Algorithmic advancements in discrete optimization

Applications to machine learning and healthcare operations

Jean Pauphilet

Abstract

To advance healthcare, operations research should address two concurrent goals: (1) Develop new algorithms for decision-making in a data-rich environment that answer key concerns from practitioners, e.g., scalability and interpretability. (2) Put its theory to the test of practice, to ensure a path towards impact. Accordingly, this thesis comprises two parts. We develop new algorithms for large-scale discrete optimization problems, with a focus on machine learning under sparsity, and implement a predictive and prescriptive approach to improve patient flow management at a large academic hospital.

1 Introduction

> "When I think of the hospital of the future, I think of a bunch of people sitting in a room full of screens and phones,"

says Toby Cosgrove, CEO of Cleveland Clinic [78]. Indeed, all major healthcare providers are rethinking how hospitals work, for the gap between populations' health needs and the care offered by systems organized around hospitals has grown ever wider. In the next ten years, hospitals will operate like air-traffic control centers whose role is to coordinate care across multiple facilities.

To support this transition and transform our healthcare system, research in operations and analytics should address two concurrent goals: First, develop new methods and algorithms for decision-making in a data rich environment, which answer key concerns from practitioners and regulators, such as reliability, interpretability and fairness. Second, put its models and algorithms to

the test of practice, to ensure a path towards implementation and impact. Accordingly, this thesis comprises two parts, which serve these two complementary objectives.

1.1 Methodological challenges in machine learning and optimization

To address the needs of practitioners in high-stake industries like healthcare, we highlight two broad research questions where algorithmic advances are needed: sparsity and interpretability in machine learning; and large-scale discrete optimization.

Sparsity and interpretability in machine learning We use models to improve our knowledge of a given phenomenon. While the amount of available data has exploded in the past decades, human cognitive ability to understand complex models has remained limited. Hence, the identification of important variables within large data sets of high dimensionality has become increasingly valuable to practitioners and decision makers. Correspondingly, the notion of sparsity, i.e., the property of a model to involve a limited number of covariates, is cardinal in high-dimensional statistics.

Algorithms for large-scale discrete optimization The combinatorial component of sparse statistical learning problems is also widely present in the operations research literature at large. Indeed, start-up costs in machine scheduling, financial transaction costs, cardinality constraints, and fixed costs in facility location problems, among others, can all be modeled with binary decision variables. In particular, many practically relevant optimization problems involve logical relationship between some continuous variables x and binary variables z of the form "$x = 0$ if $z = 0$". In addition to the presence logical constraints, real-world instances of the OR/MS problems are increasingly larger in size, driven by the widespread adoption of connected devices and remote monitoring.

1.2 Practical challenges for analytics in healthcare

To improve the quality of care and alleviate the burden on clinicians and hospital staff, healthcare operations practitioners widely agree on the need to shift from isolated improvement in each individual units to a global coordination scheme across the entire hospital. For instance, [67] identified five guiding principles, among which the utilization of advanced data analytics to "forecast patient

demand patterns", and adopt "a system-wide approach to patient flow". In essence, they advocate for the development of both predictive and prescriptive analytics to improve hospital operations. However, there are unique challenges associated with healthcare analytics in practice.

Interpretability in healthcare In order to "forecast patient demand patterns", combining patient-level information from Electronic Health Records (EHRs) with sophisticated machine learning techniques can provide welcome visibility on patient flows and inform hospital operations. Despite the richness and increasing availability of data in healthcare, predictive models are not widely deployed in practice, due to the need to create custom dataset with specific variables for each predictive task, and the need for interpretable models.

A system view of hospital operations A system-wide approach to patient flow could certainly improve hospital operations and lead to better health outcomes for the patients at lower operational costs. Empirical and modeling work has helped us better understand patient flow, the impact of operational efficiency on quality of care, and the interdependence between units. For instance, we empirically observe that delays, a common measure of operational efficiency, lead to negative health outcomes, and that congestion in the ED is often due to unavailability of beds in inpatient units. Yet, day-to-day operations remain largely designed and optimized at a unit-level, and implementing hospital-wide strategies is easier said than done.

1.3 Outline

Chapter 2, 3, and 4 present methodological contributions to the discrete optimization literature, with particular emphasis on problems emerging from machine learning under sparsity. Chapter 5 and 6 present applications and implementation of machine learning and discrete optimization methods to improve operations at a large academic hospital in Boston.

2 Sparse Regression: A Discrete Optimization Perspective

The notion of sparsity, i.e., the ability to make predictions based on a limited number of covariates, has become cardinal in statistics. The so-called cardinality constrained estimators for instance minimize prediction error while explicitly bounding the number of input variables. Though computationally expensive, they have been considered as a relevant benchmark in high-dimensional statistics. Indeed, these estimators are characterized as the solution of the NP-hard problem [60]

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) \text{ s.t. } \|\boldsymbol{w}\|_0 \leq k, \tag{1}$$

where ℓ is an appropriate convex loss function (see Table 1 for examples). The covariates are denoted by the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, whose rows are the $\boldsymbol{x}_i^\top$'s, and the response data by $\boldsymbol{y} = (y_1, ..., y_n) \in \mathbb{R}^n$. Here, $\|\boldsymbol{w}\|_0 := |\{j : w_j \neq 0\}|$ denotes the 0-pseudo norm, i.e., the number of non-zero coefficients of $\boldsymbol{w}$. For decades, such problems have thus been solved using greedy heuristics, such as step-wise regression, matching pursuits [55], or recursive feature elimination (RFE) [44]. Consequently, much attention has been directed to convex surrogate estimators which tend to be sparse, while requiring less computational effort. The Lasso estimator, in which the ℓ_0 pseudo-norm of $\boldsymbol{w}$ in (1) is replaced by its ℓ_1 norm, and initially proposed by Tibshirani [80], is widely known and used. Its practical success can be explained by three concurrent ingredients: Efficient numerical algorithms exist [28, 39, 6], off-the-shelf implementations are publicly available [38] and recovery of the true sparsity is theoretically guaranteed under admittedly strong assumptions on the data [83]. However, recent works [82, 85, 33, 76, 40, 16] have pointed out several key deficiencies of the Lasso regressor in its ability to select the true features.

Table 1: Relevant loss functions ℓ and their corresponding Fenchel conjugates $\hat{\ell}$, defined as $\hat{\ell}(y, \alpha) := \sup_u \ \alpha u - \ell(y, u)$. The observed data is continuous, $y \in \mathbb{R}$, for regression and categorical, $y \in \{-1, 1\}$, for classification. By convention, $\hat{\ell}$ equals $+\infty$ outside of its domain.

Method	Loss $\ell(y, u)$	Fenchel conjugate $\hat{\ell}(y, \alpha)$
Ordinary Least Square	$\frac{1}{2}(y-u)^2$	$\frac{1}{2}\alpha^2 + y\alpha$
Logistic loss	$\log\left(1+e^{-yu}\right)$	$-y\alpha\log(-y\alpha) + (1+y\alpha)\log(1+y\alpha)$ for $y\alpha \in [-1, 0]$
1-norm SVM - Hinge loss	$\max(0, 1-yu)$	$y\alpha$ for $y\alpha \in [-1, 0]$

Therefore, new research in numerical algorithms for solving the exact formulation (1) directly has flourished. Leveraging recent advances in mixed-integer solvers [12, 11], Lagrangian relaxation [64], cyclic coordinate descent [45], or cutting-plane methods [18], these works have demonstrated significant improvement over existing Lasso-based heuristics. Another line of research has focused on replacing the ℓ_1 norm in the Lasso formulation by other sparsity-inducing penalties which are less sensitive to noise or correlation between features. In particular, non-convex penalties such as smoothly clipped absolute deviation (SCAD) [32] and minimax concave penalty (MCP) [87] have been proposed.

Convinced that sparsity is an extremely valuable property in high-impact applications where interpretability matters, and conscious that the profusion of research on the matter might have caused confusion and provided little guidance to practitioners, we propose with the present chapter a comprehensive treatment of state-of-the-art methods for feature selection in ordinary least square and logistic regression, and make the following contributions for solving the cardinality constrained formulation (1) exactly [16, 17]:

- We provide a unified treatment of state-of-the-art methods for feature selection in statistics. More precisely, we cover the cardinality constrained formulation (1), its Boolean relaxation, the Lasso formulation and its derivatives, and the MCP and SCAD penalty.

- We formulate the cardinality constrained formulation (1) with general convex loss fuction ℓ as a binary convex optimization problem Namely, we show by invoking strong duality that the following two problems are equivalent

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \quad \sum_{i=1}^n \ell(y_i, \boldsymbol{w}^\top \boldsymbol{x}_i) + \frac{1}{2\gamma}\|\boldsymbol{w}\|_2^2 \text{ s.t. } \|\boldsymbol{w}\|_0 \leq k, \tag{2}$$

$$\min_{\boldsymbol{z}\in\{0,1\}^p : \boldsymbol{z}^\top \boldsymbol{e} \leqslant k} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad -\sum_{i=1}^n \hat{\ell}(y_i, \alpha_i) - \frac{\gamma}{2}\sum_{j=1}^p z_j \boldsymbol{\alpha}^\top \boldsymbol{X}_j \boldsymbol{X}_j^\top \boldsymbol{\alpha}, \tag{3}$$

where $\hat{\ell}(y, \cdot)$ is the Fenchel conjugate of $\ell(y, \cdot)$ and can be derived explicitely for all problems of interest (see Table 1). Our result generalizes Bertsimas et al. [18], who only address the case of OLS regression and make extensive use of the closed-form solution available in this context.

Our framework, however, extends to cases where a closed-form solution is not available and includes, in addition to linear regression, logistic regression and SVM among others. Using this mixed-integer saddle-point formulation, we propose a tractable outer-approximation algorithm to solve it. Our cutting-plane algorithm scales to data sets for which n and p are in the $10,000$s and $100,000$s respectively for regression, and $1,000$s and $10,000$s respectively for classification.

- We propose an efficient sub-gradient algorithm to solve the Boolean relaxation of (1) and provide theoretical rate of convergence for our method. We make our code freely available as a `Julia` package named `SubsetSelection` (`https://github.com/jeanpauphilet/SubsetSelectionCIO.jl`). Our algorithm scales to problems with $n, p = 100,000$ or $n = 10,000$ and $p = 1,000,000$ within minutes, as reported on Table 2, while providing high-quality estimators.

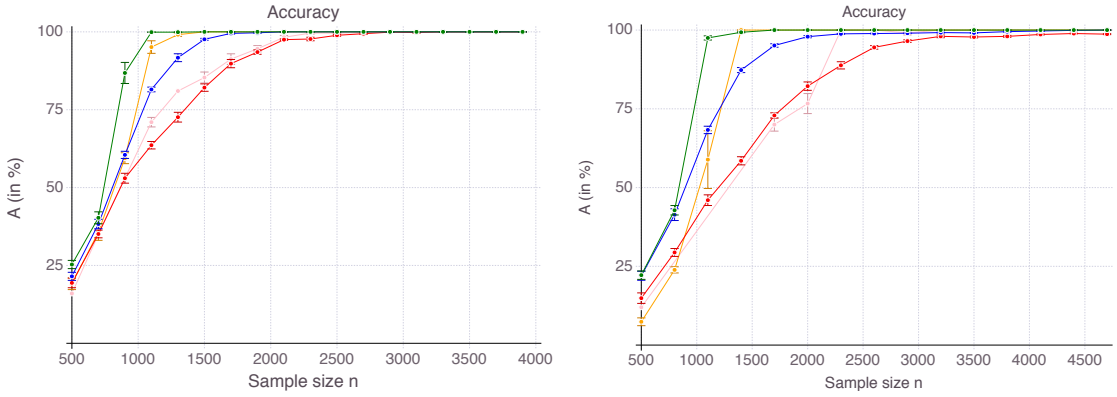
Table 2: Computational time of SS with $T_{max} = 200$ for data sets with large values of n and p, $\gamma = 2\,p/k/\max_i \|\boldsymbol{x}_i\|^2/n$. Due to the dimensionality of the data, computations where performed on 1 CPU with 250GB of memory. We provide the average computational time (and the standard deviation) over 10 experiments.

Loss function ℓ	n	p	k	time (in s)
Least Squares	$10,000$	$100,000$	100	12.90 (0.45)
Least Squares	$50,000$	$100,000$	100	28.45 (1.83)
Least Squares	$10,000$	$500,000$	100	33.00 (1.86)
Least Squares	$10,000$	$500,000$	500	43.00 (0.54)
Hinge Loss	$10,000$	$100,000$	100	37.26 (0.14)
Hinge Loss	$50,000$	$100,000$	100	160.73 (0.28)
Hinge Loss	$10,000$	$500,000$	100	157.09 (1.18)
Hinge Loss	$10,000$	$500,000$	500	59.74 (0.08)

- We compare the performance of all methods on three metrics of crucial interest in practice: accuracy - i.e, the proportion of true features which are selected - false detection rate - i.e., the proportion of selected features which are not in the true support - and computational tractability, and in various regimes of noise and correlation.

- In theory, Lasso estimators are only guaranteed to achieve 100% accuracy as $n \to +\infty$ under the so-called mutual incoherence condition (MIC). In Figure 1, we compare convergence in accuracy in settings with and without the MIC. We empirically observe what theory dictates:

Under MIC, the proportion of correct features selected converges to 1 as the sample size n increases for all methods, in all regimes of noise and correlation. Yet, on this matter, cardinality constrained and MCP formulations are the most accurate. As soon as MIC fails to hold however, ℓ_1-based estimators are inconsistent and $A < 1$, while discrete and non-convex penalties eventually perfectly recover the support.



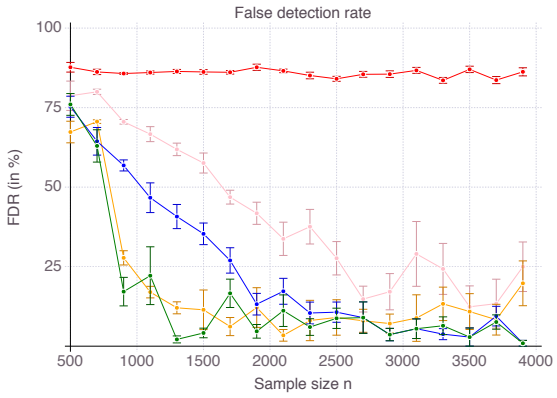
(a) MIC satisfied

(b) MIC violated

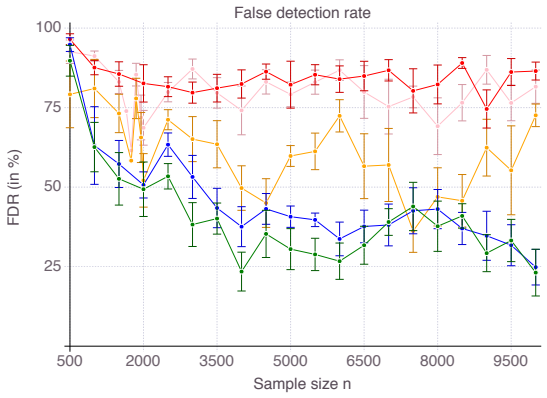
Figure 1: Accuracy as n increases, for the CIO (in green), SS (in blue with $T_{max} = 200$), ENet (in red), MCP (in orange), SCAD (in pink) with OLS loss, under a low noise regime, when the mutual incoherence condition is satisfied (left panel) or violated (right panel). We average results over 10 data sets.

- In addition, we also observe a convergence in false detection rate, namely the proportion of irrelevant features selected converging to 0 as the sample size n increases, for some but not all methods: The convex integer formulation and its Boolean relaxation are the only methods which demonstrate this behavior in low noise settings, and make the fewest false discoveries in other regimes. In our experiments (see Figure 2), Lasso-based estimators return at least 80% of non-significant features. MCP and SCAD have a low but strictly positive false detection rate (around $15 - 30\%$ in our experiments) as n increases and in all regimes.

- In terms of computational time, the integer optimization approach is unsurprisingly the most expensive option. Nonetheless, the computational cost is only one or two orders of magnitude higher than other alternatives and remains affordable in many real-world problems, even



(a) Low noise, low correlation



(b) High noise, high correlation

Figure 2: False detection rate FDR as n increases, for the CIO (in green), SS (in blue with $T_{max} = 200$), ENet (in red), MCP (in orange), SCAD (in pink) with OLS loss under the mutual incoherence condition. We average results over 10 data sets.

high-dimensional ones. Otherwise, the four remaining codes terminate in time comparable with the `glmnet` implementation of the Lasso, that is within seconds for $n = 1,000$ and $p = 20,000$.

3 A Unified Approach for Mixed-Integer Optimization

In addition to sparse regression, many important problems from the operations research literature exhibit a logical relationship between continuous variables x and binary variables z of the form "$x = 0$ if $z = 0$". Among others, start-up costs in machine scheduling problems, financial transaction costs, cardinality constraints and fixed costs in facility location problems exhibit this relationship. Since the work of Glover [42], this relationship is usually enforced through a "big-M" constraint of the form $|x| \leq Mz$ for a sufficiently large constant $M > 0$. Glover's work has been so influential that big-M constraints are now considered as intrinsic components of the initial problem formulations themselves, to the extent that textbooks in the field introduce facility location, network design or sparse portfolio problems with big-M constraints *by default*, although they are actually *reformulations* of logical constraints.

We consider optimization problems which unfold over two stages. In the first stage, a decision-maker activates binary variables, while satisfying resource budget constraints and incurring activation

costs. Subsequently, in the second stage, the decision-maker optimizes over the continuous variables. Formally, we consider the problem

$$\min_{\boldsymbol{z}\in\mathcal{Z},\boldsymbol{x}\in\mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{z} + g(\boldsymbol{x}) + \Omega(\boldsymbol{x}) \quad \text{s.t.} \quad x_i = 0 \text{ if } z_i = 0, \quad \forall i \in [n], \tag{4}$$

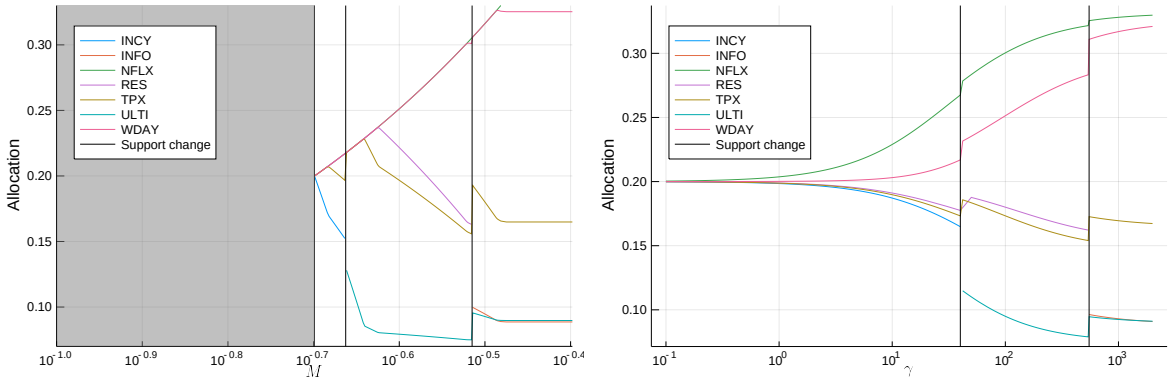
where $\mathcal{Z} \subseteq \{0,1\}^n$, $\boldsymbol{c} \in \mathbb{R}^n$ is a cost vector, $g(\cdot)$ is a generic convex function, and $\Omega(\cdot)$ is a convex regularization function of the following form:

Assumption 3.1. *In Problem* (4)*, the regularization term* $\Omega(\boldsymbol{x})$ *is one of:*

- *a big-M penalty function,* $\Omega(\boldsymbol{x}) = 0$ *if* $\|\boldsymbol{x}\|_\infty \leq M$ *and* ∞ *otherwise,*
- *a ridge penalty,* $\Omega(\boldsymbol{x}) = \dfrac{1}{2\gamma}\|\boldsymbol{x}\|_2^2$.

Conceptually, both regularization functions are equivalent to a soft or hard constraint on the continuous variables $\boldsymbol{x}$. However, they admit practical differences, as illustrated on a portfolio selection example on Figure 3: For big-M regularization, there usually exists a finite value M_0, typically unknown a priori, such that if $M < M_0$, the regularized problem is infeasible. Alternatively, for every value of the ridge regularization parameter γ, if the original problem is feasible then the regularized problem is also feasible. Consequently, if there is no natural choice of M then imposing ridge regularization may be less restrictive than imposing big-M regularization. However, for any $\gamma > 0$, the objective of the optimization problem with ridge regularization is different from its unregularized limit as $\gamma \to \infty$, while for big-M regularization, there usually exists a finite value M_1 above which the two objective values match. Yet, in practice, high values of M lead to numerical instability and provide low-quality bounds [see 5, Section 5].

Observe that the structure of Problem (4) is quite general, as the feasible set $\mathcal{Z}$ can capture known lower and upper bounds on $\boldsymbol{z}$, relationships between different z_i's, or a cardinality constraint $\boldsymbol{e}^\top \boldsymbol{z} \leq k$. Moreover, constraints of the form $\boldsymbol{x} \in \mathcal{X}$, for some convex set $\mathcal{X}$, can be encoded within the domain of g, by defining $g(\boldsymbol{x}) = +\infty$ if $\boldsymbol{x} \notin \mathcal{X}$. As a result, Problem (4) encompasses a large number of problems from the operations research literature, such as network design, facility location, sparse regression, and portfolio selection among others.



(a) Big-M regularization

(b) Ridge regularization

Figure 3: Optimal allocation of funds between securities as the regularization parameter (M or γ) increases. Data is obtained from the Russell 1000, with a cardinality budget of 5, a rank-200 approximation of the covariance matrix, a one-month holding period and an Arrow-Pratt coefficient of 1, as in Bertsimas, Cory-Wright [8]. Setting $M < \frac{1}{k}$ renders the entire problem infeasible

In this work [9], we provide three main contributions: First, we reformulate the logical constraint "$x_i = 0$ if $z_i = 0$" in a non-linear way, by substituting $z_i x_i$ for x_i in Problem (4). Second, we leverage the regularization term $\Omega(\boldsymbol{x})$ to derive a tractable reformulation of Problem (4). Finally, by invoking strong duality, we reformulate Problem (4) as a mixed-integer saddle-point problem, which is solvable via outer approximation. Precisely, we can summarize our contributions as follows:

- We identify a general class of mixed-integer optimization problems, which encompasses sparse regression, compressed sensing, sparse portfolio selection, unit commitment, facility location, network design, binary quadratic optimization, and sparse PCA as special cases.

- For this class of problems, we discuss how imposing either big-M or ridge regularization accounts for non-linear relationships between continuous and binary variables in a tractable fashion. In particular, we reformulate Problem (4) as a mixed-integer saddle point problem:

  Theorem 3.1. *Under some constraint qualification assumption, Problem* (4) *is equivalent to the following problem:*

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad \boldsymbol{c}^\top \boldsymbol{z} + h(\boldsymbol{\alpha}) - \sum_{i=1}^{n} z_i\, \Omega^\star(\alpha_i), \tag{5}$$

where $h(\boldsymbol{\alpha}) := \inf_{\boldsymbol{v}}\, g(\boldsymbol{v}) - \boldsymbol{v}^\top \boldsymbol{\alpha}$ *is, up to a sign, the Fenchel conjugate of* g *[see 52, Chap. 3.3],*

and

$$\Omega^\star(\beta) := M|\beta| \quad \textit{for the big-M penalty,}$$

$$\Omega^\star(\beta) := \tfrac{\gamma}{2}\beta^2 \qquad \textit{for the ridge penalty.}$$

Theorem 3.1 proves that regularization, either big-M or ridge, controls the convexity and smoothness of Problem (4), hence its computational tractability.

- We propose a conjunction of general-purpose numerical algorithms to solve Problem (4). The backbone of our approach is an outer approximation algorithm. Outer-approximation was proposed by Duran, Grossmann [27] and subsequently generalized by Fletcher, Leyffer [37]. Though slow in their original implementation, decomposition schemes have greatly benefited from recent improvements in mixed-integer linear solvers in the past decades, beginning with the branch-and-cut approaches of [63, 75]. In accordance with recent successful implementations of "modern" decomposition schemes [35, 36], we perform a rich root node analysis and enhance our algorithm with first-order methods to solve the Boolean relaxations and obtain improved lower bounds, certifiably near-optimal warm-starts via randomized rounding, and a discrete local search procedure.

- Finally, we demonstrate empirically that algorithms derived from our framework can outperform state-of-the-art solvers. For sparse portfolio selection, we solve to provable optimality problems one order of magnitude larger than previous attempts. On binary quadratic optimization problems with 100s of variables, we improve the objective value of the returned solution by 5 to 85% respectively, and our edge increases as the problem size increases. On network design problems, Table 3 reports the best solution found (the lower the better) after one hour, using CPLEX vs. our numerical blueprint and under both regularizations. From these results, we make two observations: First, irrespective of the regularization, our algorithm outperforms CPLEX and returns higher quality solutions. Second, we demonstrate that using ridge regularization can provide a substantial edge over using the big-M method to reformulate logical constraints.

Table 3: Best solution found after one hour on network design instances with m nodes and $(1+p)m$ initial edges. We report improvement, i.e., the relative difference between the solutions returned by CPLEX and the cutting-plane. Values are averaged over five randomly generated instances. For ridge regularization, we report the "unregularized" objective value, that is we fix $\boldsymbol{z}$ to the best solution found and resolve the corresponding sub-problem with big-M regularization. A "$-$" indicates that the solver could not finish the root node inspection within the time limit (one hour)

			Big-M			Ridge			Overall
m	p	unit	CPLEX	Cuts	Improv.	CPLEX	Cuts	Improv.	Improvement
40	1	$\times 10^{8}$	5.53	5.47	1.07%	5.97	**5.45**	8.74%	1.41%
80	1	$\times 10^{9}$	2.99	**2.94**	1.81%	3.16	2.95	6.78%	1.89%
120	1	$\times 10^{9}$	8.38	**7.82**	6.69%	–	**7.82**	–%	6.86%
160	1	$\times 10^{10}$	1.64	**1.54**	5.98%	–	**1.54**	–%	6.03%
200	1	$\times 10^{10}$	2.60	2.54	2.33%	–	**2.26**	–%	12.98%
40	2	$\times 10^{8}$	4.45	4.38	1.62%	4.76	**4.36**	8.27%	2.06%
80	2	$\times 10^{9}$	2.44	**2.31**	5.39%	2.46	**2.31**	5.97%	5.40%
120	2	$\times 10^{9}$	6.23	**5.89**	5.55%	–	**5.89**	–%	5.75%
160	2	$\times 10^{11}$	1.22	1.16	4.74%	–	**0.71**	–%	19.33%
200	2	$\times 10^{10}$	2.06	1.43	30.46%	–	**1.01**	–%	73.43%
40	3	$\times 10^{8}$	3.91	**3.85**	1.58%	4.13	**3.85**	6.73%	1.78%
80	3	$\times 10^{9}$	2.06	**1.94**	5.76%	2.04	**1.94**	5.44%	5.85%
120	3	$\times 10^{9}$	5.43	5.15	5.31%	–	**4.2**	–%	12.35%
40	4	$\times 10^{8}$	3.32	3.28	1.35%	3.53	**3.26**	7.71%	1.85%
80	4	$\times 10^{9}$	1.88	**1.77**	5.59%	–	**1.77**	–%	5.64%

4 Certifiably Optimal Sparse Inverse Covariance Estimation

In recent years, mixed-integer semidefinite optimization problems (MI-SDP) have received a lot of attention, for they naturally appear in robust optimization problems with ellipsoidal uncertainty sets [7] or as reformulations of combinatorial problems [71]. In this chapter, we extend the previous framework to one fundamental MI-SDP problem in modern multivariate analysis, namely estimating sparse inverse covariance (precision) matrices. Indeed, applications include undirected Gaussian graphical models [41], high dimensional discriminant analysis [20], portfolio allocation [30, 34], complex data visualization [81], amongst many others [see 31, for a review]. For example, in the context of undirected Gaussian graphical models, estimating the precision matrix corresponds to inferring the conditional independence structure on the related graphical model; zero entries in the precision matrix indicate that variables are conditionally independent.

Sparsity of the true precision matrix is a prevailing assumption [86, 19, 51, 29, 66]. In addition to desirable interpretability, sparsity is often needed for the problem to be well-defined. Indeed, when the

number of samples n is lower than the space dimension p, the empirical covariance matrix is singular, and thus cannot be inverted to obtain an estimate of the precision matrix. The most common method for encouraging sparsity in precision matrix estimation involves solving a ℓ_1-regularized maximum likelihood problem. The problem is convex and can be solved in high dimensions. Though this approach is tractable, solutions suffer from similar drawbacks as Lasso solutions in linear regression.

Here [13], we address these drawbacks by solving the cardinality constrained optimization problem for which the ℓ_1-regularized problem is a convex surrogate. Since the cardinality constrained problem parallels the relation the best feature selection problem plays in linear regression with Lasso, we investigate how methods developed in Chapter 2 could be naturally extended to the MI-SDP case. However, the sparse inverse covariance estimation problem constitutes a uniquely challenging instance from a theoretical and computational standpoint due to the semidefinite constraints. The main contribution of this work is to solve the cardinality constrained problem for problem sizes of interest, and compare the solutions with current approaches. Namely:

1. We prove that regularization, such as Lasso, is equivalent to a robust optimization version of inverse covariance estimation for an appropriately chosen uncertainty set, hence generalizing the seminal result of Banerjee et al. [4] and suggesting that regularization primarily encourages robustness rather than sparsity.

2. We formulate the cardinality constrained maximum likelihood problem for inverse covariance estimation as a binary optimization problem. We show that the resulting discrete optimization problem is non-smooth in general, but that adding a big-M penalty or a ridge regularization term penalty as in Chapter 3 leads to a smooth convex discrete optimization problem.

3. We propose a combination of outer-approximation algorithm and coordinate-descent methods to solve this problem. To our knowledge, this is the first time in which such a scheme is used to solve a mixed-integer nonlinear optimization problem with semidefinite constraints. It is well-known that problems of this type are notoriously hard to solve, and we observe that our approach significantly outperforms available mixed-integer nonlinear solvers. An advantage of

our approach over existing algorithms is that it provides near optimal solutions fast, and a guarantee on the solutions suboptimality if the method is terminated early.

4. We report computational results with both synthetic and real-world datasets that show that our proposed algorithm delivers near optimal solutions in a matter of seconds, and provably optimal solutions in a matter of minutes for p in the 100s and k in the 10s. The algorithm also provides high-quality solutions to problems in the $1,000$s without proof of optimality.

5. We investigate empirically statistical properties of solutions for the cardinality constrained problem. In Table 4, we compare the solutions for the two cardinality penalized formulations with big-M and ridge regularization, with the ℓ_1-regularized estimates (Glasso) and the Meinshausen and Bühlmann's approximation (MB). We observe that cardinality constrained estimates recover the sparsity pattern of the true underlying precision matrix (here, $k_{true} = 199$) with comparable accuracy as state-of-the-art but significantly lower false detection rate and improved predictive power (as measured by out-of-sample negative log-likelihood $-LL_{test}$). In addition, certifiably optimal covariance matrices are computed within 5 minutes, which, from a practical standpoint, is an affordable computational burden.

Table 4: Average performance on synthetic data with $p = 200$, $n/p = 1$, $t = 1\%$ (leading to $k_{true} = 199$), where the hyper-parameters of each formulation are chosen using the best negative log-likelihood over a validation set. We report the average performance over 10 instances (and their standard deviation)

Method	Big-M	Ridge	MB	Glasso
$k^\star$	199 (0)	199 (0)	796 (0)	796 (0)
A	95.1% (0.8)	95.1% (0.8)	99.6% (0.2)	99.5% (0.2)
FDR	4.9% (0.8)	4.9% (0.8)	67.9% (0.3)	75.1% (0.1)
$-LL_{test}$	141.39 (3.05)	141.37 (3.05)	157.11 (2.47)	162.05 (1.89)
Time (in s)	352.87 (11.12)	203.36 (39.00)	1.10 (0.04)	3.97 (0.31)

5 Predicting Inpatient Flow at a Major Hospital Using Interpretable Analytics

Healthcare offers a rewarding and impactful range of applications for analytics and operations research. For instance, combining patient-level information from Electronic Health Records (EHRs) with sophisticated predictive analytics can provide welcome visibility on patient flows and inform hospital operations. Such is the ambition of the present chapter.

In this chapter, we focus our attention on inpatients, namely patients who are admitted at the hospital and occupy a bed in an inpatient unit. For this population, patient flows can be divided into two categories: flows out of the hospital, i.e., discharges, and flows between units of the hospital. At a hospital level, a collection of work [49, 88, 54, 57] applied time-series methods to predict daily discharge volume. At a patient level, predicting discharge is associated with predicting length of stay. Being a surrogate for negative clinical outcomes as well as operational performance, length of stay has received a vivid interest in the academic literature, often in combination with hospital mortality (see [77, 84, 65, 68], and [2] for a review). Discharge destination, i.e., where the patient will be discharged to, is another important component of the discharge process. Regarding patient flows between units, the most critical ones are flows to and out of intensive care units (ICUs), for ICUs are expensive and limited resources needed by the most severe patients. In this work, we cover a range of patient flow-related predictive tasks, including predicting imminent discharges, long length of stay, discharge destination and hospital mortality, and need for an ICU bed in the next 24 hours. To the best of our knowledge, this last question has not been studied in the literature yet.

Despite the profusion of research and increasing availability of data in healthcare, predictive models are not widely deployed in practice, mainly due to the need to create custom dataset with specific variables for each predictive task. To address this issue, Nguyen et al. [61], Miotto et al. [59], Rajkomar et al. [65] proposed automatized patient representation strategies which analyze EHRs and construct relevant features in an unsupervised way using autoencoder neural networks. Since these approaches do not require an expert to manually define features, they are allegedly more

scalable. Surprisingly, however, and to the best of our knowledge, none of these approaches has been integrated within an EHR system of a real-world hospital despite their excellent predictive power on retrospective studies, including the most recent one [65]. In our opinion, they undermined three major implementation bottlenecks. First of all, the black-box nature of deep learning models impedes adoption from doctors and caregivers which are not engaged in the modeling process. Secondly, deep learning approaches are extremely expensive in terms of data, human and computing resources, and environmental costs [74]. Finally, convolutional and recurrent neural network are excellent at handling unstructured data such as medical notes. However, in practice, notes are rarely available in real-time and raise data privacy issues, especially if third-party computational resources are needed. Consequently, we believe they are better suited for retrospective clinical studies than production-ready real-time analytics.

In this chapter [15], we demonstrate how tailored modeling can be used in combination with interpretable machine learning techniques to provide accurate predictions on critical aspects of patient flows. To the best of our knowledge, our study is first of its kind to (a) address the length of stay and discharge destination prediction task for such a generic inpatient population with a unified data modeling and processing, (b) achieve state-of-the-art accuracy with a broad collection of models, including interpretable ones, (c) be fully integrated into the EHR system of a major hospital, thus demonstrating how powerful analytics can concretely impact care delivery. Specifically, our contributions can be summarized as follows:

- We propose a simple expertise-driven patient representation framework to capture the state of each inpatient as she stays in the hospital, competitive with the deep learning approaches recently proposed in the literature [61, 59, 65]. Compared to previous work, we use a hospital-centric rather than patient-centric time scale and only use features which are reliably available after admission, on a daily basis. Consequently, we successfully implement and integrate our patient representation into an EHR system and now process the data of 600 patients daily.

- From this unique set of features, we apply a broad collection of machine learning techniques

to address four length of stay-related tasks: identify same-day and next-day discharges and predict more-than-7 and more-than-14-day stays. We then investigate the question of predicting discharge destination among home, home with services, extended care facility and death. We also predict the probability for a given patient to need an intensive care bed in the next 24 hours. For all tasks, we match or surpass state-of-the-art methods with out-of-sample accuracy in the 80%+ range, even without using raw medical notes. Table 5 reports the fraction of these results corresponding to length of stay-related tasks. Sparse linear models and decision trees provide very good predictive power, together with actionable insights to practitioners thanks to their interpretability.

Table 5: Summary of the results on predicting length of stay (overall and remaining) for logistic regression (LR), CART decision trees (CART), optimal trees with parallel splits (OT), random forest (RF) and gradient boosted trees (GBT). MAE = Median Absolute Error. MRE = Median Relative Error.

	LR	CART	OT	RF	GBT
Classification: remaining length of stay < 1 day					
AUC	0.826	0.807	0.810	**0.843**	0.839
MAE in # daily discharges, no.	8.6	**6.0**	6.4	6.2	7.8
MRE in # daily discharges, %	8.7	6.0	6.5	**5.8**	7.6
Out-of-sample R^2	0.730	0.868	0.847	0.841	0.804
Classification: remaining length of stay < 2 days					
AUC	0.809	0.786	0.790	0.815	**0.822**
Classification: overall length of stay < 7 days					
AUC	0.818	0.775	0.776	0.813	**0.820**
AUC at day 1	0.827	0.795	0.797	0.828	**0.830**
AUC at day 2	**0.807**	0.752	0.752	0.800	0.804
Classification: overall length of stay < 14 days					
AUC	**0.826**	0.777	0.777	0.820	0.794

- The successful integration of our models into the EHR system of a large medical institution constitutes a salient characteristic and major achievement of our work. Figure 4 displays the machine learning-informed dashboard we built for bed management at BIDMC. In our opinion, our successful implementation illustrates that emphasis on modeling and interpretability does not hinder predictive accuracy nor scalability. On the contrary, the variety of predictive tasks we cover, with high level of accuracy, demonstrates that an expertise-driven patient representation framework can be equally powerful and versatile as neural network approaches. In addition, it

leads to more interpretable features, achieves higher engagement from the clinicians and care providers, and requires less data and computational resources. As a result, we were able to conduct the project from initial data exploration to production-level deployment in less than twelve months.

Projected Discharges

Ward	Ward Type	Ward Campus	Ward Census	Number Discharges	ICU Needed (Projected for Non-ICU Floor)
FA3	Cardiac	West	31	8	0
FA5	Cardiac	West	27	9	0
FA8	Cardiac	West	19	3	0
CC1A	Clinical	West	2	1	0
4I	Critical Care	East	9	0	5
CC5B	Critical Care	West	9	0	7
CC6B	Critical Care	West	6	0	4
CC6C	Critical Care	West	7	0	5
CC6D	Critical Care	West	7	0	5
CC7B	Critical Care	West	6	1	4
CC7C	Critical Care	West	7	0	5
CC7D	Critical Care	West	6	0	4
FA6B	Critical Care	West	8	1	7
CC1B	Emergency	West	13	12	0
11R	Med/Surg	East	35	8	0
12R	Med/Surg	East	42	13	0
5S	Med/Surg	East	24	11	0
7F	Med/Surg	East	32	5	0
7S	Med/Surg	East	17	3	0
8S	Med/Surg	East	21	5	0
CC6A	Med/Surg	West	34	10	0
CC7A	Med/Surg	West	32	8	1
FA10	Med/Surg	West	31	6	0
FA11	Med/Surg	West	29	11	0
FA2	Med/Surg	West	32	6	0
FA6A	Med/Surg	West	13	3	0
FA7A	Med/Surg	West	28	6	0
FA9A	Med/Surg	West	31	10	0
DEA1	Observation	West	5	5	0
Total			**563**	**145**	**47**

Figure 4: Screenshot of the capacity prediction tool built for the office of bed management. The dashboard displays a list of all the hospital wards with census level, expected number of discharges and expected number of ICU patients by the end of the day

- Finally, our work led to substantial operational benefits for the hospital. When it comes to predicting discharges, our models achieve a significantly lower median relative error than estimates obtained from inquiring resource nurses directly (11.5% vs. 16.0%). In addition to being less accurate, asking resource nurses of each floor about their daily predictions is a tedious process and is very sensitive to discrepancy in experience between nurses. From an

operational perspective, we analyze the impact of our models on admission delays of patient from the Emergency Department (ED) and off-service placement. A difference-in-differences analysis, reported in Table 6, reveals a significant reduction in off-service placement thanks to our tool (by 4%). Less significantly, we also observe a negative effect on (i.e., a reduction in) boarding delays.

Table 6: Difference-in-differences analysis of boarding delays and off-service placement between April and July 2019. Our predictive analytics were implemented in May-June 2019. We use April and July 2018 as control data. The variable "2019 indicator" captures the changes in activity between 2018 and 2019, "July indicator" captures the monthly seasonality between April and July, and "2019 indicator × July indicator" captures the effect of our intervention. We report estimates for each coefficient (with standard errors) and level of significance (" ", "< 0.1", "< 0.01" and "< 0.001").

	Boarding delay		Off-service placement	
	Coefficient	p-value	Coefficient	p-value
2019 indicator	0.231 (0.122)	< 0.1	0.053 (0.012)	< 0.001
July indicator	0.442 (0.115)	< 0.001	0.032 (0.011)	< 0.01
2019 indicator × July indicator	-0.120 (0.174)		-0.043 (0.017)	< 0.01
Controls: day-of-the-week, hour of the day; Observations: $5,126$				

6 Hospital-wide Patient Flow Optimization

To convert predictions on future patient discharges and flows, as developed in the previous chapter, into actionable bed placement recommendations, and improve operational efficiency of hospitals even further, one needs to develop "a system-wide approach to patient flow" [67]. In this chapter, we propose a holistic optimization approach combined with machine learning techniques to achieve this goal. Based on historical data from a large academic hospital, we demonstrate that our approach can be implemented in a real-world environment and effectively reduces delays and patient misplacement.

A central performance metric in patient flow management is delays. Indeed, delays can be used as a measure of operational efficiency as well as quality of care. Empirical work suggests that prolonged ED boarding time - the time needed for a patient in the ED to be admitted to an inpatient bed - is associated with negative health outcomes [56, 22]. Prolonged ED boarding time is usually due to unavailability of inpatient beds [69]. Consequently, better understanding and modeling of discharge

patterns are needed as well [69, 21, 26, 25]. Besides the ED, Johnson et al. [47], Long, Mathews [53], Oliveira et al. [62] empirically measure the negative consequences of prolonged intensive care unit (ICU) boarding. Finally, Green [43] surveys the potential for OR techniques in reducing hospital delays, with an emphasis on queueing models.

A general insight of queueing theory is also that resource pooling might produce better performance. Due to heterogenity in patient needs however, empirical studies have found that pooling resources in the ED can be detrimental to the patient, by increasing mortality [73, 3], readmission risk [72, 70], or overall length-of-stay [1, 72, 70]. Indeed, in an inpatient context, pooling resources leads to patient misplacement, also called off-service placement or patient overflow. Off-service placement occurs when an incoming patient is placed in a unit designated for a different service than the service required by her condition. Another related phenomenon is off-level placement [50, 23], that is, when a patient needing an ICU is placed in a general care unit.

As far as ED boarding is concerned, there is a trade-off between waiting in the ED for the right bed to become available and immediately placing the patient in another service. Thompson et al. [79], Kilinc et al. [48], Dai, Shi [24] explore this trade-off using a queueing framework and a Markov decision process model. However, as the authors acknowledge, their analysis does not scale to large medical institutions and can only be applied to a restricted number of units. Also, they do not account for inter-unit transfers of inpatients.

Our present chapter [14] falls into this last line of work but differs substantially in terms of scope and methodology: First, we adopt a holistic approach to optimize bed assignment decisions simultaneously for all inpatient units. To the best of our knowledge, no study has previously addressed the question in such breadth. Secondly, we build our analysis on data rather than stochastic modeling and distributional assumptions. Given the data rich environment that hospitals have become, we believe that queueing models are less meaningful, especially due to their stringent assumptions and the curse of dimensionality they suffer from. Finally, we integrated of our model into the EHR of a 600-bed hospital. Our solution is undergoing final calibration and testing before full implementation at the hospital by the end of the year. Our contributions can be summarized as follows:

1. We consider the entirety of the hospital and optimize patient flows at a system level, while previous work mostly focused on isolated units or a sub-network comprised of the ED and some inpatient wards. Our approach not only accounts for admission of ED patients to inpatient beds but also for outside transfers, surgical patients boarding from the post-anesthesia care units (PACUs), and patient flows within inpatient units. Figure 5 sketches the main patient flows accounted for in our framework.

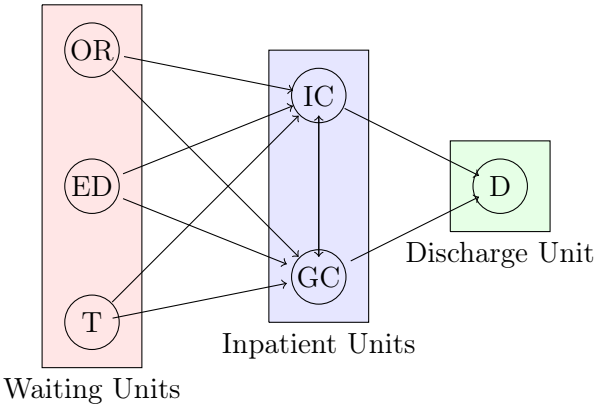


Figure 5: Schematic views of patient flows in a typical hospital. New requests for beds can either come from surgeries or scheduled admissions (OR), the emergency department (ED) or transfers from another institution (T). Inpatient units are characterized by their medical specialty or service, and their level of care. We present here two levels of care, namely intensive care (IC) and general care (GC). For simplicity, we model discharges through admission to a single virtual discharge unit (D), although multiple discharge destinations are possible.

2. We describe the location of each patient individually using integer decision variables, as opposed to stochastic queueing models which, at scale, rely on fluid model assumptions that dissolve individual movements into continuous flows. This distinction is relevant in practice because of tight capacity constraints which make each unit of capacity matter.

3. From a modeling perspective, we associate each patient with two locations, namely a physical location corresponding to the hospital unit she physically is, and a virtual location corresponding to the unit she *should be in* given her clinical need. Correspondingly, we can divide patient flows into two categories: Physical patient flows, which result from hospital management decisions to accept, place and discharge patients; Clinical flows, which are uncertain quantities.

This perspective has the double advantage of being simple and dissociating the operational from the clinical decision-making process, the later being modeled as an uncertain quantity driven by each patient's condition rather than a decision variable.

4. To account for uncertainty in the clinical trajectories, we integrate predictions obtained from machine learning techniques into an overall robust optimization formulation. While queueing models have been successfully used to model patient flows at a unit level, stochastic analysis of patients flows for hospital-wide bed assignment might be intractable due to non-stationarity of the arrival and departure processes, intricate network structure between the different units, soft pre-assignment rules of services to units, and high dimensions. In this work, we use outputs from machine learning models to build data-driven uncertainty sets for patients' clinical flows. Table 7 summarizes the main key clinical flows we are predicting, the models we used and their respective out-of-sample accuracy.

Table 7: Out-of-sample performance for all patient-flow prediction tasks, on their respective test set. We use optimal classification trees (OCT) [10] for classification tasks and regularized regression (Lasso) [80] for regression tasks

Patient category	Prediction task	Method	Metric	Value
Inpatients	Probability of discharge	OCT	AUC	0.810
	Daily discharges	OCT	Median relative error R^2	6.0% 0.847
	Probability of intensive care	OCT	AUC	0.973
	ICU census	OCT	Median relative error R^2	11.1% 0.998
ED	Bed requests	Lasso	Median absolute error Median relative error R^2	3.67 14.0% 0.910
Transfers	Bed requests	Lasso	Median absolute error Median relative error R^2	1.19 58.1% 0.805

5. In this framework, the optimal bed allocation decisions can be formulated as minimizing the mismatch between the physical flows (decisions) and the clinical ones (uncertainty), and leads to a tractable mixed-integer robust optimization problem, which we will later refer to as (H_2O). To the best of our knowledge, this formulation is novel, simple, and captures many of the

operational deficiencies observed empirically such as boarding delays and off-service placement. While robust optimization has been previously applied to hospital operations [46, 58], our work constitutes, to the best of our knowledge, the first implementation of adaptive robust techniques in this setting.

6. Finally, we demonstrate that our proposed formulation is tractable and leads to significant operational benefit. On data from a 600-bed medical center over 7 months, we solve the adaptive robust optimization problems in seconds and provide a bed assignment policy which reduces off-service placement by 33% on average, boarding delays in the emergency departments and post-anesthesia units by 30% and 19% respectively, while keeping overall occupation constant. Figure 6 summarizes the relative improvement of (H_2O) over the current bed assignment strategy for four metrics. Although there is a clear trade-off between quality of the bed assigned and time waited for the assignement, our simulations suggest that there is an opportunity for hospitals to improve on both aspects simultaneously compared to how they currently operate, by leveraging advanced prescriptive analytics. On this regard, we also demonstrate the additional benefit from using linear decision rules that allow for a more effective and flexible trade-off between waiting time and off-service placement, as displayed on Figure 7.

7 Conlusions

In this thesis, we have illustrated how interpretable machine learning and hospital operations can benefit from improvements in large-scale discrete optimization.

The first part provides two main contributions: From a modeling perspective, the use of ridge regularization, that is a non-linear yet strongly convex term, to encode logical relationships between continuous and discrete variables. From an algorithmic perspective, a generic cutting-plane strategy to numerically solve mixed-integer optimization problems at scale. Chapter 3 formally presents the class of problems that can be addressed with our framework, together with theoretical analysis and

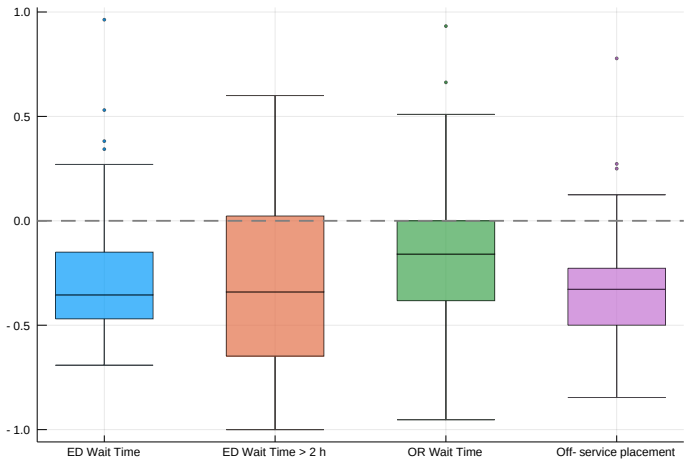


Figure 6: Boxplot for the distribution of the relative difference between the (H_2O) policy over the historical bed assignment decisions in terms of four performance metrics: (from left to right) ED wait, number of ED patients waiting more than 2 hours, OR wait, number of off-service placements. Negative values indicate a reduction, hence an improvement.

numerical experiments on many special cases. Chapter 2 focuses on sparse empirical minimization and exploits extra problem structure to design an efficient first-order heuristic. Future research could explore how our general blueprint can be tailored to other specific problem structures. We also believe that numerical ingredients like multi-threading, warm-starts, and cut sharing, could improve the implementation of the cutting-plane algorithm even further and ought to be investigated in the future. In Chapter 4, we extend this framework to a special case of mixed-integer semidefinite optimization. In our opinion, our approach could benefit an even broader class of problems in mixed-integer semidefinite and rank constrained optimization.

The second part covers the application and implementation of machine learning and discrete optimization techniques to first predict (Chapter 5) and then optimize (Chapter 6) patient flows at a large academic hospital. From our experience, we draw three conclusions: First, that data and technology already available can lead to significant operational improvement. Second, that modeling constitutes the main implementation bottleneck for practitioners. While data and computing resources become commodities, hospital managers increasingly need data-driven models that capture the majority of their daily operations and scale to the size of their system. This constitutes, in our opinion, the most promising direction for future research in the field. Finally, that optimization

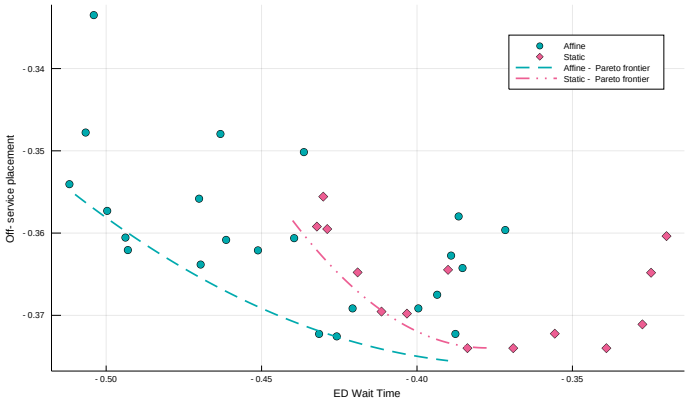


Figure 7: Trade-off waiting time in the emergency department vs. off-service placements. All quantities are relative values compared with historical placements. We compare the performance of the proposed robust affine policies (blue circles) with the static solution (pink diamonds).

is crucial to translate estimates into actionable insights and impact, as machine learning pervades EHRs and provides more predictions than what can be reasonably handled by humans.

Our main objective in this work has been to develop predictive and prescriptive analytics *in collaboration* with a medical institution. Unfortunately, research - even operations research - is often considered as disconnected from real-world problems, which rarely benefit from the research they supposedly motivated. On the other hand, calibration alone does not constitute a scientific contribution. So, defining how to properly value implementation in academic research remains an open question and a critical challenge for our community. We hope this thesis could contribute to our collective answer. It illustrates, we believe, fruitful synergies between academia and industry, with equal emphasis on methodological advancements that are relevant in practice and methodologically grounded applied operations research.

References

[1] *Alameda César, Suárez Carmen.* Clinical outcomes in medical outliers admitted to hospital with heart failure // European Journal of Internal Medicine. 2009. 20, 8. 764–767.

[2] *Awad Aya, Bader-El-Den Mohamed, McNicholas James.* Patient length of stay and mortality prediction: A survey // Health Services Management Research. 2017. 30, 2. 105–120.

[3] *Bai Anthony D., Srivastava Siddhartha, Tomlinson George A., Smith Christopher A., Bell Chaim M., Gill Sudeep S.* Mortality of hospitalised internal medicine patients bedspaced to non-internal medicine inpatient units: Retrospective cohort study // BMJ Quality and Safety. 2018. 27, 1. 11–20.

[4] *Banerjee Onureena, El Ghaoui Laurent, D'Aspremont Alexandre.* Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data // Journal of Machine Learning Research. 2008. 9, Mar. 485–516.

[5] *Beaumont Nicholas.* An algorithm for disjunctive programs // European Journal of Operational Research. 1990. 48, 3. 362–371.

[6] *Beck Amir, Teboulle Marc.* A fast iterative shrinkage-thresholding algorithm for linear inverse problems // SIAM Journal on Imaging Sciences. 2009. 2, 1. 183–202.

[7] *Ben-Tal Aharon, El Ghaoui Laurent, Nemirovski Arkadi.* Robust optimization. 2009.

[8] *Bertsimas Dimitris, Cory-Wright Ryan.* A Scalable Algorithm for Sparse Portfolio Selection // arXiv preprint arXiv:1811.00138. 2018.

[9] *Bertsimas Dimitris, Cory-Wright Ryan, Pauphilet Jean.* A Unified Approach to Mixed-Integer Optimization: Nonlinear Formulations and Scalable Algorithms // arXiv preprint arXiv:1907.02109. 2019.

[10] *Bertsimas Dimitris, Dunn Jack.* Optimal classification trees // Machine Learning. 2017. 106, 7. 1039–1082.

[11] *Bertsimas Dimitris, King Angela.* Logistic regression: From art to science // Statistical Science. 2017. 32, 3. 367–384.

[12] *Bertsimas Dimitris, King Angela, Mazumder Rahul.* Best subset selection via a modern optimization lens // Annals of Statistics. 2016. 44, 2. 813–852.

[13] *Bertsimas Dimitris, Lamperski Jourdain, Pauphilet Jean.* Certifiably optimal sparse inverse covariance estimation // Mathematical Programming. 2019.

[14] *Bertsimas Dimitris, Pauphilet Jean.* Hospital-wide Patient Flow Optimization // Submitted. 2020.

[15] *Bertsimas Dimitris, Pauphilet Jean, Stevens Jennifer, Tandon Manu.* Predicting inpatient flow at a major hospital using interpretable analytics // medRxiv preprint medRxiv:2020.05.12.20098848. 2020.

[16] *Bertsimas Dimitris, Pauphilet Jean, Van Parys Bart.* Sparse Classification: a scalable discrete optimization perspective // arXiv preprint arXiv:1710.01352. 2018.

[17] *Bertsimas Dimitris, Pauphilet Jean, Van Parys Bart.* Sparse Regression: Scalable algorithms and empirical performance // Statistical Science, to appear. 2020.

[18] *Bertsimas Dimitris, Van Parys Bart, others* . Sparse high-dimensional regression: Exact scalable algorithms and phase transitions // The Annals of Statistics. 2020. 48, 1. 300–323.

[19] *Bickel Peter J., Levina Elizaveta.* Covariance regularization by thresholding // Annals of Statistics. 2008. 36, 6. 2577–2604.

[20] *Cai Tony, Liu Weidong, Luo Xi.* A constrained $\ell 1$ minimization approach to sparse precision matrix estimation // Journal of the American Statistical Association. 2011. 106, 494. 594–607.

[21] *Chan Carri W., Dong Jing, Green Linda V.* Queues with time-varying arrivals and inspections with applications to hospital discharge policies // Operations Research. 2017. 65, 2. 469–495.

[22] *Chan Carri W., Farias Vivek F., Escobar Gabriel J.* The impact of delays on service times in the intensive care unit // Management Science. 2017. 63, 7. 2049–2072.

[23] *Chan Carri W., Green Linda V., Lekwijit Suparerk, Lu Lijian, Escobar Gabriel.* Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units // Management Science. 2019. 65, 2. 751–775.

[24] *Dai Jim G., Shi Pengyi.* Inpatient Bed Overflow: An Approximate Dynamic Programming Approach // SSRN Electronic Journal. 2017.

[25] *Dai Jim G., Shi Pengyi.* Recent Modeling and Analytical Advances in Hospital Inpatient Flow Management // Production and Operations Management. 2019.

[26] *Dong Jing, Perry Ohad.* Queueing Models for Patient-Flow Dynamics in Inpatient Wards // SSRN Electronic Journal. 2018.

[27] *Duran Marco A., Grossmann Ignacio E.* An outer-approximation algorithm for a class of mixed-integer nonlinear programs // Mathematical Programming. 1987. 39, 3. 337–337.

[28] *Efron Bradley, Hastie Trevor, Johnstone Iain, Tibshirani Robert, Ishwaran Hemant, Knight Keith, Loubes Jean Michel, Massart Pascal, Madigan David, Ridgeway Greg, Rosset Saharon, Zhu J. I., Stine Robert A., Turlach Berwin A., Weisberg Sanford, Johnstone Iain, Tibshirani Robert.* Least angle regression // Annals of Statistics. 2004. 32, 2. 407–499.

[29] *El Karoui Noureddine.* High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation // Annals of Statistics. 2010. 38, 6. 3487–3566.

[30] *Fan Jianqing, Fan Yingying, Lv Jinchi.* High dimensional covariance matrix estimation using a factor model // Journal of Econometrics. 2008. 147, 1. 186–197.

[31] *Fan Jianqing, Han Fang, Liu Han.* Challenges of Big Data analysis // National Science Review. 2014. 1, 2. 293–314.

[32] *Fan Jianqing, Li Runze.* Variable selection via nonconcave penalized likelihood and its oracle properties // Journal of the American Statistical Association. 2001. 96, 456. 1348–1360.

[33] *Fan Jianqing, Song Rui.* Sure independence screening in generalized linear models with NP-dimensionality // Annals of Statistics. 2010. 38, 6. 3567–3604.

[34] *Fan Jianqing, Zhang Jingjin, Yu Ke.* Vast portfolio selection with gross-exposure constraints // Journal of the American Statistical Association. 2012. 107, 498. 592–606.

[35] *Fischetti Matteo, Ljubić Ivana, Sinnl Markus.* Benders decomposition without separability: A computational study for capacitated facility location problems // European Journal of Operational Research. 2016. 253, 3. 557–569.

[36] *Fischetti Matteo, Ljubic Ivana, Sinnl Markus.* Redesigning benders decomposition for large-scale facility location // Management Science. 2017. 63, 7. 2146–2162.

[37] *Fletcher Roger, Leyffer Sven.* Solving mixed integer nonlinear programs by outer approximation // Mathematical Programming. 1994. 66, 1-3. 327–349.

[38] *Friedman Jerome, Hastie Trevor, Tibshirani Rob.* glmnet: Lasso and elastic-net regularized generalized linear models. 2009.

[39] *Friedman Jerome, Hastie Trevor, Tibshirani Rob.* Regularization paths for generalized linear models via coordinate descent // Journal of Statistical Software. 2010. 33, 1. 1–22.

[40] *Gamarnik David, Zadik Ilias.* High-Dimensional Regression with Binary Coefficients. Estimating Squared Error and a Phase Transition // arXiv preprint arXiv:1701.04455. 2017.

[41] *García-Puente Luis, Petrović Sonja, Sullivant Seth.* Graphical models. 5, 1. 2013. 1–7.

[42] *Glover Fred.* Improved Linear Integer Programming Formulations of Nonlinear Integer Problems. // Management Science. 1975. 22, 4. 455–460.

[43] *Green Linda V.* Using Operations Research to Reduce Delays for Healthcare // State-of-the-Art Decision-Making Tools in the Information-Intensive Age. 2008. 1–16.

[44] *Guyon Isabelle, Weston Jason, Barnhill Stephen, Vapnik Vladimir.* Gene selection for cancer classification using support vector machines // Machine Learning. 2002. 46, 1-3. 389–422.

[45] *Hazimeh Hussein, Mazumder Rahul.* Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms // arXiv preprint arXiv:1803.01454. 2018.

[46] *He Shuangchi, Sim Melvyn, Zhang Meilin.* Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach // Management Science. 2019. 65, 9. 4123–4140.

[47] *Johnson Daniel W., Schmidt Ulrich H., Bittner Edward A., Christensen Benjamin, Levi Retsef, Pino Richard M.* Delay of transfer from the intensive care unit: A prospective observational study of incidence, causes, and financial impact // Critical Care. 2013. 17, 4. R128.

[48] *Kilinc Derya, Saghafian Soroush, Traub Stephen.* Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue? // SSRN Electronic Journal. 2017.

[49] *Kim Kibaek, Lee Changhyeok, Leary Kevin J O.* Predicting Patient Volumes in Hospital Medicine : A Comparative Study of Different Time Series Forecasting Methods // Argonne National Laboratory. 2014. d, January 23. 1–13.

[50] *Kim Song Hee, Chan Carri W., Olivares Marcelo, Escobar Gabriel.* ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes // Management Science. 2015. 61, 1. 19–38.

[51] *Lam Clifford, Fan Jianqing.* Sparsistency and rates of convergence in large covariance matrix estimation // Annals of Statistics. 2009. 37, 6 B. 4254–4278.

[52] *Lauritzen Niels.* Convex optimization. 2013. 223–251.

[53] *Long Elisa F., Mathews Kusum S.* The Boarding Patient: Effects of ICU and Hospital Occupancy Surges on Patient Flow // Production and Operations Management. 2018. 27, 12. 2122–2143.

[54] *Luo Li, Xu Xueru, Li Jialing, Shen Wenwu.* Short-term forecasting of hospital discharge volume based on time series analysis // 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). 2017. 1–6.

[55] *Mallat Stephane G., Zhang Zhifeng.* Matching Pursuits With Time-Frequency Dictionaries // IEEE Transactions on Signal Processing. 1993. 41, 12. 3397–3415.

[56] *Mathews Kusum S., Durst Matthew S., Vargas-Torres Carmen, Olson Ashley D., Mazumdar Madhu, Richardson Lynne D.* Effect of emergency department and ICU occupancy on admission decisions and outcomes for critically Ill Patients // Critical Care Medicine. 2018. 46, 5. 720–727.

[57] *McCoy Thomas H., Pellegrini Amelia M., Perlis Roy H.* Assessment of Time-Series Machine Learning Methods for Forecasting Hospital Discharge Volume // JAMA network open. 2018. 1, 7. e184087.

[58] *Meng Fanwen, Qi Jin, Zhang Meilin, Ang James, Chu Singfat, Sim Melvyn.* A robust optimization model for managing elective admission in a public hospital // Operations research. 2015. 63, 6. 1452–1467.

[59] *Miotto Riccardo, Li Li, Kidd Brian A., Dudley Joel T.* Deep patient: an unsupervised representation to predict the future of patients from the electronic health records // Scientific reports. 2016. 6. 26094.

[60] *Natarajan Balas K.* Sparse approximate solutions to linear systems // SIAM Journal on Computing. 1995. 24, 2. 227–234.

[61] *Nguyen Phuoc, Tran Truyen, Wickramasinghe Nilmini, Venkatesh Svetha.* Deepr: A Convolutional Net for Medical Records // IEEE Journal of Biomedical and Health Informatics. 2017. 21, 1. 22–30.

[62] *Oliveira Ester Góes, Garcia Paulo Carlos, Citolino Filho Clairton Marcos, de Souza Nogueira Lilia.* The influence of delayed admission to intensive care unit on mortality and nursing workload: a cohort study // Nursing in Critical Care. 2019. 24, 6. 381–386.

[63] *Padberg Manfred, Rinaldi Giovanni.* Branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems // SIAM Review. 1991. 33, 1. 60–100.

[64] *Pilanci Mert, Wainwright Martin J., El Ghaoui Laurent.* Sparse learning via Boolean relaxations // Mathematical Programming. 2015. 151, 1. 63–87.

[65] *Rajkomar Alvin, Oren Eyal, Chen Kai, Dai Andrew M., Hajaj Nissan, Hardt Michaela, Liu Peter J., Liu Xiaobing, Marcus Jake, Sun Mimi, Sundberg Patrik, Yee Hector, Zhang Kun, Zhang Yi, Flores Gerardo, Duggan Gavin E., Irvine Jamie, Le Quoc, Litsch Kurt, Mossin Alexander, Tansuwan Justin, Wang De, Wexler James, Wilson Jimbo, Ludwig Dana, Volchenboum Samuel L., Chou Katherine, Pearson Michael, Madabushi Srinivasan, Shah Nigam H., Butte Atul J., Howell Michael D., Cui Claire, Corrado Greg S., Dean Jeffrey.* Scalable and accurate deep learning with electronic health records // npj Digital Medicine. dec 2018. 1, 1. 1–10.

[66] *Rigollet Philippe, Tsybakov Alexandre.* Estimation of covariance matrices under sparsity constraints // arXiv preprint arXiv:1205.1210. 2012.

[67] *Rutherford P. A., Provost L. P., Kotagal U. R., Luther K., Anderson A.* Achieving Hospital-wide Patient Flow. IHI White Paper. // Surface Science. 2017. 360, 1-3. 21–30.

[68] *Safavi Kyan C., Khaniyev Taghi, Copenhaver Martin, Seelen Mark, Zenteno Langle Ana Cecilia, Zanger Jonathan, Daily Bethany, Levi Retsef, Dunn Peter.* Development and Validation of a Machine Learning Model to Aid Discharge Processes for Inpatient Surgical Care // JAMA network open. dec 2019. 2, 12. e1917221.

[69] *Shi Pengyi, Chou Mabel C., Dai J. G., Ding Ding, Sim Joe.* Models and insights for hospital inpatient operations: Time-dependent ED boarding time // Management Science. 2016. 62, 1. 1–28.

[70] *Song Hummy, Tucker Anita, Graue Ryan, Moravick Sarah, Yang Julius.* Capacity Pooling in Hospitals: The Hidden Consequences of Off-Service Placement // SSRN Electronic Journal. 2018.

[71] *Sotirov Renata.* SDP relaxations for some combinatorial optimization problems // International Series in Operations Research and Management Science. 166. 2012. 795–819.

[72] *Stowell Andrew, Claret Pierre Geraud, Sebbane Mustapha, Bobbia Xavier, Boyard Charlotte, Genre Grandpierre Romain, Moreau Alexandre, de La Coussaye Jean Emmanuel.* Hospital out-lying through lack of beds and its impact on care and patient outcome // Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2013. 21, 1. 17.

[73] *Stretch Robert, Della Penna Nicolás, Celi Leo Anthony, Landon Bruce E.* Effect of Boarding on Mortality in ICUs // Critical care medicine. 2018. 46, 4. 525–531.

[74] *Strubell Emma, Ganesh Ananya, McCallum Andrew.* Energy and Policy Considerations for Deep Learning in NLP // arXiv preprint arXiv:1906.02243. 2019.

[75] *Stubbs Robert A., Mehrotra Sanjay.* A branch-and-cut method for 0-1 mixed convex programming // Mathematical Programming, Series B. 1999. 86, 3. 515–532.

[76] *Su Weijie, Bogdan Malgorzata, Candès Emmanuel.* False discoveries occur early on the lasso path // Annals of Statistics. 2017. 45, 5. 2133–2150.

[77] *Tabak Ying P., Sun Xiaowu, Nunez Carlos M., Johannes Richard S.* Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS) // Journal of the American Medical Informatics Association. may 2014. 21, 3. 455–463.

[78] *The Economist* . How hospitals could be rebuilt, better than before // The Economist. apr 2017.

[79] *Thompson Steven, Nunez Manuel, Garfinkel Robert, Dean Matthew D.* Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges // Operations Research. 2009. 57, 2. 261–273.

[80] *Tibshirani Robert.* Regression Shrinkage and Selection Via the Lasso // Journal of the Royal Statistical Society: Series B (Methodological). 1996. 58, 1. 267–288.

[81] *Tokuda Tomoki, Goodrich Ben, Van Mechelen Iven, Gelman Andrew, Tuerlinckx Francis.* Visualizing distributions of covariance matrices // Columbia Univ., New York, USA, Tech. Rep. 2011. 1–30.

[82] *Wainwright Martin J.* Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting // IEEE Transactions on Information Theory. 2009. 55, 12. 5728–5741.

[83] *Wainwright Martin J.* Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell 1$-constrained quadratic programming (Lasso) // IEEE Transactions on Information Theory. 2009. 55, 5. 2183–2202.

[84] *Walraven Carl van, Forster Alan J.* The TEND (Tomorrow's expected number of discharges) model accurately predicted the number of patients who were discharged from the hospital the next day // Journal of Hospital Medicine. 2018. 13, 3. 158–163.

[85] *Wang Wei, Wainwrigh Martin J., Ramchandran Kannan.* Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices // IEEE Transactions on Information Theory. 2010. 56, 6. 2967–2979.

[86] *Yuan Ming, Lin Yi.* Model selection and estimation in the Gaussian graphical model // Biometrika. 2007. 94, 1. 19–35.

[87] *Zhang Cun Hui.* Nearly unbiased variable selection under minimax concave penalty // Annals of Statistics. 2010. 38, 2. 894–942.

[88] *Zhu Ting, Luo Li, Zhang Xinli, Shi Yingkang, Shen Wenwu.* Time-Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients // IEEE Journal of Biomedical and Health Informatics. 2017. 21, 2. 515–526.