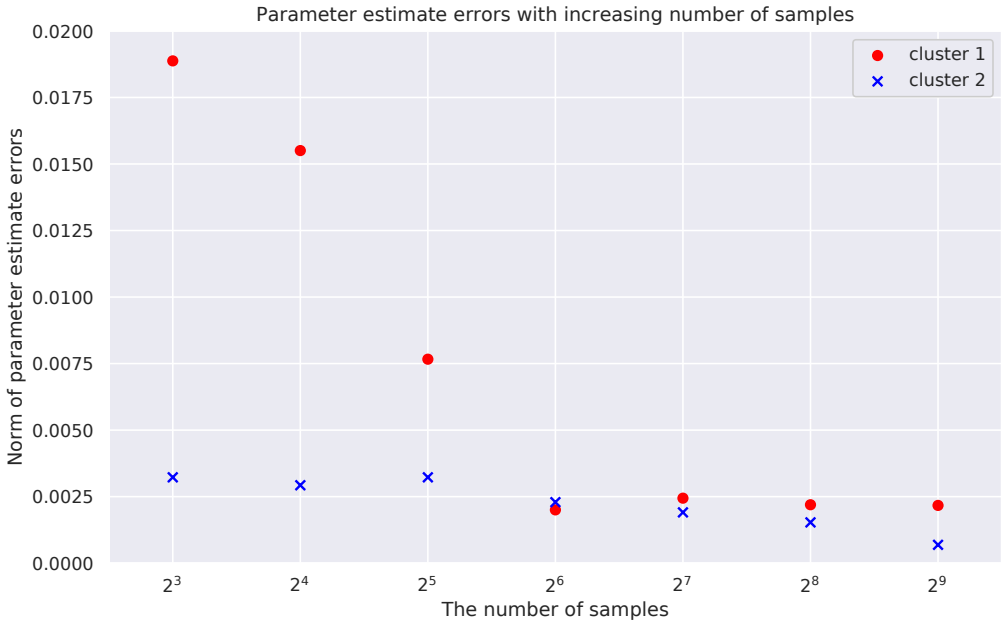

Figure 5·2: Uniform noise case.

5.6 Experimental Prediction Results

In this section, we present experimental prediction results with the prediction algorithm described in Section 5.4 on two datasets in terms of both accuracy and interpretability.

We split the datasets into 3 consecutive folds. Each fold is then used once as a test set while the 2 remaining folds form the training set. We report the average and standard deviation of mean absolute error regression loss (MAE), and R^2 (coefficient of determination) regression score (R2).

We compare different machine learning and statistical models on two datasets in

terms of both accuracy and interpretability. MAE_LP corresponds to MAE regression, i.e., $\min_{\boldsymbol{\beta}} \sum_{i\in[n]} |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$, which can be seen as MLR with only one cluster. MLR_K2 corresponds to MLR with the Random Forest classifier and $K = 2, p = 1$. MLR_K3 corresponds to MLR with the Random Forest classifier and $K = 3, p = 1$. RF corresponds to the Random Forest regression with 100 trees and hyper-parameter tuning of max_features and max_depth. Computations were performed with GUROBI 8.1.1 and python 3.7.2 on a machine with 32 Intel E5-2690 CPUs (2.90 GHz). We use the implementation of random forests included in the scikit-learn (Pedregosa et al., 2011) module.

Both our MLR method and state-of-the-art Random Forest regression perform much better than the linear MAE regression for the two datasets. However, our MLR method result is much easier to interpret than Random Forest (hundreds of different trees) and other black box models.

5.6.1 Synthetic Data

In this subsection we experiment with a synthetic dataset. Consider a model where the data $(x_i, y_i) \in \mathbb{R}^2,\ i \in [n],\ n = 512$ are generated independently by

$$
\begin{aligned}
x_i \sim & UNIFORM(-1, 1),\ i \in [n],\\
x_i = & x_i/(\sum_i x_i^2)^{0.5},\ i \in [n],\\
\epsilon_i \sim & N(0, 1),\ i \in [n]\\
y_i = & x_i + 0.01\epsilon_i,\ \text{if } x_i > 0,\ i \in [n]\\
y_i = & -x_i + 0.01\epsilon_i,\ \text{if } x_i \le 0,\ i \in [n].
\end{aligned}
$$

We next describe the data generation process. We draw the variable x_i from the uniform distribution and divide it by $(\sum_i x_i^2)^{0.5}$. Each training sample is drawn from the Cluster 1 distribution if $x_i > 0$, or from the Cluster 2 distribution if $x_i \le 0$. Given

the true regression coefficients $\beta_k, k \in [2]$, we generate the training data as follows: $y_i = x_i\beta_k + 0.01\epsilon_i$ where $\{\epsilon_i\}$ are i.i.d. random variables with zero mean and positive variance from the normal distribution. We report the average and standard deviation of mean absolute error regression loss (MAE), and R^2 (coefficient of determination) regression score (R2) for the test dataset in Table 5.1.

Table 5.1: Test MAE and R^2 for synthetic data.

Methods	avg. MAE	std MAE	avg. R2	std R2
MAE_LP	0.022	0.002	-0.036	0.009
MLR_K2	0.008	0.000	0.863	0.017
RF Regressor	0.009	0.000	0.840	0.018

5.6.2 Medical Cost Data

In this subsection we experiment with a public medical cost dataset. We predict individual medical costs billed by health insurance with age, sex (male or female), Body mass index (BMI), number of children, smoker (yes or no) and the residential region in the US (northeast, southeast, southwest, northwest). The dataset, including 1338 records, is available on GitHub (https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/insurance.csv). We convert categorical variables into dummy indicator variables, divide the target medical costs by 10000 and standardize features by subtracting the mean and scaling to unit variance before applying the methods. We report the average and standard deviation of mean absolute error regression loss (MAE), and R^2 (coefficient of determination) regression score (R2) for the test dataset in Table 5.2.

In Figure 5·3, we can see how the variable mean values differ in different clusters generated from our MLR with $K = 2$. The mean values here are after standardization and have both positive and negative values.

Table 5.2: Test MAE and R^2 for medical cost data.

Methods	avg. MAE	std MAE	avg. R2	std R2
MAE_LP	0.341	0.014	0.667	0.025
MLR_K2	0.183	0.010	0.851	0.011
MLR_K3	0.193	0.016	0.840	0.018
RF Regressor	0.188	0.007	0.854	0.009

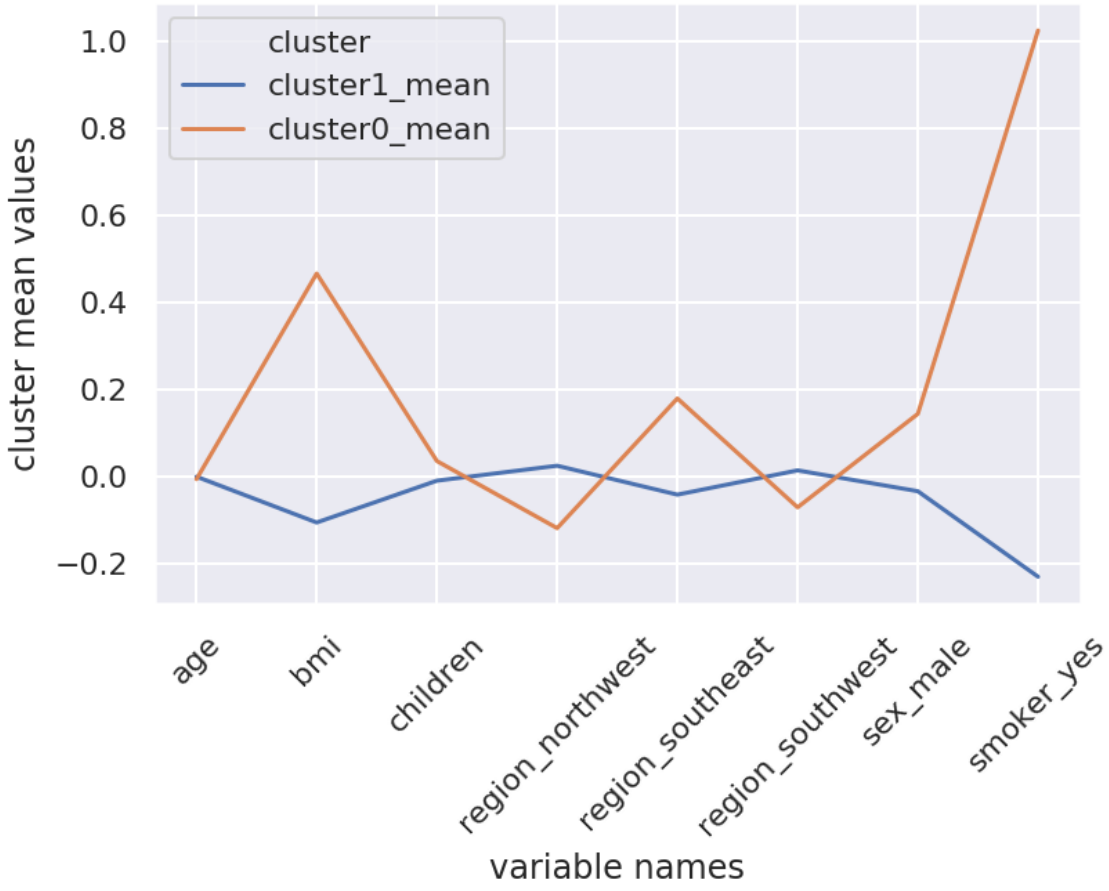

Figure 5·3: Cluster mean values for variables in every cluster.

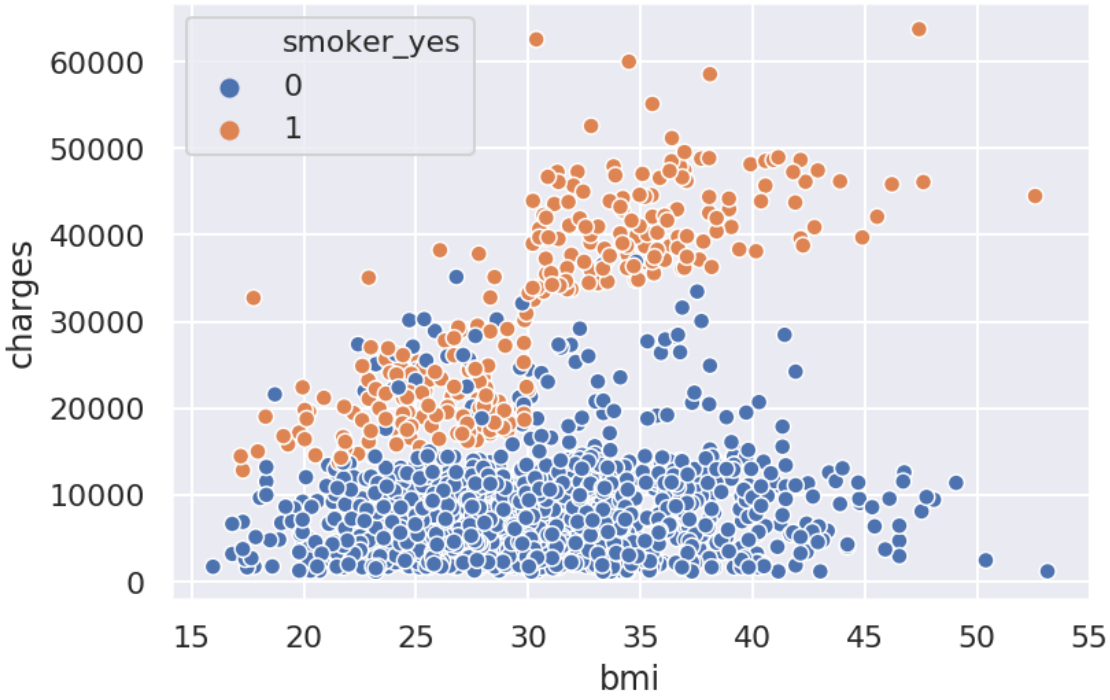

Figure 5·4: Scatter plot of costs and BMI grouped by smoking status.

In order to visualize the clusters in low dimension, we plot figures with medical costs and most different variables in different clusters, i.e., BMI and smoker. In Figure 5·4, we plot scatter figures of charges and BMI grouped by smoking status using the raw data. In Figure 5·5, we plot scatter figures of charges and BMI grouped by clusters in MLR. In Figure 5·4 and 5·5, we find that samples of cluster 0 mainly consist of smoker and higher BMI values, which is consistent with the fact that the mean values of smoker and BMI values are higher in cluster 0 in Figure 5·3.

5.7 Conclusions

We established the convergence of parameter estimates for regularized mixed linear regression models with multiple components in the noiseless case. Regularized linear regression can be seen as a specific case of MLR with a single cluster. We estab-

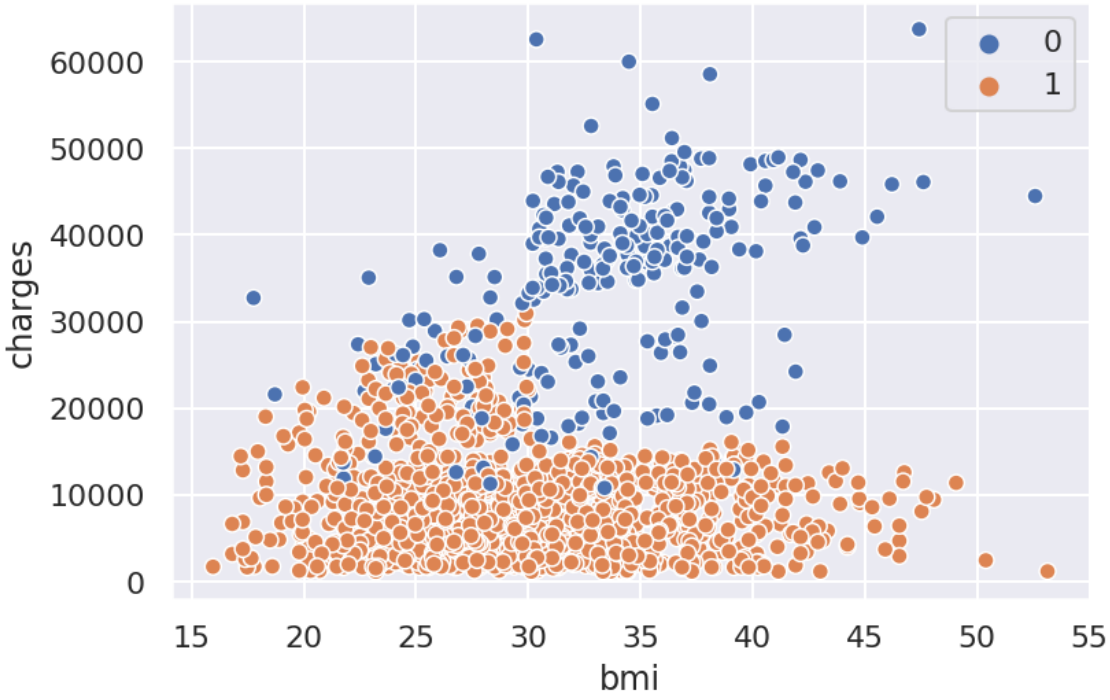

Figure 5·5: Scatter plot of costs and BMI grouped by clusters in MLR.

lish strong consistency and characterize the convergence rate of parameter estimates for such regression models subject to martingale difference noise under very weak assumptions on the data distribution.

To the best of our knowledge, our study is the first to study strong consistency of parameter estimates for mixed linear regression models under general noise conditions and general feature conditions rather than convergence with high probability. It can be used directly and applied to many areas, including but not limited to system identification and control, econometric theory and time series analysis. Besides the convergence results, our study proposes a novel method to apply MLR for practical large scale prediction problems. We present experimental prediction results and compare our prediction algorithm with mean absolute error regression and Random Forest regression in terms of both accuracy and interpretability.

Chapter 6

Conclusions

We summarize the dissertation in Sec. 6.1 and propose possible future directions in Sec. 6.2.

6.1 Summary

In this dissertation, we study and develop a variety of models and methods for predictive analytics, prescriptive analytics, designing metabolic division of labor in microbial communities, and parameter estimates for MLR.

First, we study designing metabolic division of labor in microbial communities. We develop a novel Mixed Integer Linear Program (MILP) based approach to optimally allocate the metabolic functions among organisms in a microbial ecosystem. We test the method on both a community composed of E. coli core models and a community composed of E. coli iJR904 models. In both cases, the method helps us identify the individual metabolic network topology and elucidate the interaction between species in the microbial community. It provides a new platform for the rational design of organisms and communities towards future synthetic ecology applications.

Second, we extend our scope from microbial ecosystems to the individual healthcare and study predictive analytics for 30-day hospital readmissions after general surgery discharge. We employ and develop supervised learning methods from the field of machine learning to learn efficiently and effectively from large datasets. Our methods are validated by using actual clinical data from Boston Medical Center.

Third, we study prescriptive analytics for 30-day hospital readmissions after general surgery discharge. We propose a new method, Prescriptive Support Vector Machines, and evaluate the prescriptive analytics results by different predictive machine learning methods. Our methods are validated using the National Surgical Quality Improvement Program (NSQIP) dataset.

Fourth, we introduce a general MIP formulation for MLR subject to norm-based regularization constraints. We study the consistency conditions of parameter estimates for MLR using an MIP formulation in both the noiseless case and noisy case. We propose an identifiability condition and establish that optimal solutions of the MIP converge almost surely (rather than w.h.p.) to the true parameters in the noiseless case as the sample size increases. To the best of our knowledge, our study is the first to study strong consistency of parameter estimates for MLR under general noise conditions and general feature conditions rather than convergence with high probability. Besides the convergence results, we propose a novel method to apply MLR for practical large scale prediction problems. We present experimental prediction results and compare our prediction algorithm with mean absolute error regression and Random Forest regression in terms of both accuracy and interpretability.

6.2 Future Works

A key motivation of this dissertation is that interpretable prediction models and causal inference are important in healthcare applications. In collaboration with the Boston University School of Public Health and the Boston Medical Center, we plan to develop prediction models for predicting female pregnancy using a dataset (PRESTO) obtained from surveys of female participants. We plan to use electronic health record data and develop a predictive model to predict the presence of an ovary condition, Polycystic ovary syndrome (PCOS), which is a leading cause of infertility. We will ap-

ply prediction models using machine learning, statistics, and optimization techniques for predicting female pregnancy using PRESTO data and predicting PCOS with an interpretable classification approach.

This thesis proposed several models based on Mixed Integer Program (MIP). For large scale problems, we proposed problem-specific heuristics and approximation algorithms and applied to problems using state of the art solvers, e.g., GUROBI. Relaxation techniques, e.g., Linear Program (LP) or Semidefinite Program (SDP) relaxation, can transform an NP-hard optimization problem, such as an MIP into a related problem that is solvable in polynomial time (e.g., LP/SDP). The solution to the relaxed problem can be used to gain information about the solution to the original integer program. One possible direction is in speeding up the solutions to the MIP problems considered in this dissertation.

Appendix A

Proof of equivalence of the models for SLSVM

For linear/quadratic programming models involving ℓ_1 norm or absolute values, we need to reformulate them in order to use optimization solvers, e.g., GUROBI. Two reformulation methods for $|\mathbf{w}_j|$ are as follows

- Introduce new variables z_j that satisfies $\mathbf{w}_j \leq z_j$ and $-\mathbf{w}_j \leq z_j$.
- Introduce new variables $\mathbf{w}_j^+, \mathbf{w}_j^-$, constrained to be non-negative and let $\mathbf{w}_j = \mathbf{w}_j^+ - \mathbf{w}_j^-$, and $|\mathbf{w}_j| = \mathbf{w}_j^+ + \mathbf{w}_j^-$.

For optimization solvers, e.g., GUROBI, the second reformulation method is faster than the first method due to fewer constraints. Next, we present the Quadratic Programming (QP) models for SLSVM, and prove the equivalence of all models. We use Model 3 with GUROBI in the applications involving SLSVM.

- Model 1: QP of SLSVM.

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \tfrac{1}{2}\mathbf{w}'\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i, i \in [n], \\ & \|\mathbf{w}\|_1 \leq K, \\ & \xi_i \geq 0, i \in [n]. \end{aligned}$$

- Model 2: (using $\mathbf{w}^+, \mathbf{w}^-$)

$$\min_{\mathbf{w}^+,\mathbf{w}^-,b,\xi} \quad \tfrac{1}{2}(\mathbf{w}^+ - \mathbf{w}^-)'(\mathbf{w}^+ - \mathbf{w}^-) + C\sum_{i=1}^{n}\xi_i \tag{A.1}$$

$$s.t. \quad y_i((\mathbf{w}^+ - \mathbf{w}^-)'\mathbf{x}_i + b) \geq 1 - \xi_i, i \in [n], \tag{A.2}$$

$$\sum_{j=1}^{W}(\mathbf{w}_j^+ + \mathbf{w}_j^-) \leq K, \tag{A.3}$$

$$\mathbf{w}_j^+, \mathbf{w}_j^-, \xi_i \geq 0, i \in [n], j \in [W]. \tag{A.4}$$

$$\mathbf{w}_j^+\mathbf{w}_j^- = 0, j \in [W]. \tag{A.5}$$

- Model 3: Relaxation of Model 2 by removing constraint (A.5).

$$\min_{\mathbf{w}^+,\mathbf{w}^-,b,\xi} \quad \tfrac{1}{2}(\mathbf{w}^+)'\mathbf{w}^+ + \tfrac{1}{2}(\mathbf{w}^-)'\mathbf{w}^- + C\sum_{i=1}^{n}\xi_i \tag{A.6}$$

$$s.t. \quad y_i((\mathbf{w}^+ - \mathbf{w}^-)'\mathbf{x}_i + b) \geq 1 - \xi_i, i \in [n], \tag{A.7}$$

$$\sum_{j=1}^{W}(\mathbf{w}_j^+ + \mathbf{w}_j^-) \leq K, \tag{A.8}$$

$$\mathbf{w}_j^+, \mathbf{w}_j^-, \xi_i \geq 0, i \in [n], j \in [W]. \tag{A.9}$$

Lemma 5 *The objective function of Model 2 can be simplified to the objective function of Model 3.*

$$\frac{1}{2}(\mathbf{w}^+ - \mathbf{w}^-)'(\mathbf{w}^+ - \mathbf{w}^-) + C\sum_{i=1}^{n}\xi_i = \frac{1}{2}(\mathbf{w}^+)'\mathbf{w}^+ + \frac{1}{2}(\mathbf{w}^-)'\mathbf{w}^- + C\sum_{i=1}^{n}\xi_i$$

Proof It is intuitive to remove the cross-term in the objective function since $(\mathbf{w}^+)'\mathbf{w}^- = \sum_{j=1}^{W}\mathbf{w}_j^+\mathbf{w}_j^-$ should be zero under constraint (A.5). ■

Lemma 6 *The optimal solution given by Model 3 satisfies constraint (A.5), i.e., any optimal solution of Model 3 is also a feasible solution of Model 2.*

Proof Proof by contradiction. Suppose we have reached optimal solutions $(\mathbf{w}^+, \mathbf{w}^-, b, \xi)$ given by Model 3. Suppose $\mathbf{w}_j^+ > 0, \mathbf{w}_j^- > 0$ for some j, i.e., $\mathbf{w}_j^+, \mathbf{w}_j^-$ do not satisfy

constraint (A.5). We can then define vector $\mathbf{c}$ s.t. $c_j = -min(\mathbf{w}_j^+, \mathbf{w}_j^-) < 0, c_k = 0$, for $k \neq j$, and construct a better solution $(\mathbf{w}^+ + \mathbf{c}, \mathbf{w}^- + \mathbf{c}, b, \xi)$.

Using this solution and because $\mathbf{w}_j^+, \mathbf{w}_j^-, -c_j > 0$, we have smaller objective function value, i.e.,

$$(\mathbf{w}^+)'\mathbf{w}^+ + (\mathbf{w}^-)'\mathbf{w}^- > (\mathbf{w}^+ + \mathbf{c})'(\mathbf{w}^+ + \mathbf{c}) + (\mathbf{w}^- + \mathbf{c})'(\mathbf{w}^- + \mathbf{c}) \tag{A.10}$$

The constraint (A.8) is still active, i.e.,

$$\sum_{j=1}^{W} (\mathbf{w}_j^+ + \mathbf{w}_j^-) - 2\min(\mathbf{w}_j^+, \mathbf{w}_j^-) \leq \sum_{j=1}^{W} (\mathbf{w}_j^+ + \mathbf{w}_j^-) \leq K.$$

This contradicts the fact that $\mathbf{w}^+, \mathbf{w}^-$ are optimal solutions of Model 3. ■

Theorem 7 *Model 1 is equivalent to Model 2, which is equivalent to Model 3.*

Proof It is easy to verify Model 1 is equivalent to Model 2. Any feasible solution of Model 2 is also feasible solution of Model 3, and we prove any optimal solution of Model 3 is also a feasible solution of Model 2 in Lemma 6. In conclusion, we prove the equivalence of all models for SLSVM. ■

References

Angün, E. and Altınoy, A. (2019). A new mixed-integer linear programming formulation for multiple responses regression clustering. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1634–1639. IEEE.

Bagirov, A. M., Mahmood, A., and Barton, A. (2017). Prediction of monthly rainfall in victoria, australia: Clusterwise linear regression approach. *Atmospheric Research*, 188:20–29.

Bailey, M. K., Weiss, A. J., Barrett, M. L., and Jiang, H. J. (2019). Characteristics of 30-day all-cause hospital readmissions, 2010-2016 [statistical brief# 248]. *Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from www. hcup-us. ahrq. gov/reports/statbriefs/sb248-Hospital-Readmissions-2010-2016. jsp.*

Balakrishnan, S., Wainwright, M. J., Yu, B., et al. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106(7):1039–1082.

Bertsimas, D. and King, A. (2015). Or forum—an algorithmic approach to linear regression. *Operations Research*, 64(1):2–16.

Bertsimas, D., King, A., Mazumder, R., et al. (2016). Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852.

Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2017). From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 18(1):7133–7171.

Bertsimas, D. and Shioda, R. (2007). Classification and regression via integer optimization. *Operations Research*, 55(2):252–271.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brisimi, T. S., Xu, T., Wang, T., Dai, W., Adams, W. G., and Paschalidis, I. C. (2018a). Predicting chronic disease hospitalizations from electronic health records: An interpretable classification approach. *arXiv preprint arXiv:1801.01204*.

Brisimi, T. S., Xu, T., Wang, T., Dai, W., Adams, W. G., and Paschalidis, I. C. (2018b). Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proceedings of the IEEE*, 106(4):690–707.

Brisimi, T. S., Xu, T., Wang, T., Dai, W., and Paschalidis, I. C. (2019). Predicting diabetes-related hospitalizations based on electronic health records. *Statistical methods in medical research*, 28(12):3667–3682.

Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., and Gallego, B. (2016). Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, 23(3):553–561.

Centers for Medicare & Medicaid Services (2018). Readmissions reduction program. `https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html`. Accessed: 2018-02-22.

Chaganty, A. T. and Liang, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048.

Chakravarthy, V., Ryan, M. J., Jaffer, A., Golden, R., McClenton, R., Kim, J., Press, I., and Johnson, T. J. (2018). Efficacy of a transition clinic on hospital readmissions. *The American journal of medicine*, 131(2):178–184.

Chatterjee, S. (2013). Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*.

Chen, H.-F. and Guo, L. (2012). *Identification and stochastic adaptive control*. Springer Science & Business Media.

Chen, R. and Paschalidis, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564.

Chen, Y., Yi, X., and Caramanis, C. (2014). A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604.

Chhabra, K. R., Werner, R. M., and Dimick, J. B. (2019). Clinical accountability and measuring surgical readmissions. *JAMA network open*, 2(4):e191301–e191301.

Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers. *The Annals of Mathematical Statistics*, 36(2):552–558.

Chow, Y. S. and Teicher, H. (2012). *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dai, W., Brisimi, T. S., Adams, W. G., Mela, T., Saligrama, V., and Paschalidis, I. C. (2015). Prediction of hospitalization due to heart diseases by supervised learning methods. *International journal of medical informatics*, 84(3):189–197.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2):249–282.

EcoliWiki (2018). Metabolic network reconstructions. `http://ecoliwiki.net/colipedia/index.php/Metabolic_Network_Reconstructions`. Accessed: 2018-02-22.

Edirisinghe, J. N., Weisenhorn, P., Conrad, N., Xia, F., Overbeek, R., Stevens, R. L., and Henry, C. S. (2016). Modeling central metabolism and energy biosynthesis across microbial life. *BMC genomics*, 17(1):568.

Edwards, J. S., Ibarra, R. U., and Palsson, B. O. (2001). In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125.

Elena, S. and Lenski, R. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation, 2003. *Nature Reviews. Genetics*, 4:457.

Embree, M., Liu, J. K., Al-Bassam, M. M., and Zengler, K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, 112(50):15450–15455.

Escobar, G. J., Plimier, C., Greene, J. D., Liu, V., and Kipnis, P. (2019). Multiyear rehospitalization rates and hospital outcomes in an integrated health care system. *JAMA Network Open*, 2(12):e1916769–e1916769.

Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. (1999). Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences*, 96(17):9721–9726.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friesen, M. L., Saxer, G., Travisano, M., and Doebeli, M. (2004). Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in escherichia coli. *Evolution*, 58(2):245–260.

Gitman, I., Chen, J., Lei, E., and Dubrawski, A. (2018). Novel prediction techniques based on clusterwise linear regression. *arXiv preprint arXiv:1804.10742*.

Goel, R., Patel, E. U., Cushing, M. M., Frank, S. M., Ness, P. M., Takemoto, C. M., Vasovic, L. V., Sheth, S., Nellis, M. E., Shaz, B., et al. (2018). Association of perioperative red blood cell transfusions with venous thromboembolism in a north american registry. *JAMA surgery*, 153(9):826–833.

Gonzalez, A. A., Cruz, C. G., Dev, S., and Osborne, N. H. (2016). Indication for lower extremity revascularization and hospital profiling of readmissions. *Annals of vascular surgery*, 35:130–137.

Graham, L. A., Mull, H. J., Wagner, T. H., Morris, M. S., Rosen, A. K., Richman, J. S., Whittle, J., Burns, E., Copeland, L. A., Itani, K. M., et al. (2019). Comparison of a potential hospital quality metric with existing metrics for surgical quality–associated readmission. *JAMA network open*, 2(4):e191313–e191313.

Hand, P. and Joshi, B. (2018). A convex program for mixed linear regression with a recovery guarantee for well-separated data. *Information and Inference: A Journal of the IMA*, 7(3):563–579.

Hartney, M., Liu, Y., Velanovich, V., Fabri, P., Marcet, J., Grieco, M., Huang, S., and Zayas-Castro, J. (2014). Bounceback branchpoints: Using conditional inference trees to analyze readmissions. *Surgery*, 156(4):842–848.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296.

Hoek, T. A., Axelrod, K., Biancalani, T., Yurtsev, E. A., Liu, J., and Gore, J. (2016). Resource availability modulates the cooperative and competitive nature of a microbial cross-feeding mutualism. *PLoS biology*, 14(8):e1002540.

Hoffman, G. J., Liu, H., Alexander, N. B., Tinetti, M., Braun, T. M., and Min, L. C. (2019). Posthospital fall injuries and 30-day readmissions in adults 65 years and older. *JAMA network open*, 2(5):e194276–e194276.

Hollis, R. H., Singletary, B. A., McMurtrie, J. T., Graham, L. A., Richman, J. S., Holcomb, C. N., Itani, K. M., Maddox, T. M., and Hawn, M. T. (2016). Blood transfusion and 30-day mortality in patients with coronary artery disease and anemia following noncardiac surgery. *JAMA surgery*, 151(2):139–145.

Ingraham, A. M., Richards, K. E., Hall, B. L., and Ko, C. Y. (2010). Quality improvement in surgery: the american college of surgeons national surgical quality improvement program approach. *Advances in surgery*, 44(1):251–267.

International Labour Office (2008). *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. International Labour Office.

James, J. (2013). *Medicare Hospital Readmissions Reduction Program: To Improve Care and Lower Costs, Medicare Imposes a Financial Penalty on Hospitals with Excess Readmissions*. Project HOPE.

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). Advances in flux balance analysis. *Current opinion in biotechnology*, 14(5):491–496.

Kerner, A., Park, J., Williams, A., and Lin, X. N. (2012). A programmable escherichia coli consortium via tunable symbiosis. *PLoS One*, 7(3):e34032.

Kimbrough, C. W., Agle, S. C., Scoggins, C. R., Martin, R. C., Marvin, M. R., Davis, E. G., McMasters, K. M., and Jones, C. M. (2014). Factors predictive of readmission after hepatic resection for hepatocellular carcinoma. *Surgery*, 156(4):1039–1048.

Klitgord, N. and Segrè, D. (2010a). Environments that induce synthetic microbial ecosystems. *PLoS computational biology*, 6(11):e1001002.

Klitgord, N. and Segrè, D. (2010b). The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. In *Genome Informatics 2009: Genome Informatics Series Vol. 22*, pages 41–55. World Scientific.

Klitgord, N. and Segrè, D. (2011). Ecosystems biology of microbial metabolism. *Current opinion in biotechnology*, 22(4):541–546.

Lai, T. L., Wei, C. Z., et al. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166.

Lasater, K. B. and McHugh, M. D. (2016). Reducing hospital readmission disparities of older black and white adults after elective joint replacement: the role of nurse staffing. *Journal of the American Geriatrics Society*, 64(12):2593–2598.

Le Gac, M., Brazas, M. D., Bertrand, M., Tyerman, J. G., Spencer, C. C., Hancock, R. E., and Doebeli, M. (2008). Metabolic changes associated with adaptive diversification in escherichia coli. *Genetics*, 178(2):1049–1060.

Li, B. Y., Urish, K. L., Jacobs, B. L., He, C., Borza, T., Qin, Y., Min, H. S., Dupree, J. M., Ellimoottil, C., Hollenbeck, B. K., et al. (2019). Inaugural readmission penalties for total hip and total knee arthroplasty procedures under the hospital readmissions reduction program. *JAMA network open*, 2(11):e1916008–e1916008.

Li, Y. and Liang, Y. (2018). Learning mixtures of linear regressions with nearly optimal complexity. *arXiv preprint arXiv:1802.07895*.

Low, L. L., Lee, K. H., Ong, H., Eng, M., Wang, S., Tan, S. Y., Thumboo, J., and Liu, N. (2015). Predicting 30-day readmissions: performance of the lace index compared with a regression model among general medicine patients in singapore. *BioMed research international*, 2015: 169870.

Luo, Z. and Chou, E. (2006). Pavement condition prediction using clusterwise regression. *Transportation Research Record: Journal of the Transportation Research Board*, (1974):70–77.

McHugh, J. P., Foster, A., Mor, V., Shield, R. R., Trivedi, A. N., Wetle, T., Zinn, J. S., and Tyler, D. A. (2017). Reducing hospital readmissions through preferred networks of skilled nursing facilities. *Health Affairs*, 36(9):1591–1598.

Mee, M. T., Collins, J. J., Church, G. M., and Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences*, page 201405641.

Min, L. and Hoffman, G. J. (2019). Predicting readmissions—with a twist. *JAMA Network Open*, 2(10):e1912399–e1912399.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, 3(2):e00036–12.

Nielsen, B. (2005). Strong consistency results for least squares estimators in general vector autoregressions with deterministic terms. *Econometric Theory*, 21(3):534–561.

Ojala, M. and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863.

Orth, J. D., Fleming, R. M., and Palsson, B. O. (2010a). Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide. *EcoSal plus, 4(1). DOI: 10.1128/ecosalplus.10.2.1*.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010b). What is flux balance analysis? *Nature biotechnology*, 28(3):245–248.

O'Brien, E. J., Monk, J. M., and Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, 161(5):971–987.

Pande, S., Merker, H., Bohl, K., Reichelt, M., Schuster, S., De Figueiredo, L. F., Kaleta, C., and Kost, C. (2014). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME journal*, 8(5):953.

Paoletti, S., Juloski, A. L., Ferrari-Trecate, G., and Vidal, R. (2007). Identification of hybrid systems a tutorial. *European journal of control*, 13(2-3):242–260.

Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, 26(2):174–187.

Park, Y. W., Jiang, Y., Klabjan, D., and Williams, L. (2017). Algorithms for generalized clusterwise linear regression. *INFORMS Journal on Computing*, 29(2):301–317.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petrigliano, F. A., Bezrukov, N., Gamradt, S. C., and SooHoo, N. F. (2014). Factors predicting complication and reoperation rates following surgical fixation of proximal humeral fractures. *Journal of Bone and Joint Surgery. American Volume*, 96(18):1544–1551.

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003). An expanded genome-scale model of escherichia coli k-12 (i jr904 gsm/gpr). *Genome biology*, 4(9):R54.

Rosenthal, A. Z., Qi, Y., Hormoz, S., Park, J., Li, S. H.-J., and Elowitz, M. B. (2018). Metabolic interactions between dynamic bacterial subpopulations. *eLife*, 7:e33099.

Rosenzweig, R. F., Sharp, R., Treves, D. S., and Adams, J. (1994). Microbial evolution in a simple unstructured environment: genetic differentiation in escherichia coli. *Genetics*, 137(4):903–917.

Rowell, K. S., Turrentine, F. E., Hutter, M. M., Khuri, S. F., and Henderson, W. G. (2007). Use of national surgical quality improvement program data as a catalyst for quality improvement. *Journal of the American College of Surgeons*, 204(6):1293–1300.

Rozen, D. E. and Lenski, R. E. (2000). Long-term experimental evolution in escherichia coli. viii. dynamics of a balanced polymorphism. *The American Naturalist*, 155(1):24–35.

Rozen, D. E., Schneider, D., and Lenski, R. E. (2005). Long-term experimental evolution in escherichia coli. xiii. phylogenetic history of a balanced polymorphism. *Journal of Molecular Evolution*, 61(2):171–180.

Sauer, U., Cameron, D. C., and Bailey, J. E. (1998). Metabolic capacity of bacillus subtilis for the production of purine nucleosides, riboflavin, and folic acid. *Biotechnology and bioengineering*, 59(2):227–238.

Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. (1997). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765.

Schreiber, J. (2017). Pomegranate: fast and flexible probabilistic modeling in python. *The Journal of Machine Learning Research*, 18(1):5992–5997.

Shou, W., Ram, S., and Vilar, J. M. (2007). Synthetic cooperation in engineered yeast populations. *Proceedings of the National Academy of Sciences*, 104(6):1877–1882.

Spencer, C. C., Bertrand, M., Travisano, M., and Doebeli, M. (2007). Adaptive diversification in genes that regulate resource use in escherichia coli. *PLoS genetics*, 3(1):e15.

Spencer, C. C., Tyerman, J., Bertrand, M., and Doebeli, M. (2008). Adaptation increases the likelihood of diversification in an experimental bacterial lineage. *Proceedings of the National Academy of Sciences*, 105(5):1585–1589.

Squires, D. and Anderson, C. (2015). Us health care from a global perspective: spending, use of services, prices, and health in 13 countries. *The Commonwealth Fund*, 15:1–16.

Tan, S. Y., Low, L. L., Yang, Y., and Lee, K. H. (2013). Applicability of a previously validated readmission predictive index in medical patients in singapore: a retrospective study. *BMC health services research*, 13(1):366.

Thommes, M., Wang, T., Zhao, Q., Paschalidis, I. C., and Segrè, D. (2018). Designing metabolic division of labor in microbial communities. *bioRxiv*, page 442376.

Thommes, M., Wang, T., Zhao, Q., Paschalidis, I. C., and Segrè, D. (2019). Designing metabolic division of labor in microbial communities. *mSystems*, 4(2):e00263–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tosoian, J. J., Hicks, C. W., Cameron, J. L., Valero, V., Eckhauser, F. E., Hirose, K., Makary, M. A., Pawlik, T. M., Ahuja, N., Weiss, M. J., et al. (2015). Tracking early readmission after pancreatectomy to index and nonindex institutions: a more accurate assessment of readmission. *JAMA surgery*, 150(2):152–158.

Treves, D. S., Manning, S., and Adams, J. (1998). Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of escherichia coli. *Molecular biology and evolution*, 15(7):789–797.

Tsoi, R., Wu, F., Zhang, C., Bewick, S., Karig, D., and You, L. (2018). Metabolic division of labor in microbial systems. *Proceedings of the National Academy of Sciences*, page 201716888.

van Gestel, J., Vlamakis, H., and Kolter, R. (2015). From cell differentiation to cell collectives: Bacillus subtilis uses division of labor to migrate. *PLoS biology*, 13(4):e1002141.

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., Austin, P. C., and Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, 182(6):551–557.

Vest, J. R., Kern, L. M., Silver, M. D., Kaushal, R., and investigators, H. (2015). The potential for community-based health information exchange systems to reduce hospital readmissions. *Journal of the American Medical Informatics Association*, 22(2):435–442.

Vidal, R. (2008). Recursive identification of switched arx systems. *Automatica*, 44(9):2274–2287.

Vlamakis, H., Kolter, R., et al. (2015). Division of labor in biofilms: the ecology of cell differentiation. *Microbiology spectrum*, 3(2):MB–0002.

Wang, H., Robinson, R. D., Johnson, C., Zenarosa, N. R., Jayswal, R. D., Keithley, J., and Delaney, K. A. (2014). Using the lace index to predict hospital readmissions in congestive heart failure patients. *BMC cardiovascular disorders*, 14(1):97.

Wang, J. and Klein, H. G. (2010). Red blood cell transfusion in the treatment and management of anaemia: the search for the elusive transfusion trigger. *Vox sanguinis*, 98(1):2–11.

Wang, T. and Paschalidis, I. C. (2019a). Convergence of parameter estimates for regularized mixed linear regression models. *arXiv preprint arXiv:1903.09235*.

Wang, T. and Paschalidis, I. C. (2019b). Prescriptive cluster-dependent support vector machines with an application to reducing hospital readmissions. In *2019 18th European Control Conference (ECC)*, pages 1182–1187. IEEE.

Wei, C. et al. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *The Annals of Statistics*, 15(4):1667–1682.

White, E. K. (2018). The best supervisor. *Nature*, 562(7726):297–298.

Whitlock, E. L., Kim, H., and Auerbach, A. D. (2015). Harms associated with single unit perioperative transfusion: retrospective population based analysis. *BMJ: British Medical Journal*, 350:h3037.

Wintermute, E. H. and Silver, P. A. (2010). Emergent cooperation in microbial metabolism. *Molecular systems biology*, 6(1):407.

Xu, T., Brisimi, T. S., Wang, T., Dai, W., and Paschalidis, I. C. (2016). A joint sparse clustering and classification approach with applications to hospitalization prediction. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 4566–4571. IEEE.

Yen, I. E.-H., Lee, W.-C., Zhong, K., Chang, S.-E., Ravikumar, P. K., and Lin, S.-D. (2018). Mixlasso: Generalized mixed regression via convex atomic-norm regularization. In *Advances in Neural Information Processing Systems*, pages 10891–10899.

Yi, X. and Caramanis, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575.

Yi, X., Caramanis, C., and Sanghavi, S. (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621.

Yi, X., Caramanis, C., and Sanghavi, S. (2016). Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*.

Yin, D., Pedarsani, R., Chen, Y., and Ramchandran, K. (2018). Learning mixtures of sparse linear regressions using sparse graph codes. *IEEE Transactions on Information Theory*.

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences*, page 201421834.

Zhao, Q., Segrè, D., and Paschalidisy, I. C. (2016). Optimal allocation of metabolic functions among organisms in a microbial ecosystem. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pages 7063–7068. IEEE.

Zhong, K., Jain, P., and Dhillon, I. S. (2016). Mixed linear regression with multiple components. In *Advances in neural information processing systems*, pages 2190–2198.

Zomorrodi, A. R. and Segrè, D. (2017). Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nature Communications*, 8(1):1563.

CURRICULUM VITAE