PARAMETER INVARIANT STATISTICS AND THEIR APPLICATION TO CLINICAL DECISION SUPPORT

Alexander Steven Roederer

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation	Co-Supervisor of Dissertation
Insup Lee	C. William Hanson III
Professor, Computer and Information Science	Professor, Anesthesiology and Critical Care

Graduate Group Chairperson

Lyle Ungar, Professor, Computer and Information Science

Dissertation Committee:

Camillo J. Taylor, Professor, Computer and Information Science

Lyle Ungar, Professor, Computer and Information Science

Yoseph Barash, Professor, Computer and Information Science

Christopher Yang, Associate Professor, Information Science, Drexel University

PARAMETER INVARIANT STATISTICS AND THEIR APPLICATION TO CLINICAL DECISION SUPPORT

For my friends, who fill me with determination.

Acknowledgements

Though one person is listed as author of this document, no thesis is truly the work of a single person; I am indebted to many people who have supported me during my graduate career.

First, I wish to thank my advisor, Insup Lee, for taking a chance on me when I was fresh out of undergraduate school. Your support enabled this thesis to exist, and has changed my life immeasurably for the better. I cannot thank you enough. I would also like to thank my co-advisor, Bill Hanson, whose wisdom and medical expertise were invaluable to me. I am grateful to CJ Taylor, the chair of my committee, and the rest of my committee—Lyle Ungar, Yoseph Barash, and Christopher Yang—for their support and helpful advice in honing the focus of this document. I am also greatly indebted to James Weimer, who served as the inspiration for this thesis, and whose guidance was indispensable when I felt unsure how to proceed along the way.

I am forever grateful to my family: my parents, Estela and Steve; my sister, Adrianna; and my grandparents, Mima and Pipe. From a young age they instilled in me the ambition, drive, and love of knowledge that made me who I am today.

During the course of my PhD I have been lucky enough to have met many amazing people who I now consider not just colleagues, but friends. Thanks to Andrew King for being my workout partner and for being a constant source of ideas, some of them good,

all of them interesting. Thanks to David Weiss and to Jennifer Gillenwater, who were both generous enough to share their profound machine learning wisdom whenever I would pop into their cubicles with questions. Much thanks to Katie Gibson for being a source of levity during late nights at the office when I needed it, which was frequently.

To my friend Laura: my deepest gratitude for always keeping a watchful eye on me, making sure I did not forget to eat, and for reaching out on one of the most trying nights of my life. You really are the Remus Lupin in my life. To my friend, Christa: I cherish your constant friendship, your sage wisdom, and the jokes we craft together. Thank you for always being there to lend an ear when I needed it, and for being my confidant when I could turn to no one else. I cannot express how important you are to me.

Finally, to Colin Fanning: you have become an integral part of my life and a constant source of joy. Your love and support these last few years has helped me get through the most pear-shaped of times. Thank you. 

Statisticians have a different version of
The Boy Who Cried Wolf.

Saturday Morning Breakfast Cereal, by Zach Weinersmith, July 9th, 2016. Used with permission.

ABSTRACT

PARAMETER INVARIANT STATISTICS AND THEIR APPLICATION TO CLINICAL DECISION SUPPORT

Alexander Steven Roederer

Insup Lee

C. William Hanson III

The proliferation of digital medical device technology in the modern hospital has led to an explosion in the amount of available patient data. This data deluge presents opportunities for improved patient care, but in the current clinical environment it also creates significant challenges. One approach to alleviating the data deluge burden has been to develop clinical decision support systems, which automatically analyze data and provide the results to clinicians.

With an eye to improving clinical decision support systems, in this thesis we describe parameter invariant classification, a technique for detecting changes in patient state. The technique requires little data, is resilient to interpatient variability and noise, and can achieve a good detection rate while guaranteeing a constant rate of false positives over each individual in the population. It does so by avoiding direct estimation of a patient's parameters; instead, it considers two time-invariant linear systems (each describing a possible patient state) and selects which is more likely to have generated the data at any given time. This selection process is designed to be invariant to nuisance parameters— parameters that do not help to distinguish between the candidate systems.

The predictability advantages of parameter invariance come at a cost: the average performance of a parameter invariant classifier is usually lower than that of other classifiers. One reason is any parameters left out of the nuisance parameter set with a low signal-to-noise ratio decrease the accuracy of the classifier. To remedy this, we describe an algorithm

in which parameter invariant statistics are calculated over many possible feature subspaces. Then, feature selection is used to choose the combination of statistics which achieves the best performance, selecting for subspaces with high signal-to-noise ratios. We prove that a classifier trained over these features performs no worse than a parameter invariant classifier created from a single statistic.

Finally, we apply parameter invariant classification to a number of challenging problems in the medical domain: detection of hypovolemia in critically ill patients, automated meal detection in diabetic patients, and detection of pulmonary shunts in infants. We demonstrate promising performance in each scenario.

Contents

List of Tables

List of Illustrations

Chapter 1

Introduction

This work is motivated by the challenge of monitoring the clinical health of the critically ill patient. In it, we describe how clinical decision support systems can aid clinicians in using large amounts of available patient data to improve care, and how their usefulness is sometimes limited by the large number of false positives they produce. To address this, we describe parameter invariant statistics, statistics designed to be invariant to specific groups of nuisance transformations in the parameters of a model. We show how these statistics can be used to perform detection that bounds the false positive rate over all individuals in a population. We posit how machine learning techniques can be used to boost detectors' performance while preserving the bound on the false positive rate, and apply the proposed techniques to a number of different real-world applications from critical care monitoring.

During the course of a day, medical professionals are constantly making decisions about the care of their patients. These decisions are difficult, in large part, because they are made in the face of uncertainty about what is actually going on in a patient's body. While various *medical devices* have long been used to diagnose or treat disease, recent advances in digital technologies have led to an explosion in the number and kind of medical devices

available. These medical devices are collecting unprecedented amounts of digital data on many different aspects of patient physiology. (Frost & Sullivan (2012) and Manyika et al. (2011) both describe this shift toward "big data" in medicine.) The hope has been that this data could be harnessed by clinicians to help reduce uncertainty in decision making, thereby improving quality of care and lowering costs (Raghupathi and Raghupathi, 2014).

While there is widespread consensus that rapid digitization of huge amounts of patient data presents numerous potential opportunities, it also creates major challenges. (There are dozens of published reports detailing these opportunities and challenges. See, for example, Berner (2009); Celi et al. (2010); Groves et al. (2013); Manyika et al. (2011). The report by Frost & Sullivan (2012) states this particularly clearly:

> While data is being hailed as the key to improving health outcomes and reducing healthcare costs, the sheer volume of data is so overwhelming that most organizations are unable to take full advantage of it with their current resources. Managing and harnessing the analytical power of these large datasets, however, is vital to the success of all healthcare organizations.

While the amount of available patient data increased dramatically over the past several decades, the way in which clinicians interact with this data has failed to change significantly. The deluge of medical data quickly made the limitations of paper-based information management clear (Chaudhry et al., 2006), and this coupled with mandatory government requirements lead to a rapid shift toward storing patient data digitally in *electronic medical records* (EMRs) (Raghupathi and Raghupathi, 2014).[1] While EMRs have become ubiquitous within the United States, they serve primarily to *store* patient data; it is usually difficult or impossible to access, visualize, or analyze data from within the EMR system. Without

[1] Electronic medical records are also called electronic health records (EHRs). The terms are used interchangeably in the literature.

additional technology, EMRs are "essentially just copies of paper-based records stored in electronic form" (Bennett and Doub, 2011), but containing an unwieldy amount of data.

Clinicians today are not only responsible for physical care of their patients, but are also responsible for coordinating and comprehending patient data over time and among multiple providers and settings, which requires processing and managing vast amounts of information (Peleg and Tu, 2006). It is unsurprising, then, that studies have revealed large numbers of preventable medical errors stemming from bad data, and sizable gaps between best practices and actual practices (Committee On Quality Of Healthcare In America, 2001; Kohn et al., 2000); clinicians are now expected to manage, validate, and process larger and larger amounts of patient information, in addition to their existing duties.

One approach to alleviating the burden of the data deluge has been to develop computerized *clinical decision support (CDS) systems*: computer-based tools designed to automatically analyze data and provide the results of this analysis to clinicians. For example, CDS systems can alert clinicians when the patient requires attention, automatically check for conflicting medications, or provide pertinent summaries of a patient's course of stay. More advanced CDS systems can go beyond simple visualization or rudimentary checks on data. They can apply state-of-the-art machine learning techniques to patient data in real time, performing analyses which would be impossible for clinicians to perform by hand. (See Greenes (2014); Osheroff et al. (2005); Rinott et al. (2012) for detailed definitions of clinical decision support systems.) Equipped with the results of these analyses, clinicians can be alleviated of the burden of data management while gaining the ability to make more informed decisions about the care of their patients.

When designed, implemented, and deployed in areas where they are needed, use of even rudimentary CDS systems has had dramatic, positive results (Hunt et al., 1998). Garg et al. (2005) did a systematic review on controlled trials testing effects of computerized clinical decision support systems, and showed that 64 percent of the CDS systems studied im-

proved practitioner performance.[2] A systematic review by Chaudhry et al. (2006) showed improvements in quality and efficiency of care at four benchmark institutions through the use of "health information technology" (a wider category of technological solutions that includes CDS systems). A more recent study by Kilsdonk et al. (2013) showed that a user-centered CDS system could better support healthcare practitioners than an equivalent paper-based guideline. And a meta-analysis of CDS randomized control trials on alerts and reminders by Trowbridge and Weingarten (2001) showed they can alter clinician decision making, reduce medication errors, and promote preventive screening. The true goal of CDS systems is not just to duplicate current practice, but to equip clinicians to harness medical data in novel and innovative ways.

1.1 Challenges in Developing Clinical Decision Support Systems

Though some simple CDS systems have become widespread in clinical practice, more complex systems which take advantage of signal processing or machine learning techniques are rare, and even more rarely see widespread use. There are a number of reasons for this.

First, the amount of good medical data available to learn from is usually very small. Though their numbers are growing, large databases of collected retrospective medical data are still rare, as routinely capturing and storing patient data digitally is very new to medicine. Data that does exist is often very noisy due to motion artifacts or poor electrical connections, and treatments administered by clinicians may further alter the signal. If a patient has a unique combination of conditions (*comorbidities*) or suffers from a particularly

[2]Though only 13 percent reported improved patient outcomes, improved practitioner performance may lead to safer, more effective care in a period of time too long to capture by a short trial study.

rare condition, even similar patient data may not exist. Patient physiologic signals violate many simplifying assumptions often made of data under analysis. All measurements taken from a single patient are correlated temporally, and many popular traditional learning techniques are not directly suitable for such data.

Many traditional learning techniques assume individuals are drawn from a single distribution with fixed parameters, but human physiological dynamics are not uniform over a population of patients. Each new individual who enters a hospital system has a unique set of underlying parameters that dictate their physiologic behavior. This intersubject variability violates many traditional machine learning assumptions. Some popular machine learning techniques (such as online learning) can be used to tune model performance over a small amount of an individual's data in order to boost performance on later data from that individual, in effect attempting to create a patient-specific model. However, in a medical setting, a CDS must be able to produce consistent, useful results as soon as they are needed. Systems cannot wait for patients to produce data to learn from; by the time enough data has been collected for training, the patient may have suffered injury or death. Thus, many current approaches to creating models that perform well over every individual are inadequate for clinical care.

Finally, because CDS systems are intended for use in a clinical environment, extremely rigorous evaluation of their effectiveness is vital, sometimes even requiring clinical trials. System performance must be very high over all individuals; for CDS system results to be useful, they must not only be accurate (as failure might lead to injury or the death of a patient) but also precise; a high rate of false positives will fatigue caretakers and cause them to ignore the system's recommendations over time (Kesselheim et al., 2011). Trust in the results of the system has been found to be a major determinant in the success of CDS systems (Alexander, 2006). Thus, the results of the analysis must engender the trust of clinicians by producing consistent, predictable results over all patients, and by being

transparent. As patient status can change quickly, results must also be rendered in a timely fashion to be useful. Most current approaches fail to meet these criteria.

Clinical decision support systems that do not address these challenges risk making the data deluge problem worse by becoming another source of information that clinicians ignore. Indeed, one of the most pressing issues for the simple CDS systems in widespread use today, such as bedside threshold alarm systems, drug-drug interaction alert systems, and computerized physician order entry systems, is their huge number of false positives. (For examples, see Clark et al. (2006); Koppel et al. (2005); Weingart et al. (2003), respectively.) Extremely high rates of false positives fatigue and burden caretakers and lead to the systems being ignored, considered nuisances, or turned off entirely.Without any predictable guarantee as to the overall performance of a CDS system, and with the tendency for such systems to perform extremely poorly on challenging patients, clinicians' trust is easily lost.

1.2 Proposed Approach

The ideal analysis technique for use in CDS systems would learn from small amounts of patient data, be resilient to interpatient variability and noise, and would achieve high accuracy while *limiting the number of false positive alarms produced over each patient*. In this thesis, we describe such an analysis technique. We begin by describing how to calculate statistics based on likelihood ratios, which are invariant to interpatient variability. We then describe how these *parameter invariant statistics* can be used to construct *parameter invariant classifiers* which can detect changes in a patient's current state using only a small amount of data, even in scenarios where interpatient variability is high. This is achieved by avoiding direct estimation of a patient's parameters. Instead, a parameter invariant classifier considers two classes of time-invariant linear systems, each of which is hypothesized to describe one of two possible patient states, and selects which of these two model classes is

more likely to have generated the data at any given moment in time. This selection process is designed to be invariant to *nuisance parameters*—parameters that, based on the data, do not help to distinguish between the two candidate classes. In doing so, this method can *guarantee a constant rate of false positive detections across all individuals in the population*.

Figure 1.1 illustrates the constant false positive rate property. A standard classifier (represented in orange in the figure) may achieve good average performance (the solid orange line), but individuals (represented by orange X markers) may each achieve different true and false positive rates. While these values can be adjusted based on how the classifier's parameters are tuned (moving the X markers along the dotted lines), individual performance will still vary across the population. A classifier with the constant false positive rate property (represented in purple in the figure) guarantees that for all individuals in the population (represented by purple asterisks) and for all choices of parameters, the false positive rate will be constant over all individuals. Tuning the parameters of such a classifier can change the false positive rate and true positive rate (the shaded purple area), but the false positive rate of all individuals will change together.

The predictability advantages of parameter invariance come at a cost, however. The average performance of a parameter invariant classifier is usually lower than that of a non-parameter invariant classifier. One reason for this: any parameters not included in the nuisance parameter set that contain a low signal-to-noise ratio will decrease the accuracy of the statistic. To remedy this issue, this thesis describes an algorithm in which "two-sided" test statistics are calculated over a number of possible feature subspaces. Then, feature selection is used to choose the combination of statistics which achieve the best performance, in essence selecting for subspaces with high signal-to-noise ratios. We prove that a classifier trained over these features achieves a true positive rate no worse than a parameter invariant classifier created from a single statistic. This work proposes applying

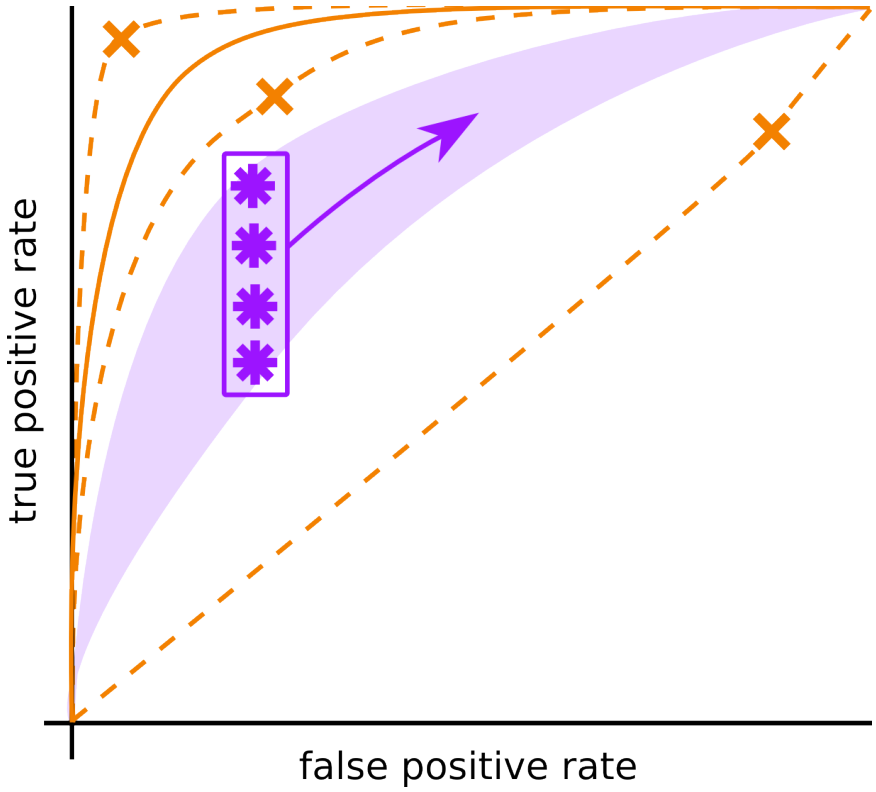


Figure 1.1: A theoretical illustration of the constant false alarm rate principle using a receiver operator curve. Traditional classifier performance is shown in orange; while average population-level performance (the solid line) may seem good, each individual may have a completely different true positive/false positive trade-off rate. A classifier with a constant false alarm rate is shown in purple; here, all individuals are guaranteed to have a fixed rate of false positive, while true positive rate varies.

a number of existing feature selection techniques to these features, and explores how the statistic's properties are preserved through the selection technique.

Finally, we apply parameter invariant classification to a synthetic data set (to clearly illustrate performance results) and to a number of challenging problems in the medical do-

main, including work on detection of hypovolemia in critically ill patients (an extension of the work in Roederer et al. (2015)), automated meal detection in diabetic patients (extending work in Chen et al. (2015b)), and detection of pulmonary shunts in infants (an extension of the work in Ivanov et al. (2015)).

The concept of parameter invariance itself is not novel; it originated in work by the Doppler signal detection community. In Scharf and Friedlander (1994) the methodology behind parameter invariant statistics is described, and further works such as Bon et al. (2008); Kraut and Scharf (1999) have built upon it. However, this work presents a novel generalization of parameter invariance—the parameter invariant statistic—and incorporates it into a machine learning framework and a new feature generation/selection algorithm. Our numerous applications of the technique to a number of medical domains are likewise novel.

The primary contribution of this thesis is to fully develop parameter invariant statistics for use in event detection in the medical domain. We begin with the simplest formulation, and show how concepts from the machine learning literature can be incorporated to address some of the shortcomings of the technique. We also include in the work a number of medical applications.

1.3 Contributions

We summarize the contributions of this work as follows:

- **Parameter Invariant Classifiers.** We introduce parameter invariant statistics from the signal processing literature to a machine learning framework, resulting in parameter invariant classifiers. We show how these classifiers satisfy our constant false alarm rate performance metric, while more traditional machine learning classifiers do not.

- **Parameter Invariant Features.** It is often difficult to determine *a priori* which parameters are nuisance parameters, but selecting the wrong subset of parameters as nuisance parameters can lead to suboptimal classification. To remedy this, we define a novel set of features based on parameter invariant statistics. We describe the challenges associated with identifying the best feature in such a large feature set. To address this, we propose a greedy feature selection heuristic. We describe the asymptotic properties of learning over such a feature set, and we test these features over a synthetic dataset to demonstrate improvement in performance.

- **Applying Parameter Invariance to Healthcare.** We apply parameter invariant classifiers to three unique medical classification problems. For some of these problems, traditional techniques exist that maximize performance but produce large numbers of false positives for some individuals in the population. We show that parameter invariant classification provides good performance while maintaining a fixed rate of false positives over all individuals. We further show that in each case, using the proposed parameter invariant feature set further improves detection rate while maintaining the fixed false positive rate property.

1.4 Thesis Overview

The organization of the thesis is as follows:

- **Chapter 2 - Related Work.** In this chapter, we explore related work that attempts to address the challenge of identifying changes in data over time. We include selected examples of algorithms that attempt to learn full system parameters, and selected examples of algorithms which do not. Algorithms that attempt to learn full system parameters include those that learn them from data and those that use medical knowl-

edge to directly model physiology. Algorithms that do not learn full system parameters include discriminative classifiers and unsupervised/semisupervised clustering methods.

- **Chapter 3 - Classification Via Maximum Likelihood and Maximal Invariance.** In this chapter, we present a precise definition of the problem we wish to solve: how to achieve a constant false positive classification rate when a population is made up of individuals whose parameters vary. We compare and contrast the conventional maximum likelihood and maximally invariant classifier approaches to the problem, and describe why they are sometimes insufficient.

- **Chapter 4 - Parameter Invariant Classification.** This chapter defines parameter invariance and parameter invariant statistics, and shows how they address the insufficiencies highlighted in the preceding chapter. This chapter describes the generalized parameter invariant statistic that can be calculated for any linear time-invariant system over a specified set of nuisance transformations. We show how a parameter invariant statistic can be used for *detection*. We provide proofs that using these statistics for detection provides an asymptotically bounded false positive rate. We also describe the "two-sided" test extension that allows for lower bounds on detection performance.

- **Chapter 5 - Parameter Invariant Subspace Selection.** This chapter focuses on boosting the true positive rate of a parameter invariant detector. We describe an algorithm to generate a number of parameter invariant statistics over all subspaces of the original problem and, treating these statistics as features, select the subset of them that produces the best performance. We prove that using this subspace feature set will produce a classifier no worse than using only the original parameter invariant

statistic over the full space. We propose improving the runtime of this algorithm by using a greedy feature selection technique.

- **Chapter 6 - Applications.** We apply both the original parameter invariant detector and the subspace feature set parameter invariant detector to both synthetic data and a number of clinical applications. We show that in all cases, maximum likelihood methods have false positive rates that vary wildly across the population, while the parameter invariant detector provides near-constant false positive rates over all patients, with modest decreases in detection rate when both are tuned to achieve the same average rate of false positive. We also show that the proposed expanded feature set boosts detection performance while maintaining the desired constant false positive rate.

- **Chapter 7 - Conclusion and Future Work.** We close with a review of the work, and a discussion of possible future extensions.

Chapter 2

Related Work

At a very abstract level, patients can be thought of as systems operating in one of two modes: a "healthy" mode, which describes normal operation, or an "illness" mode, which describes unusual, undesirable, or unexpected operation. The goal of clinical monitoring is to ascertain which mode the patient is currently in at any given moment, and predict if and when a change in modes will occur. A number of different techniques exist in the literature to solve problems of this type. In this section, we provide a high-level overview of such techniques, which can be broadly categorized based on whether they assume the existence of an accurate *first-principles model* of the underlying system, and whether they assume the existence of good quality *training data* for the patient. This categorization is illustrated in Figure 2.1.

When a highly accurate model of the underlying physical system is available, model-adaptive control techniques can detect mode switches, even when little training data is available (the upper left quadrant of Figure 2.1). If an accurate model is available and training data is plentiful and of high quality, model-adaptive control allows for further refinement, resulting in highly accurate, robust systems (the upper right quadrant). When no

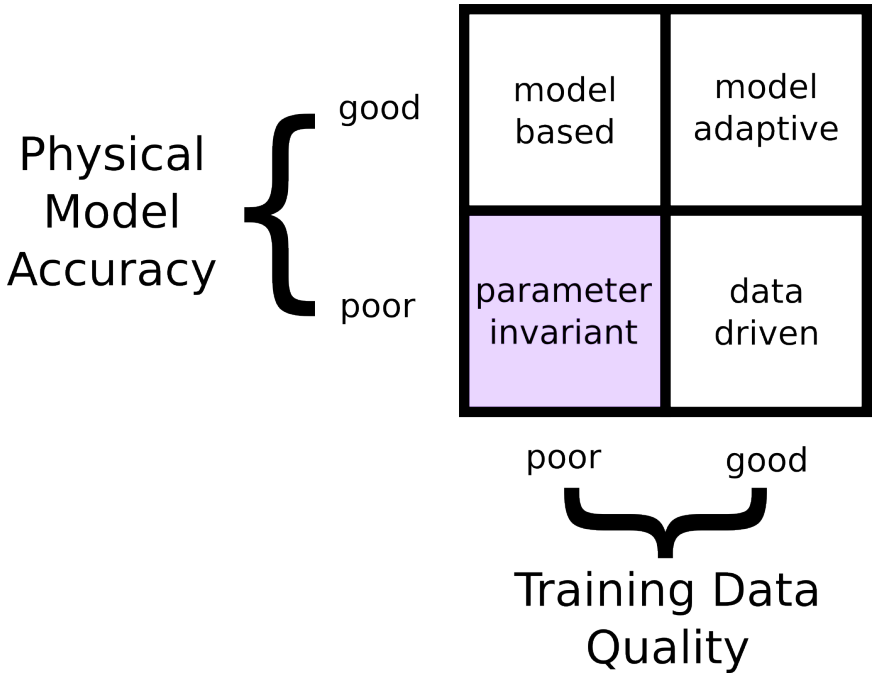


Figure 2.1: A conceptual diagram illustrating categorization of existing techniques for classifying which mode a system is in. The vertical axis distinguishes techniques based on whether they rely on use of a physically accurate model. The horizontal axis distinguishes techniques based on whether they rely on a good quality training data set.

specific model of the underlying physical system is available, but training data is plentiful, data-driven machine learning techniques can fit generic models to the data (the lower right quadrant).[3]

In all cases, developed models have parameters which must be selected. Usually these parameters are learned from some quantity of existing population-level training data. How-

[3]In reality, techniques do not fall strictly into one quadrant or another; they exist on a spectrum based on the degree to which the model is hand-specified based on some physical understanding of the world, and the richness of the training data necessary to achieve an accurate model.

ever, model parameters learned over a population may not perform well over all individuals in that population, especially if intra-individual variance is high. Some effort has been made in adapting models to be more patient specific, but such efforts usually require considerable amounts of training data to be available for each individual, which may not be feasible in a medical setting.

Few techniques exist in the lower left quadrant, suitable for situations with both poor models and poor quality of training data, and still fewer techniques provide any sort of performance guarantees over individuals without requiring considerable amounts of training data over each of them. The remainder of this document seeks to define such a technique.

In this chapter, we review existing work in model-based/model-adaptive control systems, as well as data-driven model development through time-series analysis and machine learning techniques. We describe theoretical foundations for creating models, and implementations of those models in the medical domain. In all cases, we focus on how others have approached the problem of detecting or predicting changes in a system's underlying mode. Finally, we describe techniques for adapting models to be patient-specific. This chapter is not intended to be an exhaustive list of all relevant techniques. Instead, we seek to highlight a number of popular approaches from each category, in order to underscore how the work of this thesis fills a gap in the broader signal processing and machine learning context.

2.1 Creating and Using Models of Physical Systems

In this section, we consider techniques which fall in the upper half of the table depicted in Figure 2.1; that is, we consider the construction and use of high quality time-series models of physical systems, both when training data is of poor quality, and when it is of high quality. In either case, model parameters must be estimated in order to capture the

behavior of the underlying system generating the observed data; we discuss challenges in this estimation process.

In particular we will focus on work that seeks to create models based on human physiology. Theoretically, a complete, perfect physiological model could exactly determine a patient's current mode. However, such a model would be intractably complex, so instead, an approximate model of some kind is used. The level of approximation necessary for a model to be practical varies based on application. Sometimes these models still attempt to capture the fundamental nature of physiology and biology in some way, with the assumption that such models achieve better performance with less effort than would be required to fit a generic, non-physiologically motivated model. In all cases, however, once a model has been fit, the model itself should contain some portion which specifies which mode the model, and thus the patient, is likely to be in based on current input and output observations.

2.1.1 Creating Models from Physiology

There have been extensive efforts to capture the behavior of human physiology in terms of mathematical models for at least a century. The first major modern texts on the subject, spurred by developments in systems engineering and control theory in the 1950s and 1960s, included those by Grodins (1963); Milhorn (1966); Milsum (1966). In the past several decades, a dramatically improved ability for researchers to gather physiologic data at unprecedented scale has led to a rapid acceleration of the field (Ottesen et al., 2004).

In the control theory literature, creating a model that conforms to observed data is called *system identification*. System identification techniques can be broadly classified as white-box (*i.e.,* techniques that attempt to derive their model from first principles, such as directly mimicking physiology, as described in Section 2.1.1), black-box (*i.e.,* techniques that are purely data-driven, see Juditsky et al. (1995); Sjöberg et al. (1995); Suykens and Vande-

walle (2012) for surveys and Section 2.2 for further discussion), or gray-box (*i.e.,* techniques that mix elements of white-box and black-box approaches; see Bohlin and Graebe (1995); Kristensen et al. (2004); Tulleken (1993)). System identification is a well-studied area with major historical work; Aström and Eykhoff (1971) provides a thorough survey for "identifying" models for both linear and non-linear systems. More recent work by Ljung and Gunnarsson (1990) describes how system identification can use tracking to identify time-varying dynamics.

Modeling of physiological systems is often performed using multi-scale compartmental modeling (Cobelli and Carson, 2008). Compartmental models consist of physiological compartments (*e.g.,* organ, blood, cell) where each compartment contains a state that interacts with neighboring compartments through differential or time-differenced dynamics (*i.e.,* a state-space model). A fundamental challenge of compartmental modeling is balancing real-world accuracy and model observability. High-fidelity models may accurately capture patient physiology, but their parameters may not be observable through available training data.[4] Additionally, highly accurate models can be difficult to develop, requiring extensive knowledge of human physiology.

For these reasons, most physiologically informed models used in practice are simplified for usability. However, even these approximations can be difficult to work with, as their form may be unusual. While some models (such as the artificial pancreas models) can be constructed with understood input/output behaviors, for more complex systems, model inputs may be unclear, and models with access only to output data are often less accurate. Also, it is rare that model parameters are selected from expert knowledge; usually, some amount of patient data is required to determine parameter values. Complex models can

[4]Consider, for example, a model of the heart that incorporates components representing cells and their functions. As current medical technologies cannot observe the behavior and interactions of individual cells in a living patient, there would be no way to calculate the parameters of these components.

feature large numbers of parameters, which must be tuned using large amounts of data, a further benefit to using a simplified model.

A complete investigation of the field is beyond the scope of this work. Instead, we briefly discuss some prominent examples of physiologic first-principles models that have seen successful use.

2.1.2 Prominent Physiologic Models from the Literature

One of the most successful physiological modeling efforts has been around the cardiac pacemaker (Bogdan et al., 2012; Mond and Proclemer, 2011; Noble, 1984, 2002), with the safety of many pacemakers currently on the market established through testing with detailed electrocardiovascular models. A particularly effective mechanism for testing involves modeling contractions of the chambers of the heart using timed automata (Jiang et al., 2012). There have also been successful efforts to create detailed models of not just the electrophysiology of the heart, but the entire heart itself (Trayanova, 2011).

Similarly, successful efforts to develop an "artificial pancreas" (Albisser et al., 1974; Bequette, 2005; Cobelli et al., 2011) lead to the proven ability to control diabetes through closed-loop administration of insulin in both Type 1 and Type 2 diabetics (Breton et al., 2012; Magni et al., 2007; Weinzimer et al., 2008). This has led to the modern proliferation of wearable continuous glucose pumps. Both simple linear pancreas models and more complex non-linear versions (such as the Bergman model, described in Bergman et al. (1979), and T1DMS (Dalla Man et al., 2014; The Epsilon Group, 2016)) have seen use in validation and testing.

2.2 Learning Models Directly from Data

In this section, we consider techniques which fall in the lower half of the table depicted in Figure 2.1; that is, techniques which eschew use of an "accurate" first-principles physiological model, and instead attempt to learn a useful model directly from observed data. These techniques can come from a number of different domains, including signal processing, time series analysis, machine learning. When sufficient training data is available, such techniques can produce models more easily than first-principles modeling, with models achieving good performance while being robust to noise.

First, we describe techniques from the control theory and signal processing literature which fit generic time series models to data, then use information about the model's fit to determine the patient's mode. Next, we touch on generative models, which eschew temporality and instead specify a joint probability distribution over observations and modes. Finally, we briefly mention the extensive work that exists in discriminative classification, which seeks only to produce probabilities for each mode conditional on the observations.

2.2.1 Generic Modeling and Control

Instead of attempting to construct detailed physical models, many works tailor more generic time-series models to observed physiologic behavior. Models constructed in this way often provide more straightforward parameter estimation and stronger theoretical guarantees about performance; depending on their simplicity, they may also require less patient data for parameter tuning. They are also usually less laborious to create; generating such a model is often simply a matter of solving some constrained optimization problem over model parameters. Young (1981) provides a survey of parameter estimation techniques in the signal processing literature before 1980, including both historical analog methods and more relevant digital methods. Hamilton (1995) provides a somewhat more recent treat-

ment, though focuses on econometrics. Vibergt (1995) presents an overview of subspace-based system identification methods, including QR and singular value decompositions. Work by Quinn et al. (2009) applies these concepts to physiologic signals, modeling infant ECG signals using ARIMA models, adding an "X-factor" term to capture otherwise unmodeled data variation among other innovations. They successfully use this model, in a factorial framework, to successfully detect changes in infants' condition.

While techniques in this area have similarities to the work presented in this thesis, performance guarantees are not provided on an individual level but are instead measured in a limiting case over the population. Widely-known work by Box and Jenkins (1994) introduced procedures for estimating the parameters of an autoregressive moving average model. These procedures achieve good fit with a modest amount of data, but require explicit retraining over each system they are applied to, and performance degrades as an individual's behavior diverges from the population. More complex general models such as neural networks have been used to learn general time series models, as in Tang et al. (1991), with mixed results. Learning the appropriate model parameters is difficult over training data drawn from a diverse set of systems, and for network-based models, determining the best topology for the network can be difficult. Reduced-order models developed with black-box techniques may be highly observable, and work best when data is plentiful. Unfortunately, in scenarios in which data is limited and variance between patients is very high, it can be difficult to learn precise parameters that generalize well.

2.2.2 Learning Generative Models from Data

When systems being identified are assumed to be linear systems, and system parameters are assumed to follow a Gaussian distribution, state estimators such as Hidden Markov Models (Rabiner, 1989) and Kalman filters (Kalman, 1960) have been shown to be optimal estima-

tors of an underlying autoregressive process. Since linear and Gaussian assumptions are common, such techniques have long been popular for identifying changes in a system's underlying state by estimating their parameters. For example, Zhu (2011) use hidden Markov models to detect anomalies in patients' blood glucose levels, and Smith and West (1983) apply multiprocess Kalman filters to monitor serum creatinine in renal transplant patients.

In the machine learning literature, hidden Markov models, Kalman filters, and dynamic Bayesian networks techniques have been generalized under the framework of graphical models; Koller et al. (2007) present an elegant and exhaustive description. Learning over these models has been well developed, especially for linear systems (see Jordan et al. (1999) for a tutorial introduction). In the machine learning community, these (and similar) models are described as *generative*, as they provide a full probabilistic model of all variables in a system.

Many older generative modeling techniques tend to assume data is generated from a discrete, fixed set of classes that do not change over time, an assumption often violated in medical data. More recent advances in generative models have attempted to address this issue with the introduction of continuous state estimators for dynamical systems. An early forebear of this was the work of Carter and Kohn (1994), who use a Gibbs sampler to infer a linear state space model that is a mixture of normals and coefficients that can switch over time. More recent examples include the work of Kim and Pavlovic (2007) and work on continuous density hidden Markov models by Sha and Saul (2006). Hierarchical Bayesian modeling, developed by Kass and Steffey (1989) and Gelman (1995) has seen success in application to medical data. Fox et al. (2011, 2008) present a hierarchical Bayesian method for using hidden Markov models to learn switching linear dynamical system models. Saria et al. (2010) use hierarchical Dirichlet processes to define a time series topic model, and apply this work to personalized risk stratification on heart rate data from premature infants.

2.2.3 Avoiding Estimation of Model Parameters with Discriminative Classifiers

Techniques described in the previous section attempt to model the full joint probability distribution over the data; that is, $P(x, y)$. In contrast, there has been extensive work in determining which class y the data x belongs to, without attempting to model the underlying distribution. The simplest of these approaches involves extracting n characteristics from the data $\boldsymbol{X}$ of interest and fitting a simple model which *discriminates* the class $\boldsymbol{y}$ from these characteristics:

$$\boldsymbol{y} = \begin{bmatrix} 1 & x_{0,0} & \dots & x_{n,0} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{0,i} & \dots & x_{n,i} \end{bmatrix} \boldsymbol{w}.$$

The result is a probability distribution for $\boldsymbol{y}$ conditional on having observed some $\boldsymbol{X}$, rather than the full joint distribution from Section 2.2.2.

Discriminative Classifiers with Temporal Features

Discriminative classifiers consider discrete sets of features $\boldsymbol{X}_i$ with no inherent structure over these features or between feature sets. As such, the ability of the discriminant classifier to distinguish between classes comes in large part from the inclusion of the appropriate features. If different kinds of "systems" can only be distinguished by their behavior as it changes over time, as is the case when considering patient health, then this information must be extracted from the data and included in the features considered if the classifier is to achieve good performance. Studies have been done comparing generative and discriminative models, and despite their simplicity, discriminative models have been shown to perform well at detecting deterioration, as in Clifton et al. (2011) who compared performance of Gaussian mixture models to support vector machines in detecting deterioration

in vital signs.

Numerous methods of capturing temporal information by processing or filtering data over time have been proposed. Dietterich (2001) provide a survey of machine learning over sequential data. As Verduijn et al. (2007) discuss, feature extraction can be guided by domain knowledge or purely by available data. A popular way of capturing this information is by including Fourier or wavelet coefficients as features; Shoeb (2003) use such features for detecting patient-specific seizure onset. Slow feature analysis, another possible source of features, is made feasible for use in Zafeiriou et al. (2013).

Clustering Time Series

Efforts have also been made to classify time series through clustering. Such techniques are attractive because they do not require physiologic data to be labeled. As gold-standard labeled data is often scarce in medicine, this presents obvious benefits. Liao (2005) presents a thorough survey of time series clustering techniques. By defining similarity kernels between individual time series, new time series can be classified by looking for similar existing samples and grouping them. Similarities between time series can also be defined using extracted morphological features to represent the signal, as in Saria et al. (2011), which presents a medical detection problem application. Unsupervised learning has been successfully employed to learn over such feature spaces in the medical domain without patient models or parameters, as in Wiens and Guttag (2010), which describes accurate detection of arrhythmias in ECG signal.

2.2.4 Hybrid Generative-Discriminative Models

There have also been efforts in bridging the gap between generative and discriminative models. One approach is the use of kernels which provide the advantages of generative

models to discriminative classifiers. The Fisher kernel, developed by Jaakkola and Haussler (1999), is an early example. More recently, Moreno et al. (2003) developed an alternative to the Fisher kernel which establishes kernel distance based on Kullback-Leibler divergence, allowing them to naturally handle variable length sequence data. Jebara and Howard (2004) developed probability product kernels, which map data points in the input space to distributions over the sample space, and set the inner product to be the integral of the product of pairs of distributions. The underlying generative models considered by each of these approaches vary from kernel to kernel. These kernels, though promising, have seen limited application to medical data sets.

2.3 Creating Patient-Specific Models

Many of the techniques discussed in this chapter share a common assumption: that patients behave according to a shared distribution or a known, fixed set of distributions. However, patient data from different individuals often shows significant variability. Patients may have different comorbidities, different physiological profiles, different medical histories, and may be on different medications; two patients are rarely alike, even when they are suffering from the same primary affliction. White-box techniques require knowledge of the underlying model, and black-box techniques operate under the assumption that the model is fixed with parameters that can be learned over a sufficient quantity of training data. In medical contexts, however, models are difficult to construct, the underlying model for each patient considered may differ, and each subject in the training set may only have a single piece of training data associated with them (such as a single medical exam or hospital stay), creating a much more challenging scenario.

Approaches do exist in the literature to address these problems. More general, population level models may be adapted using information specific to an individual, or individuals

very similar to them, creating a more personalized model. In this section, we describe a select few of these approaches and their application to patient state detection.

If sufficient training data from an individual is available, machine learning approaches can be used to train a local model on that data. This local model can then be used directly. For example, Zhang and Szolovits (2008) develop a patient-specific alarm algorithm by training a neural network over eight hours of patient vital sign numerics. Aforementioned work by Wiens and Guttag (2010), "online" approach allowed unsupervised reclustering based on few initial heart beats of each patient, which significantly improved performance. Shoeb et al. (2004) detect onset of epileptic seizures by using wavelet decomposition on EEG signal to extract features, then classifying those features using support vector machines. Their technique uses a significant amount of training data from each patient but produces good results.

Local models may also be creatively combined with a more general model. In work by Visweswaran et al. (2010) and Visweswaran and Cooper (2010), the authors create Bayesian networks with local structure and use Bayesian model averaging to combine them with networks with global structure. They then use these models to predict sepsis and heart failure. Their local structure was based only on data routinely available at time of admission, which allows the model to be applied immediately on a new patient.

Local models may also be created by selecting training data from individuals similar to the individual in question, based on some similarity metric. Ng et al. (2015) do just this, using a locally supervised metric learning similarity measure, then training logistic regression models to create personalized risk profiles for diabetes onset.

Finally, newer work in the field of *transfer learning* attempts to bring more nuisance to the process of creating local models, breaking down "auxiliary" data sources (that is, the population-level data) into its information theoretic components, and combining this with a local model. Gong et al. (2015) apply this technique to develop risk models for cardiac

surgeries that are tailored to individual patients.

In all cases, these techniques require either an initial training phase over the individual (meaning they cannot be used in critical monitoring scenarios), must be trained as they go (using online learning, which may lead to poor initial performance), or they are limited to use only of information available on admission, such as demographics. In all cases, no guarantees on performance can be given, especially when physiological diversity is high. But robustness to physiological diversity, even in circumstances with limited data, is key to most effective clinical decision support (Lee et al., 2012; Lee and Sokolsky, 2010; Sokolsky et al., 2011). In the next chapter, we formally define this problem and provide a more thorough description of the inadequacy of current solutions, with an eye toward providing a more satisfactory solution.

Chapter 3

Classification Via Maximum Likelihood and Maximal Invariance

This chapter provides a precise formulation of the problem of interest: how to robustly classify systems when those systems are taken from two varying populations while limiting the number of false positive classifications. We establish the metrics we use to evaluate the performance of candidate solutions to the problem. We then describe two simple solutions: classification by maximum likelihood, and classification by testing against the likelihood ratio. These will serve as the basis for discussion and development in the rest of the thesis. We show how the likelihood ratio can be made invariant to common nuisance transformations on data, define maximal invariance, and show that, though desirable, maximal invariance is not always achievable.

Let us return to our previously defined abstraction of a patient as a complex system with two modes: a "healthy" mode that describes normal operation, and an "illness" mode that describes unusual, undesirable, or unexpected operation.[5] As we saw in Chapter 2, creat-

[5]The broad class of *cyber-physical* systems (systems with both "cyber" and "physical" components) are

ing a model to represent such a system is a common approach for determining which mode the system is in. However, determining exact parameters of the models in question can be difficult. Even when each individual system conforms to a simple model, the parameters of those models can vary widely over the population (*i.e.,* different patients' bodies behave differently in response to similar stimuli). In these scenarios, it is often infeasible to collect the large amount of data from each individual system required to accurately estimate its parameters. Under these combined challenges, well-established parameter estimation methods (like maximum likelihood) can produce models which generalize poorly. This can lead to massive false positive rates for certain individuals in the population, even for systems with good average performance, which is unacceptable in safety-critical scenarios.

In our abstract scenario, however, knowing the exact parameterization of the underlying system is not as important as knowing which mode the system is in. A medical alert system does not necessarily need to learn the full parameterization of a patient's physiological dynamics to determine if the patient is healthy or ill; there are likely sets of parameters which tend to characterize sick patients. However, these parameters may be mixed in with other parameters that do not aid in characterizing illness.

In this chapter, we begin by giving a precise mathematical formulation of our problem definition, which we frame as a hypothesis testing problem. Our performance metrics reflect this problem framing: we aim to maintain a specified upper bound on false positive rate over all individuals, while maximizing true positive rate for each individual over all possible parameter values. Such a test is defined as *uniformly most powerful*. We provide

often thought of in this way. For example, automobile systems vary widely in their internal workings, but they can be broadly categorized as either operating normally, or operating in a way that produces suboptimal driving (a flat tire, a loose timing belt) or prevents driving altogether (a dead battery). Smart buildings may be in an ideal state, or may suffer from some malfunction (broken climate control thermostat, inoperable windows, leaks in plumbing, etc).

a high-level discussion of maximum likelihood-based classifiers and maximally invariant-based classifiers. Maximum likelihood classifiers select the most likely estimate of the unknown parameters of a model. In contrast, maximally invariant classifiers first divide the unknown parameters into *test parameters* and *nuisance parameters*, where test parameters denote parameters that provide discriminatory information for testing, while nuisance parameters provide no discriminatory information. Once the test and nuisance parameters are identified, the maximally invariant classifiers seek to eliminate the effects of the nuisance parameters from the decision space while simultaneously maximizing the classifier's power, leading to a uniformly most powerful invariant classifier. Maximally invariant classifiers are designed to be invariant to specific sets of nuisance transformations on the data; thus, in this section, we also present the common nuisance transformations over population data that we will focus on in the remainder of this work.

We also present the shortcomings of each approach with respect to the aforementioned problem definition. Specifically, maximum likelihood-based classifiers cannot guarantee bounds on the false positive rate, while maximally invariant statistics—which do provide those bounds—do not typically exist.[6]

Much of this chapter follows the formulation given in Scharf and Demeure (1991). However, we simplify and clarify notation to present a more cohesive introduction to the original work that follows in Chapter 4.

[6]There exist special cases where maximum likelihood and maximally invariant tests are optimal solutions to Equation 3.5 (as discussed at the end of this section); however, these special cases are unlikely to manifest in clinical applications.

3.1 Problem Definition

In this section, we explicitly define our problem of interest: robust classification for varying populations. We describe the model and testing scenario, introduce performance metrics, and then state our robust classification optimization problem.

3.1.1 Model and Test Formulation

Consider a population of systems $\mathcal{V}$, where each individual system in the population $v \in \mathcal{V}$ conforms to a model with parameters $\boldsymbol{\theta}_v$ and produces some observable data $\boldsymbol{x}_v$ according to that model. In this work, we will focus on the simple linear model

$$\boldsymbol{x}_v = \boldsymbol{F}_v \boldsymbol{\theta}_v + \sigma_v \boldsymbol{n} \tag{3.1}$$

where $\boldsymbol{x}_v$ is a series of measurements from an individual system, $\boldsymbol{F}_v$ is a dynamics matrix which captures the relationship between the measurements and the unknown parameters $\boldsymbol{\theta}_v = \begin{bmatrix} \theta_1 & \dots & \theta_J \end{bmatrix}^\top$, and $\sigma_v \in \{\sigma \in \mathbb{R} | \sigma > 0\}$ scales a zero-mean noise, $\boldsymbol{n}$.[7]

Assume this population of systems can be divided into two disjoint subpopulations, or *classes* of systems, $\mathcal{V}_0$ and $\mathcal{V}_1$ ($\mathcal{V} = \mathcal{V}_0 \cup \mathcal{V}_1, \quad \mathcal{V}_0 \cap \mathcal{V}_1 = \varnothing$), each with its own set of possible parameters Θ_0 and Θ_1, where

$$v \in \mathcal{V}_y \Leftrightarrow \boldsymbol{\theta}_v \in \Theta_y.$$

For later convenience, we also assume that each of these classes $y \in \{0, 1\}$ is associated with a known parameter-indexing set, $\mathcal{I}_y \subseteq \mathcal{I} = \{1, \dots, J\}$, such that

$$\Theta_y = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^J, \ \forall j \in \mathcal{I}, \ j \notin \mathcal{I}_y \rightarrow \theta_j = 0\}.$$

[7] In this work we will make no assumption regarding the distribution of the noise other than that it is zero-mean.

That is, $\mathcal{I}_y$ denotes the elements of $\boldsymbol{\theta}$ that can be non-zero under each class. In this work, we will consider the scenario where each element of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_j$, is either known to be zero, or varies without restriction over $\mathbb{R}$.

We then pose the binary classification problem

$$\mathcal{H}_0 : \boldsymbol{\theta}_v \in \Theta_0 \;\text{ vs. }\; \mathcal{H}_1 : \boldsymbol{\theta}_v \in \Theta_1 \tag{3.2}$$

that aims to classify an individual as belonging to one of the two subpopulations.[8]

Note that this classification problem takes a form similar to that found in a standard regression model, with two major differences: first, $\boldsymbol{\theta}$ varies not just between classes, $y \in \{0, 1\}$, but also between individuals, $v \in \mathcal{V}$. Second, the dynamics matrix $\boldsymbol{F}$ also varies between individuals $v \in \mathcal{V}$.[9] This intersubject variation makes attempting to directly estimate the parameters for each individual difficult when the available training data is scarce—as is often the case when developing CDS systems for patient monitoring.[10]

The goal of this work is to develop a function $\hat{y}(\boldsymbol{x}_v)$ that is a binary classifier

$$\hat{y}(\boldsymbol{x}_v) = \begin{cases} 0 & \text{if } v \in \mathcal{V}_0 \\ 1 & \text{if } v \in \mathcal{V}_1 \end{cases} \tag{3.3}$$

[8]Note that this definition simplifies by assuming the parameter spaces are disjoint; they rarely are. For further discussion, see Section 3.3.

[9]When this intersubject variation in $\boldsymbol{F}$ is not present, the problem can be solved simply with any number of standard learning techniques.

[10]An alternate, but roughly equivalent formulation that may help illustrate the problem involves splitting the "model" $\boldsymbol{F}, v$ into two parts, $\boldsymbol{F}_{0,v}$ and $\boldsymbol{F}_{1,v}$ and writing the hypotheses as

$$\mathcal{H}_0 : \boldsymbol{x}_v = \boldsymbol{F}_{0,v}\boldsymbol{\theta}_v + \sigma_v \boldsymbol{n}$$
$$\mathcal{H}_1 : \boldsymbol{x}_v = \boldsymbol{F}_{1,v}\boldsymbol{\theta}_v + \sigma_v \boldsymbol{n}$$

with the classification problem being determining which of these two separate models more accurately describes the data.

and that achieves good performance despite high interpatient variance, as common in clinical applications. In the next section, we define the performance metrics we will use to define "good performance."

3.1.2 Performance Metrics

In clinical settings, sensitivity and specificity are common measures of classifier performance. The false positive rate is traditionally calculated as the total number of individual false positives divided by the number of individuals in $\mathcal{V}_0$, and is equivalent to one minus *specificity*. True positive rate is traditionally calculated as the total number of true positives divided by the number of samples in $\mathcal{V}_1$, and is commonly called *sensitivity*. Under this scheme, high rates of true positive and low rates of false positive do not necessarily indicate that good performance is achieved over all individuals in the population.

In this work, we quantify the performance of classifiers in terms of false positive and true positive rates. For an individual, $v \in \mathcal{V}$, let X_v represent the random variable "generating" the observed data $\boldsymbol{x}_v$. A false positive occurs when $\boldsymbol{\theta}_v \in \Theta_0$ but $\hat{y}(\boldsymbol{x}_v) = 1$. A true positive occurs when $\boldsymbol{\theta}_v \in \Theta_1$ and $\hat{y}(\boldsymbol{x}_v) = 1$.

False Positive Rate

If $\mathcal{H}_0$ is *composite* (*i.e.,* the set Θ_0 contains more than one element) then each individual $v \in \mathcal{V}_0$ can achieve a different false positive rate, $P_{\boldsymbol{\theta}_v}[\hat{y}(X_v) = 1]$. Thus, for a given candidate classifier $\hat{y}$, we consider an upper bound on false positive rate over all possible sets of parameters from the "negative" class and all individuals who belong to that negative class, $\mathcal{V}_0$:

$$P_{FP}(\hat{y}) = P_{\boldsymbol{\theta}_v}[\hat{y}(X_v) = 1] \leq \sup_{v \in \mathcal{V}_0,\, \boldsymbol{\theta} \in \Theta_0} P_{\boldsymbol{\theta}}[\hat{y}(X_v) = 1] = \sup_{v \in \mathcal{V}_0,\, \boldsymbol{\theta} \in \Theta_0} E_{\boldsymbol{\theta}}[\hat{y}(X_v)].$$

If the hypothesis $\mathcal{H}_0$ is *simple* (*i.e.,* $\Theta_0 = \{\boldsymbol{\theta}_0\}$), then $\boldsymbol{\theta}_v = \boldsymbol{\theta}_0$ and the above equation can be simplified:

$$P_{FP}(\hat{y}) = \sup_{v \in \mathcal{V}_0} P_{\boldsymbol{\theta}_0}[\hat{y}(X_v) = 1] = \sup_{v \in \mathcal{V}_0} E_{\boldsymbol{\theta}_0}[\hat{y}(X_v)]$$

for all $v \in \mathcal{V}_0$.

True Positive Rate

If $\mathcal{H}_1$ is composite, each possible $\boldsymbol{\theta}_v \in \Theta_1$ achieves a different true positive rate $P_{\boldsymbol{\theta}_v}[\hat{y}(X_v) = 1]$. We thus consider a lower bound on the true positive rate for any individual, $v \in \mathcal{V}_1$,

$$P_{TP}(\hat{y}) = P_{\boldsymbol{\theta}_v}[\hat{y}(X_v) = 1] \geq \inf_{v \in \mathcal{V},\, \boldsymbol{\theta} \in \Theta_1} P_{\boldsymbol{\theta}}[\hat{y}(X_v) = 1] = \inf_{v \in \mathcal{V},\, \boldsymbol{\theta} \in \Theta_1} E_{\boldsymbol{\theta}}[\hat{y}(X_v)].$$

Similar to the false positive rate, if $\mathcal{H}_1$ is simple ($\Theta_1 = \{\boldsymbol{\theta}_1\}$), then

$$P_{TP}(\hat{y}) = P_{\boldsymbol{\theta}_v}[\hat{y}(X_v) = 1] = E_{\boldsymbol{\theta}_1}[\hat{y}(X_v)]$$

as there is only one possible value for $\boldsymbol{\theta}_v$ in Θ_1.

3.1.3 Optimization Problem

As we have discussed, in many real-world scenarios, maintaining an upper bound on false positive rate over all individuals while maximizing the true positive rate is often a desirable property. In statistical hypothesis testing, a test which achieves this property is called *uniformly most powerful* (see Scharf and Demeure, 1991).

Uniformly Most Powerful (UMP) Classifier. A classifier $\hat{y}$ is a uniformly most powerful classifier of false positive rate α ($UMP_\alpha(\hat{y})$) if and only if it achieves a false positive rate

of α, and its true positive rate is uniformly greater than the true positive rate of any other test $\hat{y}'$ whose false positive rate is less than or equal to α. That is,

$$UMP_\alpha(\hat{y}) \Leftrightarrow \forall \hat{y}' \in \mathcal{Y}_\alpha,\ \forall v \in \mathcal{V},\ \forall \boldsymbol{\theta} \in \Theta_1,\ \ [\hat{y} \in \mathcal{Y}_\alpha \wedge E_{\boldsymbol{\theta}}\left[\hat{y}(X_v)\right] \geq E_{\boldsymbol{\theta}}\left[\hat{y}'(X_v)\right]]$$

where $\mathcal{Y}_\alpha = \{\hat{y} \mid P_{FP}(\hat{y}) \leq \alpha\}$ is the set of tests with false positive rates less than α.

Thus, for a given hypothesis testing problem, if a limit on false positive rate is a required property, finding a UMP classifier is ideal.

For illustrative purposes, let us consider a sample system consisting of two variables, x_1 and x_2, where each measurement variable is equal to one of two parameters plus some amount of zero-mean Gaussian noise:

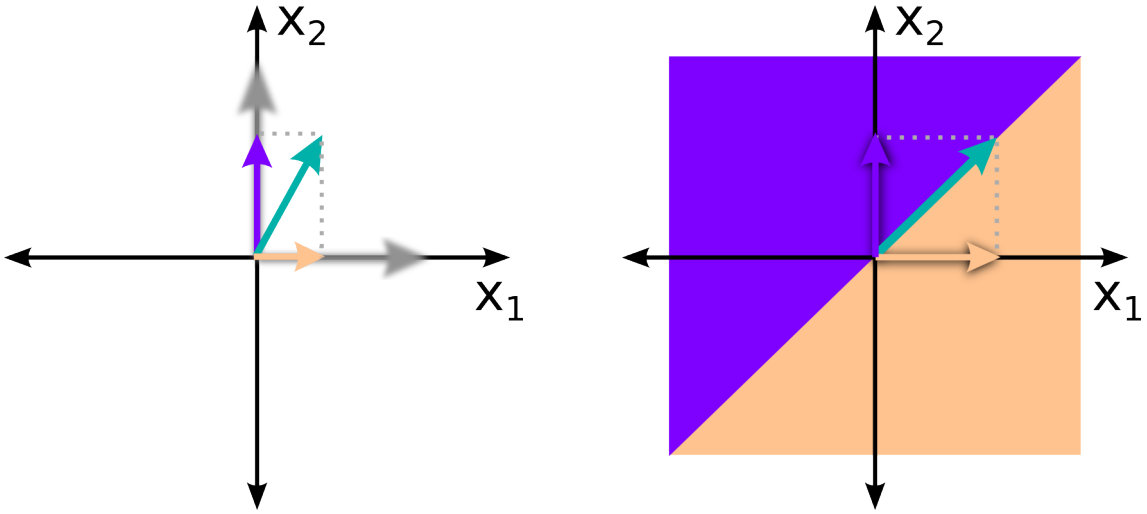
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{3.4}$$

where

$$\mathcal{H}_0 : (\theta_1, \theta_2) \in \Theta_{0,1} \times \Theta_{0,2}$$
$$\mathcal{H}_1 : (\theta_1, \theta_2) \in \Theta_{1,1} \times \Theta_{1,2}$$
$$n_k \sim N[0, 1].$$

This system is illustrated in Figure 3.1a; system observations (depicted as a teal vector) are the result of the sum of some underlying system parameters (depicted in orange for θ_1 and purple for θ_2) plus noise (depicted in gray).

If both $\mathcal{H}_0$ and $\mathcal{H}_1$ are simple hypotheses, the Neyman-Pearson lemma (see Scharf and Demeure, 1991) guarantees that a UMP classifier exists and can be designed by using an ordinary likelihood ratio test over (θ_0, θ_1) and setting an appropriate threshold cutoff for the value of the ratio. For example, in our sample system above, if our null and test hypotheses



(a) Sample system illustration.

(b) Sample system in the case of simple hypotheses. UMP classification regions shown via shaded background.

Figure 3.1: Graphs of the simple sample system used for illustrative purposes throughout this work. Measured output is represented by the teal arrow, made up of θ_1 (light orange) and θ_2 (darker purple) components, plus noise. The Neyman-Pearson lemma guarantees that a classifier based on the diagonal shading in the right figure is a UMP classifier. ($\mathcal{H}_0$ is the lighter orange region while $\mathcal{H}_1$ is the darker purple region.)

are

$$\mathcal{H}_0 : (\theta_1, \theta_2) \in \{1\} \times \{0\}$$
$$\mathcal{H}_1 : (\theta_1, \theta_2) \in \{0\} \times \{1\}$$

then the likelihood ratio test

$$\hat{y}(\boldsymbol{x}_v) = \begin{cases} 0 & \text{if } x_2 < x_1 + \eta \\ 1 & \text{if } x_2 \geq x_1 + \eta \end{cases}$$

is a UMP classifier, with η chosen to achieve a desired false positive rate. This result is illustrated for our sample system in Figure 3.1b.

While the Neyman-Pearson lemma is a powerful result, requiring simple hypotheses is overly restrictive. In order to achieve a UMP test in general, we must be able to design a likelihood ratio test for every possible θ without knowing θ ahead of time. In fact, it can be shown that a UMP test exists if and only if the likelihood ratio test for every $\theta \in \Theta$ can be completely defined without knowledge of the values of θ. It follows that if a UMP test exists, using it produces performance as good as what can be achieved when θ is known (as proven in Van Trees (2004)). However, in most cases it is difficult or impossible to define a test that works for all values of $\theta \in \Theta$ without knowledge of θ's specific values.[11]

Since UMP classifiers may not exist for all problems, in this work, we consider other solutions to the following relaxed classifier design problem:

Fixed False Positive Rate Classifier Optimization Problem.

$$\max_{\hat{y} \in \mathcal{Y}_\alpha} P_{TP}(\hat{y}) \tag{3.5}$$

Under this paradigm, if the UMP classifier exists, it is still optimal. A classifier $\hat{y}$ which satisfies Equation 3.5, when applied to the hypothesis testing problem in Equation 3.2 for population-varying system parameters, maximizes the minimum true positive rate while ensuring bounded false positive rate—a highly desirable property in clinical settings.[12]

In the following section, we consider one popular family of classifiers: those based on maximum likelihood. We explore whether they can achieve this property.

[11] In particular, "two-sided" tests over many distributions, which test whether θ is inside or out of some parameter range, are known not to exist. The exponential family of distributions is one such set of distributions; no "two-sided" test exists for it. In general, this occurs because values of θ greater than the parameter range and those less than the parameter range must be treated differently.

[12] In some clinical scenarios, it may be preferential to bound the minimum true positive rate and then minimize the false positive rate, which is equivalent to switching the hypotheses' labels in Equation 3.2.

3.2 Maximum Likelihood-Based Classifiers

In this subsection we provide a high-level summary of maximum likelihood classifiers for arbitrary systems. Given observed data points $\boldsymbol{x} = (x_1, \ldots, x_n)$ generated from a random variable $\boldsymbol{X}$ with a $\boldsymbol{\theta}$-parameterized distribution $f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$, where Θ is some parameter space, we can consider the observations as fixed and write the joint density function as varying over the space of parameters, producing the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$:

$$f(\boldsymbol{x}_1, \ldots, x_n|\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{x}).$$

The maximum likelihood estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ that maximizes L:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; \boldsymbol{x}).$$

Depending on the model for f being used, it may be possible to calculate the maximum likelihood estimate directly from the data. If no closed-form solution exists, numerical optimization methods can be employed. We described some of these methods in Section 2.2.3.

3.2.1 Generalized Likelihood Ratio

For the composite binary classification problem considered in this work, the parameters are known to be drawn from one of two parameter spaces, Θ_0 or Θ_1. We can thus directly calculate the maximum value of the likelihood function under each hypothesis (by exhaustively considering all possibilities) and consider the ratio. This is the *generalized likelihood ratio*, and it serves as a measure of class membership likelihood.

Generalized Likelihood Ratio. For parameters $\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1$, the generalized likelihood ratio of the measurement $\boldsymbol{x}$ is

$$L_R(\boldsymbol{x}) = \frac{\max_{\boldsymbol{\theta}_1 \in \Theta_1} L(\boldsymbol{\theta}_1; x)}{\max_{\boldsymbol{\theta}_0 \in \Theta_0} L(\boldsymbol{\theta}_0; x)}. \tag{3.6}$$

When the generalized likelihood ratio is compared to a threshold, η, the *generalized likelihood ratio test* (GLRT) results.[13]

Generalized Likelihood Ratio Test (GLRT).

$$\hat{y}(\boldsymbol{x}) = \begin{cases} 0, \ L_R(x) < \eta \\ 1, \ L_R(x) \geq \eta \end{cases} \tag{3.7}$$

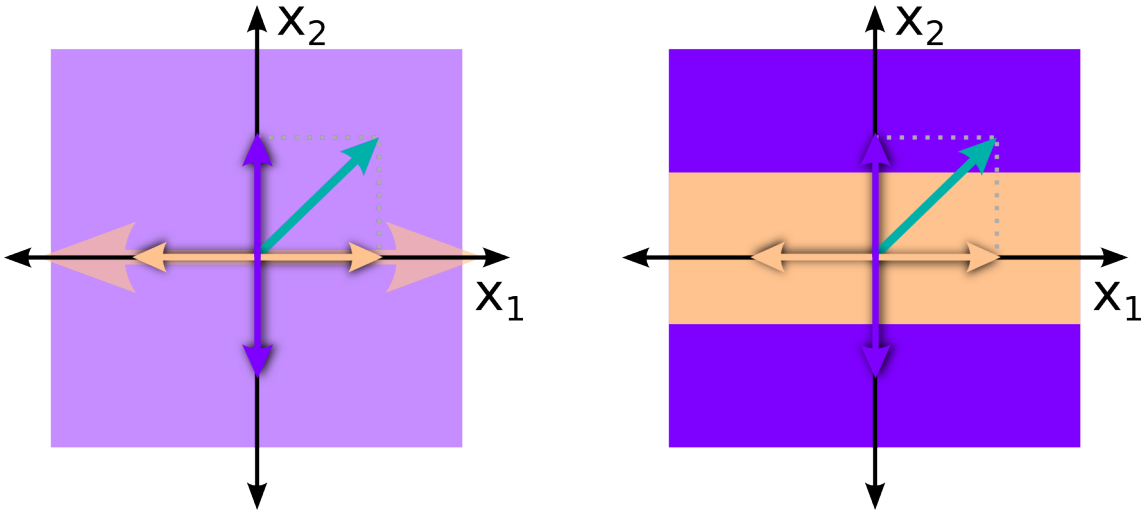
Let us consider again the sample system described in Equation 3.4, with null and candidate hypotheses defined as follows:

$$\mathcal{H}_0 : (\theta_1, \theta_2) \in \{\mathbb{R}\} \times \{0\}$$
$$\mathcal{H}_1 : (\theta_1, \theta_2) \in \{\mathbb{R}\} \times \{\mathbb{R}\}.$$

This scenario is illustrated in Figure 3.2a. Because the hypotheses here are not simple, direct application of the Neyman-Pearson lemma is impossible. Instead, we can consider the GLRT, $L_R(\boldsymbol{x})$:

$$\begin{aligned} L_R(\boldsymbol{x}) &= \frac{\max_{\boldsymbol{\theta}_1 \in \Theta_1} L(\boldsymbol{\theta}_1; x)}{\max_{\boldsymbol{\theta}_0 \in \Theta_0} L(\boldsymbol{\theta}_0; x)} \\ &= \max_{\boldsymbol{\theta}_1 \in \Theta_1} ln \ L(\boldsymbol{\theta}_1; x) - \max_{\boldsymbol{\theta}_0 \in \Theta_0} ln \ L(\boldsymbol{\theta}_0; x) \\ &= \frac{1}{2} \min_{\boldsymbol{\theta}_0 \in \mathbb{R} \times \{0\}} \sum_{k=1}^{2} (x_k - \theta_{0,k})^2 - \frac{1}{2} \min_{\boldsymbol{\theta}_1 \in \mathbb{R} \times \mathbb{R}} \sum_{k=1}^{2} (x_k - \theta_{1,k})^2 \\ &= \frac{x_2^2}{2} \end{aligned}$$

[13]For simplicity, we assign the boundary case (where $\boldsymbol{L}_R(\boldsymbol{x}) = \eta$) a value of 1. When we consider $\boldsymbol{x} \in \mathbb{R}^m$, and the probability of $\boldsymbol{L}_R(\boldsymbol{x}) = 1$ is 0, this boundary case is irrelevant. When $\boldsymbol{x}$ is taken from a countable set, the boundary is dealt with by assigning $\hat{y}$ some value γ when $\boldsymbol{x}$ falls on the boundary, where γ is equal to the probability of $\boldsymbol{x}$ actually falling on the boundary; this removes any potential biasing effects. For notational simplicity, however, we neglect this detail in the remainder of the text.



(a) Sample system with non-simple hypotheses.

(b) UMP solution provided by the GLRT.

Figure 3.2: Graphs of the sample system, updated to show the additional classification challenge presented by non-simple hypotheses. The shading on the right figure displays the classification solution provided by the GLRT.

and thus the GLRT in this scenario is

$$\hat{y}(\boldsymbol{x}) = \begin{cases} 0, & \frac{x_2^2}{2} < \eta \\ 1, & \frac{x_2^2}{2} \geq \eta \end{cases}$$

with η chosen to achieve a desired constant rate of false positive. This solution is illustrated geometrically in Figure 3.2b.

When used for classification, the GLRT has many attractive properties. For some families of distributions, it is UMP.[14] Even when not UMP, in applications where the change in parameters over the population is small, it tends to work well (*e.g.,* Kraut and Scharf (1999) established that for Gaussian distributions without invariance, the GLRT is UMP).

[14]For example, for any distribution with a monotone likelihood ratio, the Karlin-Rubin theorem can be applied; see Scharf and Demeure (1991).

However, as the estimated parameter values of the GLRT diverge from the true values—a likely occurrence when training data is scarce—the performance of the GLRT suffers. As with any maximum likelihood based classification approach, inability to achieve a good estimate of the likelihood translates to poor performance. The following lemma concerns the potential for the GLRT (or any other classifier) to be feasible for the classifier design problem in Equation 3.5.

Lemma 3.2.1 (Classifier Feasibility)**.** *Let $f_\theta(t)$ denote the θ-parameterized distribution of a statistic, $t(x)$, under the $\mathcal{H}_0$ hypothesis (i.e. $\theta \in \Theta_0$), for which a classifier $\hat{y}(\boldsymbol{x}) \in \{0, 1\}$ is designed such that $t(x) \geq \eta \longleftrightarrow \hat{y}(\boldsymbol{x}) = 1$. Then*

$$\exists \theta, \theta' \in \Theta_0 | f_{\theta'}(t) \neq f_\theta(t) \longleftrightarrow \exists \eta, \alpha, \int_\eta^\infty f_\theta(t)dt > \alpha \ \wedge \ \int_\eta^\infty f_{\theta'}(t)dt \leq \alpha.$$

Proof. From left to right,

$$\begin{aligned}
&\exists \theta, \theta' \in \Theta_0, f_{\theta'}(t) \neq f_\theta(t) \\
\longrightarrow &\exists \theta, \theta' \in \Theta_0, \ \exists \eta, \int_\eta^\infty f_{\theta'}(t)dt < \int_\eta^\infty f_\theta(t)dt \\
\longrightarrow &\exists \theta, \theta' \in \Theta_0, \ \exists \eta, \alpha, \int_\eta^\infty f_\theta(t)dt > \alpha \ \wedge \ \int_\eta^\infty f_{\theta'}(t)dt \leq \alpha
\end{aligned}$$

From right to left,

$$\begin{aligned}
&\exists \theta, \theta', \in \Theta \ \exists \eta, \alpha, \int_\eta^\infty f_\theta(t)dt > \alpha \ \wedge \ \int_\eta^\infty f_{\theta'}(t)dt \leq \alpha \\
\longrightarrow &\exists \theta, \theta', \in \Theta_0 \ \exists \eta, \int_\eta^\infty f_{\theta'}(t)dt < \int_\eta^\infty f_\theta(t)dt \\
\longrightarrow &\exists \theta, \theta', \in \Theta_0 \ \exists \eta, \int_\eta^\infty \left(f_\theta(t) - f_{\theta'}(t)\right) dt > 0 \\
\longrightarrow &\exists \theta, \theta', \in \Theta_0 \ f_\theta(t) \neq f_{\theta'}(t)
\end{aligned}$$

□

The above lemma demonstrates that the false positive rate is arbitrarily satisfied (*i.e.,* satisfied regardless of the maximum false positive rate) if and only if the distribution of the test statistic, under the null hypothesis, is invariant to the parameters. Since the GLRT as defined in Equation 3.7 (like most classifiers that focus on maximizing likelihood) has no mechanism to ensure this property, in applications where parameters can significantly affect the outputs—such as clinical monitoring and classification—the GLRT is likely to yield a false positive rate that varies significantly over the population. Such variance can be observed in a number of medical applications in Chapter 6.

3.3 Maximally Invariant Classifiers

As we have discussed, due to the GLRT's dependence on specific parameter estimates, the GLRT is unlikely to bound the false positive rate in populations with significant parameter variance between individuals. In this section, we turn our attention to maximally invariant statistics, and how to design classifiers incorporating them which achieve a bounded false positive rate.

Because it is likely that the parameter spaces of the two hypotheses will overlap, designing maximally invariant statistics requires separating the parameters into *nuisance parameters* and *test parameters* under each hypothesis:

Nuisance and Test Parameters. Given parameter sets under each hypothesis, Θ_0 and Θ_1, then

$$\hat{\Theta}_N = \Theta_0 \cap \Theta_1, \;\; \hat{\Theta}_0 = \Theta_0 \backslash \Theta_1, \;\; \text{and} \;\; \hat{\Theta}_1 = \Theta_1 \backslash \Theta_0,$$

denote the nuisance parameters, test parameters under $\mathcal{H}_0$, and test parameters under $\mathcal{H}_1$, respectively.

The nuisance parameters, $\hat{\Theta}_N$, are non-discriminatory; while they affect measurements through the model, their inclusion in both Θ_0 and Θ_1 indicates that they cannot provide any useful testing information. Allowing non-discriminatory information to enter the classifier through the measurements is a major contributor to poor classification performance.

As discussed in the previous subsection, the GLRT attempts to remove dimensions with non-discriminatory information by "removing" the MLE estimate of the nuisance parameters from the measurements through normalization; however, when training data is scarce, estimation accuracy decreases and more non-discriminatory information adds noise to the ratio value. To provide robust classifier performance in the presence of nuisance parameters, maximally invariant statistics eliminate all (and only all) possible effects the nuisance parameters can have on the measurements, thus guaranteeing the statistic and resulting classifier are unaffected by the nuisance parameters. In the remainder of this section, as motivation for our proposed technique, we examine the effects of the nuisance parameters and describe common types of nuisance transformations. We then discuss maximally invariant statistics, which can be used to create *uniformly most powerful invariant* (UMPI) tests which are always feasible for the classifier design problem in Equation 3.5 and are equivalent to the UMP test, if it exists.

3.3.1 Nuisance Transformations

In this section we consider sets of transformation functions induced by the nuisance parameters $\hat{\Theta}_N$. These *nuisance transformation sets* determine the impact of the parameters on the measurement space. These sets of nuisance transformations act as sets of endomorphisms on the measurement space, determining the impact that changes in parameters have on the measurements. Figure 3.3 attempts to illustrate this. Sets of possible parameters θ in parameter space Θ are mapped to observable measurements $\boldsymbol{x}$ via the dynamics matri-

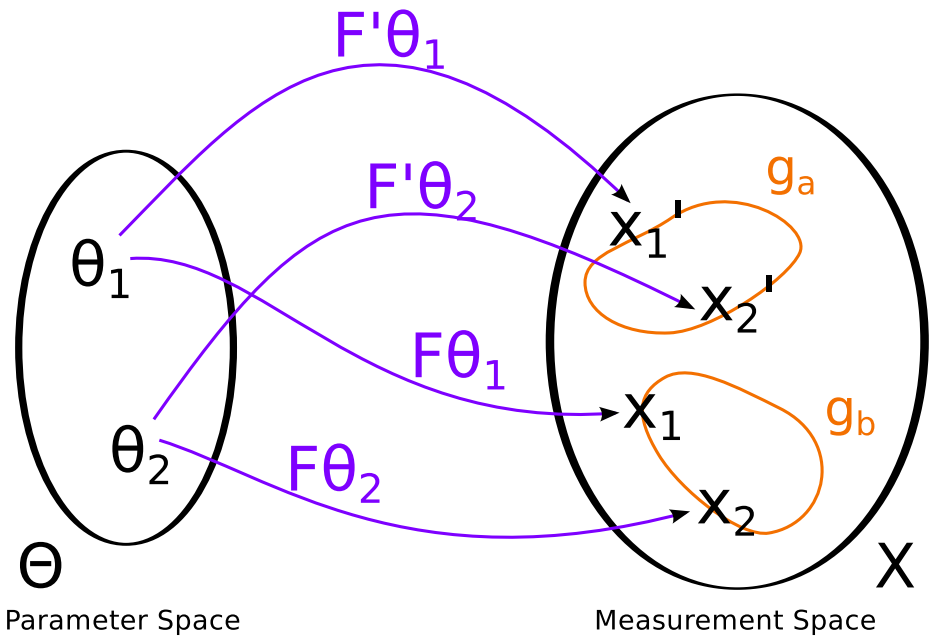
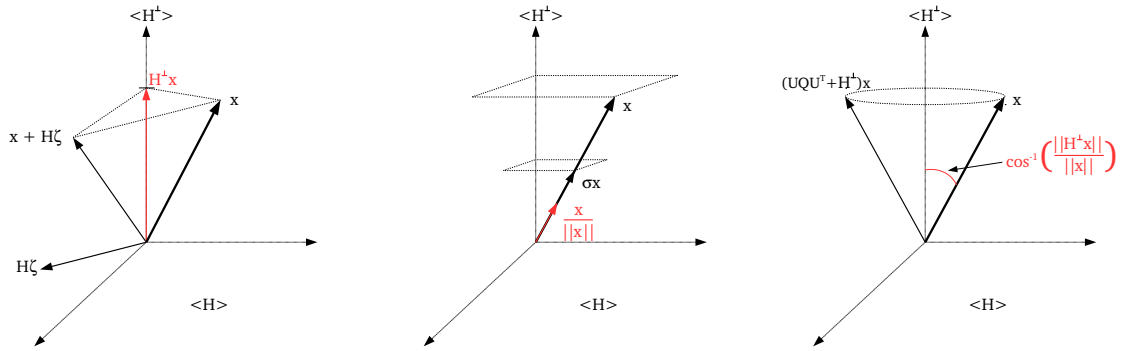


Figure 3.3: An illustration of how nuisance transformations (g_a and g_b, shown in orange) act as endomorphisms in the measurement space (X) which capture measurement perturbations that manifest when changes in the parameter space (Θ) occur. The $\boldsymbol{F}$ matrices ($\boldsymbol{F}$ and $\boldsymbol{F}'$, shown in purple) define the dynamics of a system, mapping parameters to measurements.

ces $\boldsymbol{F}$. In the image, we can imagine $\boldsymbol{F}$ models the dynamics of a healthy patient, while $\boldsymbol{F}'$ models the dynamics of a sick patient. θ_1 and θ_2 represent the parameters of two different patients. Each patient manifests different measurements in the measurement space depending on the model which applies to them ($\boldsymbol{x}_1$ and $\boldsymbol{x}_1'$ if patient θ_1 were healthy or sick, respectively, and $\boldsymbol{x}_2$ and $\boldsymbol{x}_2'$ if patient θ_2 were healthy or sick, respectively). Differences in the values of θ can be thought of as manifesting, through the choice of $\boldsymbol{F}$, as nuisance transformation on the measurements, g.

In this work, we focus on three particular types of nuisance transformations common

across population-sampled data in real-world applications: transformations pertaining to translation, transformations of scale, and transformations of rotation. While these sets of transformations are not exhaustive of all possible transformations, they present three broad classes and can be thought of as representative of translations in Cartesian (translation) and polar coordinate frames (rotation and scale). In the remainder of this work we will attempt to create classifiers that are invariant to these transformations in order to generate robust classifiers for real-world healthcare monitoring. The two types of transformations, along with terms invariant to them, are illustrated in Figure 3.4. (In this work, specific sets of transformations are represented by script $\boldsymbol{G}$ [that is, $\mathcal{G}$].)



(a) Translation (b) Scaling (c) Rotation

Figure 3.4: Visual representation of the types of transformations included in the transformation set $\mathcal{G}_y$. Terms invariant to the transformations are shown in red.

Translation

A *translation* occurs when output data is (partially) generated through a known process with unknown parameters. A translation is illustrated in Figure 3.4a. Output data $\boldsymbol{x}$ is summed with a vector $\boldsymbol{H\zeta}$ of unknown magnitude and direction (in a known subspace,

$\langle \boldsymbol{H} \rangle$) such that $\boldsymbol{x} + \boldsymbol{H}\boldsymbol{\zeta}$ results.[15] From the figure, we observe that the vector $\boldsymbol{H}^{\perp}\boldsymbol{x}$ is invariant to any translation in the subspace $\langle \boldsymbol{H} \rangle$, as $\boldsymbol{H}^{\perp}\boldsymbol{H} = 0$. For the simple model in Equation 3.1, a translation occurs via the nuisance parameter $\boldsymbol{\zeta}$ as

$$\mathcal{G}_t = \{g \mid g(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{H}\boldsymbol{\zeta}\}.$$

Rotation

A *rotation* occurs when a translation is performed on a signal in a subspace without affecting the magnitude of the signal. A nuisance rotation denotes that only the *magnitude* of the transformed measurements contains remaining discriminatory information.[16] The effect of a rotation is illustrated in Figure 3.4c. From the figure, we observe that the angle between the cone of rotation and the subspace $\langle \boldsymbol{H} \rangle^{\perp}$, namely $\cos^{-1}\left(\frac{\|\boldsymbol{H}^{\perp}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\right)$ is invariant to an arbitrary rotation $\boldsymbol{Q}$ in $\langle \boldsymbol{H} \rangle$. Letting $\mathcal{R}$ be the set of rotation matrices

$$\mathcal{R} = \{\boldsymbol{R} \mid \boldsymbol{R}^{-1} = \boldsymbol{R}^{\top}, |\boldsymbol{R}| = 1\},$$

an output $\boldsymbol{x}$ is rotated in the subspace $\langle \boldsymbol{H} \rangle$ by applying a rotation, $\boldsymbol{Q} \in \mathcal{R}$, in the subspace of $\langle \boldsymbol{H} \rangle$, denoted as $\boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top}$ (where $\boldsymbol{U}\boldsymbol{U}^{\top} = \boldsymbol{H}$). For the model in Equation 3.1, rotation occurs via the unknown noise coefficient, which manifests as the nuisance parameters $\boldsymbol{Q}$:

$$\mathcal{G}_r = \left\{g \;\middle|\; g(\boldsymbol{x}) = \left(\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top}\right)\boldsymbol{x}\right\}$$

where $\boldsymbol{V} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\top}$.

[15] When the known subspace is a single dimension (*i.e.,* the subspace is a vector), a translation consists of an unknown magnitude in the direction of the subspace vector.

[16] An example of a nuisance rotation occurs when monitoring for an event in a window of time; whether the measurements deviate from a nominal value early or late in the window (*i.e.,* direction) may be much less informative than the magnitude of change.

Scale

Scaling occurs when the outputs are generated by a process which multiplies them by an unknown magnitude.[17] The effect of scaling is captured in Figure 3.4b, where an output $\boldsymbol{x}$ is scaled by σ such that $\sigma\boldsymbol{x}$ results.

From the figure, we observe that the normalized vector $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ is invariant to this arbitrary scaling. For the model in Equation 3.1, scaling is a direct result of the unknown noise coefficient, σ, which induces the group of transformations

$$\mathcal{G}_s = \{g \mid g(\boldsymbol{x}) = \sigma\boldsymbol{x},\ \sigma \geq 0\}.$$

Combined Transformation Set $\mathcal{G}_y$

By composing functions from the three sets, we can consider a new set of transformations $\mathcal{G}_y$ induced by the nuisance parameters $(\boldsymbol{\zeta}, \sigma, \boldsymbol{Q}) \in \hat{\Theta}_N$ as

$$\begin{aligned} \mathcal{G}_y &= \{g_t \circ g_r \circ g_s \mid g_s \in \mathcal{G}_s,\ g_r \in \mathcal{G}_r,\ g_t \in \mathcal{G}_t\} \\ &= \left\{g \;\middle|\; g(\boldsymbol{x}) = \sigma\left(\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top\right)\boldsymbol{x} + \boldsymbol{H}\boldsymbol{\zeta}\right\} \end{aligned} \tag{3.8}$$

where $\boldsymbol{V} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I} - \boldsymbol{H}^\perp$. This combined set of transformations is of particular interest because it combines two common types of nuisance transformations and thus encompasses a broad class of nuisance transformations likely to appear in data. For example, as previously mentioned, when monitoring for an event in a window of time, whether the measurements deviate from a nominal value early or late in the window (*i.e.,* direction) may be much less informative than the magnitude of change; the time of event induces a nuisance rotation in the data. A sudden shift in the position of a sensor, perhaps due to a patient's movement, manifests as a nuisance translation. The patient's data will be shifted

[17] Scaling captures the effects of unknown uncertainty in a process (*e.g.,* the variance of metabolism in a diabetic population).

uniformly in a subspace of unknown magnitude and direction, but such a shift would not be relevant to the detection problem.

As mentioned in the previous paragraphs, the transformations of translation, rotation, and scale correspond to real-world noise that routinely occurs in medical data. In fact, for the model in Equation 3.1, these transformations are sufficiently general to cover all effects of nuisance parameters. Even if all assumptions do not hold, most noise should be captured by such a general set. However, this is not true for all models. Development of transformation sets and corresponding invariance operators capable of expressing more complex transformations and effects is a subject of ongoing research. As mentioned in Section 3.1.1, the theoretical foundations of this work assume parameters are either known to be zero or vary without restriction over $\mathbb{R}$. Though this is likely not to be true in practice (*e.g.,* physiologic signals are rarely sensibly negative), deviations from this formulation will still produce useful statistics, and practical applications have produced good performance. It can be shown that there are less restrictive formulations for $\boldsymbol{\zeta}, \sigma$, and $\boldsymbol{Q}$ which still meet conditions that will be necessary in the remainder of this work. We leave these formulations for future work.

3.3.2 Maximally Invariant Statistics

When a population consists of systems impacted by a nuisance transformation set, such as the set $\mathcal{G}_y$ defined in Equation 3.8, and no UMP test is apparent, a possible approach to building a classifier with good performance is to design a statistic (and a corresponding classifier which uses that statistic) that achieves the best performance possible (*maximal*) while being unchanged by the entire nuisance transformation set (*invariant*). Such a statistic is referred to in Scharf and Friedlander (1994) as a *maximally invariant statistic*.

Maximality. A statistic $t(\boldsymbol{x})$ is maximal with respect to a nuisance transformation set $\boldsymbol{G}$

(that is, $maximal(t, \boldsymbol{G})$) if and only if

$$t(\boldsymbol{x}) = t(\boldsymbol{x}') \longrightarrow \exists\ \boldsymbol{g} \in \boldsymbol{G} \mid \boldsymbol{x}' = \boldsymbol{g}(\boldsymbol{x}).$$

Invariance. A statistic $t(\boldsymbol{x})$ is invariant with respect to a nuisance transformation set $\boldsymbol{G}$ (that is, $invariant(t, \boldsymbol{G})$) if and only if

$$\forall \boldsymbol{g} \in \boldsymbol{G},\ \ t(\boldsymbol{g}(\boldsymbol{x})) = t(\boldsymbol{x}).$$

Maximal Invariance. A statistic $t(\boldsymbol{x})$ is maximally invariant with respect to a nuisance transformation set $\boldsymbol{G}$ (that is, $maxinv(t, \boldsymbol{G})$) if and only if

$$maxinv(t, \boldsymbol{G}) \Leftrightarrow maximal(t, \boldsymbol{G}) \ \ \wedge \ \ invariant(t, \boldsymbol{G}).$$

Maximally invariant statistics are designed to be maximal and invariant over specific sets of nuisance transformations, $\boldsymbol{G}$, so the type of nuisance transformations assumed to be present in the data impact the design of the statistic. Note that maximally invariant statistics are not unique, but a maximally invariant statistic is sufficient for all other invariant statistics. In cases where a maximally invariant statistic exists and has a monotone likelihood ratio, the Karlin-Rubin theorem (see Scharf and Demeure, 1991) ensures the statistic can be used to design a *uniformly most powerful invariant* classifier:

Uniformly Most Powerful Invariant Classifier. Given a statistic $t(\boldsymbol{x})$ which is known to be maximally invariant to the nuisance transformation set $\boldsymbol{G}$, (that is, $maxinv(t, \boldsymbol{G})$), the classifier

$$\hat{y}_t(x) = \begin{cases} 1,\ t(\boldsymbol{x}) \geq \eta \\ 0,\ t(\boldsymbol{x}) < \eta \end{cases}$$

is uniformly most powerful invariant (UMPI):

$$UMPI(\hat{y}_t) \Leftrightarrow \hat{y}_t \in \mathcal{Y}_{\alpha,invariant}$$
$$\wedge \forall \hat{y}' \in \mathcal{Y}_{\alpha,invariant} \; \forall \, v \in \mathcal{V} \; \forall \, \boldsymbol{\theta} \in \Theta_1, \; E_{\boldsymbol{\theta}} \left[\hat{y}_t(X_v)\right] \geq E_{\boldsymbol{\theta}} \left[\hat{y}'(X_v)\right]$$

where $\mathcal{Y}_{\alpha,invariant} = \{\hat{y} \mid \hat{y} \in \mathcal{Y}_\alpha \; \wedge \; \forall \boldsymbol{g} \in \boldsymbol{G}, \; \hat{y}(\boldsymbol{x}) = \hat{y}(\boldsymbol{g}(\boldsymbol{x}))\}$.

In words, a classifier $\hat{y}_t$ which is created using a statistic invariant to the nuisance transformation set $\boldsymbol{G}$ is UMPI for false positive rate α if and only if it achieves a false positive rate of α and if its true positive rate is uniformly greater than the true positive rate of any other test $\hat{y}'$ which is also invariant to $\boldsymbol{G}$ and has false positive rate less than or equal to α. Note that the UMPI test is not necessarily optimal for the problem in Equation 3.5 (since there could exist a more powerful test that achieves the desired false positive rate but is not invariant to $\boldsymbol{G}$). This again draws attention to the relationship between maximum likelihood and maximally invariant based classifiers: maximum likelihood–based classifiers tend to have better detection rates, but provide no guarantees on false positive rates; maximally invariant-based classifiers provide guaranteed false positive rates, but tend to have lower true positive rates.

However, as the following lemmas prove, there are several necessary conditions for the existence of a maximally invariant statistic. Note that the contrapositive of these lemmas means that, if any of these conditions do not hold, a maximally invariant statistic cannot exist.

Lemma 3.3.1 (Identity and Maximal Invariance)**.** *The inclusion of the identity endomorphism $e(x) = x$ in $\boldsymbol{G}$ is a necessary condition for the existence of a maximally invariant statistic $maxinv(t, \boldsymbol{G})$.*

Proof. By maximality, with $x = x'$,

$$t(x) = t(x) \Longrightarrow \exists \; g \in \boldsymbol{G}, x = g(x).$$

We then let $g(x) = e(x)$. □

Lemma 3.3.2 (Inverse and Maximal Invariance)**.** *The inclusion of inverse endomorphisms* $\forall g \in \boldsymbol{G}\ \exists g' \in \boldsymbol{G} \mid g'(g(x)) = x$ *is a necessary condition for the existence of a maximally invariant statistic* $maxinv(t, \boldsymbol{G})$.

Proof. Let $g(x) = x'$. Then

$$\begin{aligned} & t(x) = t(g(x)) \\ & \implies \exists\ g' \in \boldsymbol{G}, x = g'(x') \\ & \implies x = g'(g(x)). \end{aligned}$$

□

Theorem 3.3.3 (Orbital Uniqueness and Maximal Invariance)**.** *Let* $\mathcal{O}(x)$ *be the* orbit *of* x*, defined as*

$$\mathcal{O}(x) = \{g(x), \forall g \in \boldsymbol{G}\}.$$

Then, orbit uniqueness

$$\forall x_1, x_2, x_1 \neq g(x_2) \implies \mathcal{O}(x_1) \cap \mathcal{O}(x_2) = \varnothing$$

is a necessary condition for the existence of a maximally invariant statistic $maxinv(t, \boldsymbol{G})$.

Proof. By contradiction, assume $x_1 \neq x_2$, and

$$\exists\ \bar{x} \in \mathcal{O}(x_1) \cap \mathcal{O}(x_2) \wedge \exists\ \hat{x} \in \mathcal{O}(x_1), \hat{x} \notin \mathcal{O}(x_2).$$

Then,

$$\begin{aligned} & \bar{x} \in \mathcal{O}(x_1) \cap \mathcal{O}(x_2) \\ & \implies \exists\ g_1 \in \boldsymbol{G} \mid \bar{x} = g_1(x_1) \wedge \exists\ g_2 \in \boldsymbol{G} \mid \bar{x} = g_2(x_2). \end{aligned}$$

By the invariance property of t,

$$\begin{aligned} t(x_1) &= t(g_1(x_1)) \\ &= t(\bar{x}) \\ &= t(g_2(x_2)) \\ &= t(x_2). \end{aligned}$$

By maximality,

$$t(x_1) = t(x_2) \implies \exists\ \bar{g} \in \boldsymbol{G} \mid x_1 = \bar{g}(x_2).$$

But then,

$$\begin{aligned} & \hat{x} \in \mathcal{O}(x_1) \\ \implies & \exists\ \hat{g} \in \boldsymbol{G} \mid \hat{x} = \hat{g}(x_1) \\ \implies & t(\hat{x}) = t(\hat{g}(x_1)) \\ & = t(x_1) \\ & = t(\bar{g}(x_2)) \\ & = t(x_2) \end{aligned}$$

and by maximality,

$$\begin{aligned} t(\hat{x}) = t(x_2) &\implies \exists\ g \in \boldsymbol{G} \mid \hat{x} = g(x_2) \\ &\implies \hat{x} \in \mathcal{O}(x_2) \end{aligned}$$

which is a contradiction. Thus, either $x_1 = x_2$, or $\mathcal{O}(x_1) \cap \mathcal{O}(x_2) = \varnothing$. □

Additionally, as the following lemma suggests, finding the maximally invariant statistics that result in UMPI tests over the hypothesis testing problem can rarely be done in practice, because the existence of a maximally invariant statistic places a strong restriction on the sets of transformations induced by the parameters under each hypothesis.

Lemma 3.3.4 (Existence of a maximally invariant statistic implies equivalence of null and nuisance parameters)**.** *The existence of a statistic which is maximally invariant to* $\boldsymbol{G}$ *under both* $\mathcal{H}_0$ *and* $\mathcal{H}_1$ *implies that the null parameter set is equal to the nuisance parameter set.*

Proof. Given $maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N}) \wedge maxinv(t, \boldsymbol{G}_{\hat{\Theta}_0})$:

$$\begin{aligned} maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N}) &\implies max(t, \boldsymbol{G}_{\hat{\Theta}_N}) \wedge inv(t, \boldsymbol{G}_{\hat{\Theta}_N}) \\ &\implies t(x) = t(g(x)) \ \forall g \in \boldsymbol{G}_{\hat{\Theta}_N} \\ &\implies t(x) = t(x') \text{ with } g(x) = x' \\ &\implies \exists g \in \boldsymbol{G}_{\hat{\Theta}_N} | g(x) = x' \end{aligned}$$

Now, consider some $g' \in \boldsymbol{G}_{\hat{\Theta}_0} \setminus \boldsymbol{G}_{\hat{\Theta}_N}$, and let $g'(x) = x''$. Then

$$\begin{aligned} maxinv(t, \boldsymbol{G}_{\hat{\Theta}_0}) &\implies max(t, \boldsymbol{G}_{\hat{\Theta}_0}) \wedge inv(t, \boldsymbol{G}_{\hat{\Theta}_0}) \\ &\implies t(x) = t(g(x)) \forall g \in \boldsymbol{G}_{\hat{\Theta}_0} \\ &\implies t(x) = t(g'(x)) \\ &\implies t(x) = t(x'') \\ &\implies \exists g' \in \boldsymbol{G}_{\hat{\Theta}_0} | g'(x) = x'' \end{aligned}$$

but also

$$\begin{aligned} max(t, \boldsymbol{G}_{\hat{\Theta}_N}) \wedge t(x) = t(x'') &\implies \exists g \in \boldsymbol{G}_{\hat{\Theta}_N} | g(x) = x'' \\ &\implies g'(x) = g(x) \\ &\implies g' \in \boldsymbol{G}_{\hat{\Theta}_N} \\ &\implies \boldsymbol{G}_{\hat{\Theta}_0} = \boldsymbol{G}_{\hat{\Theta}_N} \end{aligned}$$

□

Lemma 3.3.4 demonstrates that the existence of a maximally invariant statistic itself implies that the set of transformations induced by the null test parameters is the same as the set induced by the nuisance parameters. This is a very strong restriction. In practical applications, where there is large uncertainty in the parameters for both hypotheses and where most models only approximate the real world, it is therefore unlikely that a maximally invariant statistic can be found. This motivates our work in the next chapter, where we will consider how to create a near–maximally invariant statistic.

3.4 Summary

In this chapter, we provided background and precise formulation our problem of interest: robust classification of systems while limiting false positive classifications when systems are taking from two populations with varying dynamics. We defined the essential metrics for this problem. We also defined common nuisance transformations, and showed that maximal invariance, though a desirable property in the face of these nuisance transformations, is rarely achievable in practice.

Chapter 4

Parameter Invariant Statistics for Linear Time-Invariant Systems

Motivated by the shortcomings of the maximum likelihood and maximally invariant approaches presented in Chapter 3, this chapter introduces *parameter invariant* (PAIN) statistics, a class of statistics which can be used as part of a likelihood ratio test to create robust binary classifiers over linear time-invariant systems. We show that these classifiers maintain a constant rate of false positives over each individual in the population, while achieving good true positive performance though they are not necessarily UMP/UMPI. They achieve this good performance by being invariant to both known nuisance parameters and the test parameters under $\mathcal{H}_0$. This has the effect of reducing the parameter space and eliminating noise, as well as providing consistency in false positive rates. In scenarios where a UMPI classifier exists, we prove the PAIN statistic converges to the UMPI classifier.

First, we describe the systems of interest to us, linear time-invariant systems, which under common conditions have no UMP test. We then describe how to formulate statistics which are invariant to specified groups of nuisance transformations imposed by parameters.

Then, we show how to calculate such a statistic for the group of transformations of interest described in Section 3.3.1.

4.1 Linear Time-Invariant Systems

Though the techniques we have described thus far can be applied to systems that behave according to any sort of distribution,[18] in this and subsequent chapters we will narrow our focus to consider linear time-invariant (LTI) systems. Our interest in LTI systems stems from their ease of analysis (compared to time-varying and/or nonlinear systems), and the fact that they serve as good approximations for most real-world systems, especially over relatively small windows (a process known as linearization, discussed at length in Bentler and Dijkstra (1985)).

Consider again our population $\mathcal{V}$ of systems. This time, let each individual system in the population, $v \in \mathcal{V}$, be a system whose input-output response obeys a multiple-input single-output LTI model:

$$x_v(k) = \sum_{j=1}^{N} a_{v,j} x_v(k-j) + \sum_{l=1}^{L} \sum_{i=1}^{N} b_{v,l,i} u_{v,l}(k-i) + \sigma_v n_v(k).$$

Here, k represents the time step, N is the order of the model, x_v is the output, $u_{v,l}$ is the l-th input, and n_v is a noise. The model parameters $\boldsymbol{a}_v = \begin{bmatrix} a_{v,1} & \dots & a_{v,N} \end{bmatrix}^\top$, $\boldsymbol{b}_{v,l} = \begin{bmatrix} b_{v,l,1} & \dots & b_{v,l,N} \end{bmatrix}^\top$, and σ_v are non-stochastic for each individual $v \in \mathcal{V}$, but vary with unknown randomness across the population $\mathcal{V}$. In contrast to standard assumptions, we do not assume priors over these parameters.

From each individual system in the population, we consider a window of T sequential measurements and assume, without loss of generality, that the first measurement corre-

[18] Single Gaussian random variables provide the simplest example.

sponds to $k = N + 1$. Then, we define $\boldsymbol{x}_v = \begin{bmatrix} x_v(N+1) & \dots & x_v(N+T) \end{bmatrix}^\top$, such that the time-concatenated measurements are modeled as

$$\boldsymbol{x}_v = \boldsymbol{A}_v \boldsymbol{a}_v + \sum_{l=1}^{L} \boldsymbol{B}_{v,l} \boldsymbol{b}_{v,l} + \sigma_v \boldsymbol{n} \tag{4.1}$$

where

$$\begin{aligned}
\boldsymbol{A}_v &= \begin{bmatrix} x_v(N) & \dots & x_v(1) \\ \vdots & \ddots & \vdots \\ x_v(N+T-1) & \dots & x_v(T) \end{bmatrix} \\
\boldsymbol{B}_{v,l} &= \begin{bmatrix} u_{v,l}(N) & \dots & u_{v,l}(1) \\ \vdots & \ddots & \vdots \\ u_{v,l}(N+T-1) & \dots & u_{v,l}(T) \end{bmatrix}, \\
\boldsymbol{n} &= \begin{bmatrix} n(N+1) & \dots & n(N+T) \end{bmatrix}^\top.
\end{aligned}$$

For simplicity, we then concatenate the input/output matrices and the parameter matrices, which allows us to write the model in Equation 4.1 in the form described in Equation 3.1:

$$\boldsymbol{x}_v = \boldsymbol{F}_v \boldsymbol{\theta}_v + \sigma_v \boldsymbol{n},$$

where

$$\begin{aligned}
\boldsymbol{F}_v &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}_1 & \dots & \boldsymbol{B}_L \end{bmatrix} \\
\boldsymbol{\theta}_v &= \begin{bmatrix} \boldsymbol{a}^\top & \boldsymbol{b}_1^\top & \dots & \boldsymbol{b}_L^\top \end{bmatrix}^\top.
\end{aligned}$$

Recall that Equation 3.1 takes the familiar form of a standard regression model, with one major difference: θ and σ vary with the individual. This key difference motivates the remainder of the work. In the next section, we describe how to develop a statistic which achieves good classification performance despite this variance.

4.2 Parameter Invariant Statistics

Recall from Definition 3.3 the three sets of parameters defined by the binary hypothesis testing problem in Equation 3.2: $\hat{\Theta}_N$, $\hat{\Theta}_0$, and $\hat{\Theta}_1$, denoting the nuisance parameters, test parameters under $\mathcal{H}_0$, and test parameters under $\mathcal{H}_1$, respectively. These sets of parameters impose transformation groups $\boldsymbol{G}_{\hat{\Theta}_N}$, $\boldsymbol{G}_{\hat{\Theta}_0}$, and $\boldsymbol{G}_{\hat{\Theta}_1}$ on the data. As part of discussion surrounding Lemma 3.3.4, we showed that achieving maximal invariance to only the nuisance transformation group is rarely possible. In this section, we consider an alternate to the maximally invariant statistic: the more general *parameter invariant* (PAIN) statistic, which is maximally invariant to both the nuisance *and* the null test parameters, and thus discriminates based only on the event test parameters.

Parameter Invariant Statistic. Given transformation groups $\boldsymbol{G}_{\hat{\Theta}_0}$, and $\boldsymbol{G}_{\hat{\Theta}_N}$, let $\boldsymbol{G}_{\hat{\Theta}_N,\hat{\Theta}_0}$ $= \{g_{N,0}(\boldsymbol{x}) = g_N(g_0(\boldsymbol{x})) \mid g_N \in \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{g}_0 \in \boldsymbol{G}_{\hat{\Theta}_0}\}$. The statistic t is *parameter invariant* to $\boldsymbol{G}_{\hat{\Theta}_N}$ and $\boldsymbol{G}_{\hat{\Theta}_0}$ ($PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})$) if and only if it is maximally invariant to $\boldsymbol{G}_{\hat{\Theta}_N,\hat{\Theta}_0}$:

$$PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0}) \;\; \Leftrightarrow \;\; maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N,\hat{\Theta}_0}). \tag{4.2}$$

The compositional nature of a PAIN statistic provides several attractive properties.

Lemma 4.2.1 (Properties of a PAIN statistic)**.** *A PAIN statistic has the following properties:*

$$\begin{aligned} PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0}) &\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_0}) \\ PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0}) &\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_N}) \\ \hat{\Theta}_0 = \varnothing \;\; \wedge \;\; PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0}) &\longrightarrow maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N}) \end{aligned}$$

Proof. For the first property, $PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})$ implies

$$\begin{aligned}
&\longrightarrow maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) && \text{(PAIN definition)} \\
&\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) && \text{(maxinv definition)} \\
&\longrightarrow \forall g_{N,0} \in \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}, t(g_{N,0}(\boldsymbol{x})) = t(\boldsymbol{x}) && \text{(invariant definition)} \\
&\longrightarrow \forall g_0 \in \boldsymbol{G}_{\hat{\Theta}_0}, \forall g \in \boldsymbol{G}_{\hat{\Theta}_N}, t(g_N(g_0(\boldsymbol{x}))) = t(\boldsymbol{x}) && (\boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) \text{ definition} \\
&\longrightarrow \forall g_0 \in \boldsymbol{G}_{\hat{\Theta}_0},\ t(I(g_0(\boldsymbol{x}))) = t(\boldsymbol{x}) && \text{identity axiom, let } g_N(\boldsymbol{x}) = I(\boldsymbol{x}) \\
&\longrightarrow (\forall g_0 \in \boldsymbol{G}_{\hat{\Theta}_0},\ t(g_0(\boldsymbol{x})) = t(\boldsymbol{x}) \\
&\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_0}).
\end{aligned}$$

For the second property, $PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})$ implies

$$\begin{aligned}
&\longrightarrow maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) && \text{(PAIN definition)} \\
&\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) && \text{(maxinv definition)} \\
&\longrightarrow \forall g_{N,0} \in \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}, t(g_{N,0}(\boldsymbol{x})) = t(\boldsymbol{x}) && \text{(invariant definition)} \\
&\longrightarrow \forall g_0 \in \boldsymbol{G}_{\hat{\Theta}_0}, \forall g_N \in \boldsymbol{G}_{\hat{\Theta}_N}, t(g_N(g_0(\boldsymbol{x}))) = t(\boldsymbol{x}) && (\boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}) \text{ definition} \\
&\longrightarrow \forall g_N \in \boldsymbol{G}_{\hat{\Theta}_N},\ t(g_N(I(\boldsymbol{x}))) = t(\boldsymbol{x}) && \text{identity axiom, let } g_0(\boldsymbol{x}) = I(\boldsymbol{x}) \\
&\longrightarrow \forall g_N \in \boldsymbol{G}_{\hat{\Theta}_N},\ t(g_N(\boldsymbol{x})) = t(\boldsymbol{x}) \\
&\longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_N}).
\end{aligned}$$

For the third property, let $PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0}) \longrightarrow invariant(t, \boldsymbol{G}_{\hat{\Theta}_N})$ from the second

property above be *Claim A*. Then, $\hat{\Theta}_0 = \varnothing \ \wedge \ PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})$ implies

$\longrightarrow maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0})$ (PAIN definition)

$\longrightarrow maximal(t, \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0})$ (maxinv definition)

$\longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \rightarrow \exists g_{N,0} \in \boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0} \mid g_{N,0}(\boldsymbol{x}) = \boldsymbol{x}'$ (maximality)

$\longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \rightarrow \exists g_N \in \boldsymbol{G}_{\hat{\Theta}_N}, g_0 \in \boldsymbol{G}_{\hat{\Theta}_0} \mid g_N(g_0(\boldsymbol{x})) = \boldsymbol{x}'$ ($\boldsymbol{G}_{\hat{\Theta}_N, \hat{\Theta}_0}$ definition)

$\longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \rightarrow \exists g_N \in \boldsymbol{G}_{\hat{\Theta}_N} \exists g_0 \in I \mid g_N(g_0(\boldsymbol{x})) = \boldsymbol{x}'$ ($\hat{\Theta}_0 = \varnothing \rightarrow \boldsymbol{G}_{\hat{\Theta}_0} = \varnothing$)

$\longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \rightarrow \exists g_N \in \boldsymbol{G}_{\hat{\Theta}_N} \mid g_N(I(\boldsymbol{x})) = \boldsymbol{x}'$

$\longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \rightarrow \exists g_N \in \boldsymbol{G}_{\hat{\Theta}_N} \mid g_N(\boldsymbol{x}) = \boldsymbol{x}'$

$\longrightarrow maximal(t, \boldsymbol{G}_{\hat{\Theta}_N})$

$\longrightarrow maximal(t, \boldsymbol{G}_{\hat{\Theta}_N}) \ \wedge \ PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})$

$\longrightarrow maximal(t, \boldsymbol{G}_{\hat{\Theta}_N}) \ \wedge \ invariant(t, \boldsymbol{G}_{\hat{\Theta}_N})$ (*Claim A*)

$\longrightarrow maxinv(t, \boldsymbol{G}_{\hat{\Theta}_N})$ (maxinv definition).

□

The above properties indicate that when the endomorphisms induced by the nuisance parameters contains the identity endomorphism, the PAIN statistic is invariant to the nuisance parameters and to the null test parameters, and in the case that there are no null test parameters $\hat{\Theta}_0$, the PAIN statistic is also maximally invariant with respect to the nuisance parameters.

4.2.1 Classification using Parameter Invariant Statistics

Employing the newly developed PAIN statistic, a *parameter invariant classifier* can be constructed in a manner similar to that of the GLRT:

PAIN Classifier. Given a statistic $t \in \{t|PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})\}$, then

$$\hat{y}(\boldsymbol{x}) = \begin{cases} 1 & t(\boldsymbol{x}) \geq \eta \\ 0 & t(\boldsymbol{x}) < \eta \end{cases}$$

is a PAIN classifier.

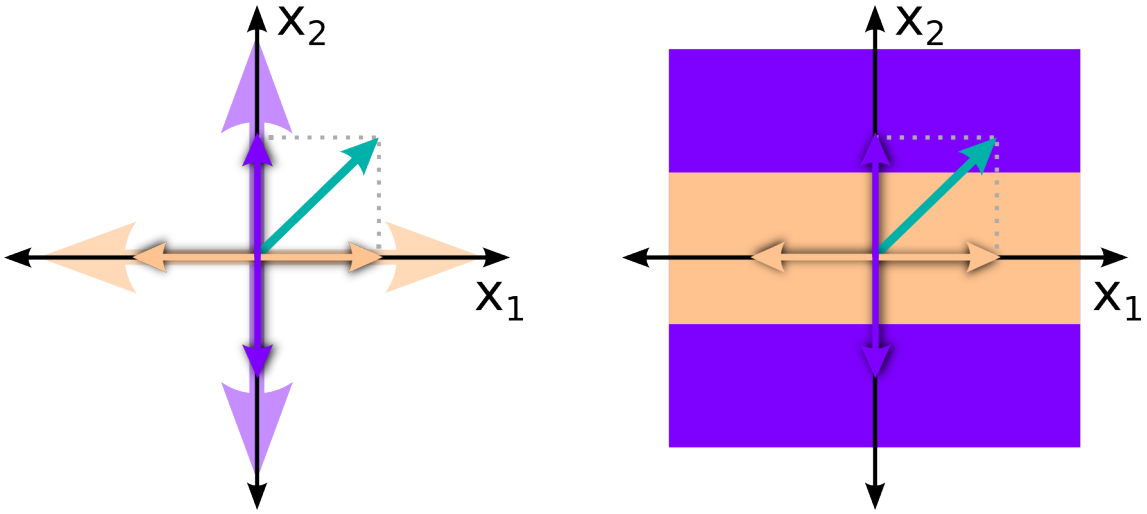
A PAIN classifier uses a PAIN statistic akin to how a UMPI test uses a maximally invariant statistic; however, while the maximally invariant statistic may not always exist, the PAIN statistic can be calculated reliably.

Consider once again the sample system in Equation 3.4, this time with a null and candidate hypotheses defined as follows:

$$\mathcal{H}_0 : (\theta_1, \theta_2) \in \{\mathbb{R}\} \times \{0\}$$
$$\mathcal{H}_1 : (\theta_1, \theta_2) \in \{0\} \times \{\mathbb{R}\}.$$

This scenario is illustrated in Figure 4.1a. Attempting to create a GLRT results in a test over the difference of an unknown non-central chi-squared distribution and a central chi-squared distribution. As the non-central chi-squared distribution contains unknown parameters, establishing a constant rate of false positive is impossible. Instead, we can consider the sets of nuisance parameters

$$\boldsymbol{G}_{\hat{\Theta}_N} = \left\{ g \middle| g\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\}$$
$$\boldsymbol{G}_{\hat{\Theta}_0} = \left\{ g \middle| g\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} c, \forall c \in \mathbb{R} \right\}$$
$$\boldsymbol{G}_{\hat{\Theta}_1} = \left\{ g \middle| g\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} c, \forall c \in \mathbb{R} \right\}$$



(a) Sample system with non-simple, non-overlapping hypotheses.

(b) Solution provided by PAIN classification.

Figure 4.1: Graph of the sample system, updated to show the additional classification challenge presented by non-simple, non-overlapping hypotheses.

and employ Equation 4.2 to produce a PAIN test. This test is in fact identical to the GLRT in Figure 3.2a.

$$\hat{y}(\boldsymbol{x}) = \begin{cases} 1, & \frac{x_2^2}{2} \geq \eta \\ 0, & \frac{x_2^2}{2} < \eta \end{cases}$$

with η chosen to achieve a desired constant rate of false positive. The test is shown in Figure 4.1b.

Figure 4.2 shows the relationship between PAIN classification and the other classification schemes discussed in Chapter 3 in terms of the properties they preserve. At the top is the ideal, the uniformly most powerful (UMP) test with a constant false alarm rate. Below this is a uniformly most powerful invariant test (UMPI). When neither of these two tests can be constructed, the typical choice is to move toward a maximum likelihood-based

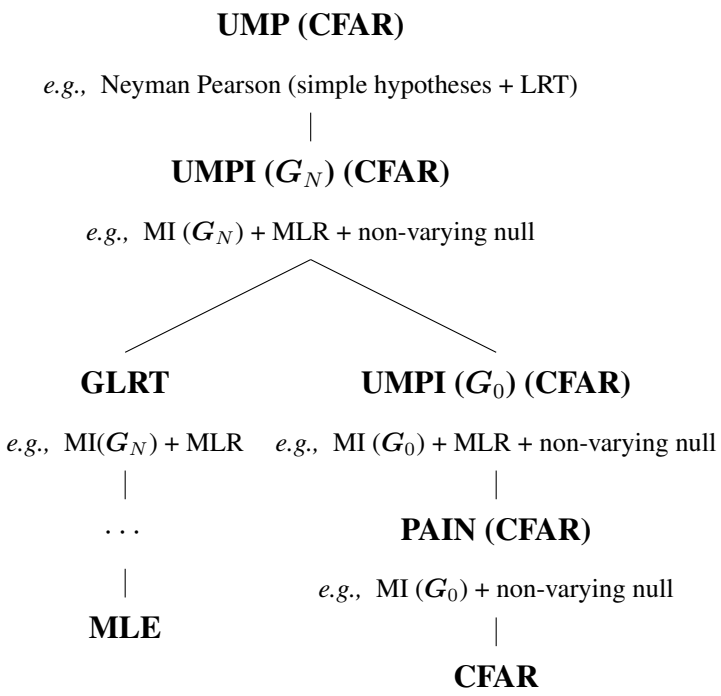


Figure 4.2: A hierarchy of properties preserved by the statistical tests described in this work. Moving down the tree indicates a weaker test. MLR indicates the statistic has a monotone likelihood ratio. MI($\boldsymbol{G}$) indicates a statistic is maximally invariant to $\boldsymbol{G}$. CFAR indicates a statistic achieves a constant false alarm rate over all individuals in the population.

method, such as the GLRT. Alternately, we propose maintaining the constant false alarm rate property via maximal invariance. Typically, this means attempting to calculate a maximally invariant statistic with a monotone likelihood ratio, which allows the construction of a UMP or UMPI test. Our alternative, the PAIN statistic, falls under the umbrella of maximal invariance, but does not impose the restrictions that the existence of an UMPI test does.

Though the PAIN approach does not assume data is distributed in any particular way, we can make strong claims about the PAIN classifier's performance in certain cases. In particular, in a number of common cases, we can show that a PAIN classifier achieves a constant false alarm rate (CFAR). For example, in Lemma 4.3.4, we demonstrate that the PAIN statistic we propose in Section 4.3 is CFAR over Gaussian data. Even in practical cases where data does not conform to a distribution for which CFAR can be proven, PAIN classifiers usually still approach a bounded rate of false alarm over all individuals in the population. In either case, however, achieving this property comes at a cost in true positive rate, P_{TP}, as projecting out noise dimensions frequently results in lost signal (see Section 5.1).

Two Sided PAIN Classifier

To improve the performance of a classifier based on a PAIN statistic, we first observe that switching the hypotheses results in a different PAIN statistic, one that is invariant to a group of transformations imposed by parameters from $\hat{\Theta}_1$ rather than $\hat{\Theta}_0$. We write both statistics as

$$t_0 \in \{t | PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_0})\}$$
$$t_1 \in \{t | PAIN(t, \boldsymbol{G}_{\hat{\Theta}_N}, \boldsymbol{G}_{\hat{\Theta}_1})\}.$$

Then, we can use both of the statistics simultaneously. We refer to this test as a two-sided PAIN classifier:

Two-sided PAIN Classifier. A two-sided PAIN classifier $\hat{y}$ satisfies

$$\hat{y}(\boldsymbol{x}) = \begin{cases} 1 & t_0(\boldsymbol{x}) \geq \eta_0 \wedge t_1(\boldsymbol{x}) \geq \eta_1 \\ 0 & t_0(\boldsymbol{x}) < \eta_0 \wedge t_1(\boldsymbol{x}) < \eta_1 \end{cases} \tag{4.3}$$

	test 0 accepts $\mathcal{H}_0$	test 0 rejects $\mathcal{H}_0$
test 1 accepts $\mathcal{H}_1$	low power	accept $\mathcal{H}_1$
test 1 rejects $\mathcal{H}_1$	accept $\mathcal{H}_0$	bad model(s)

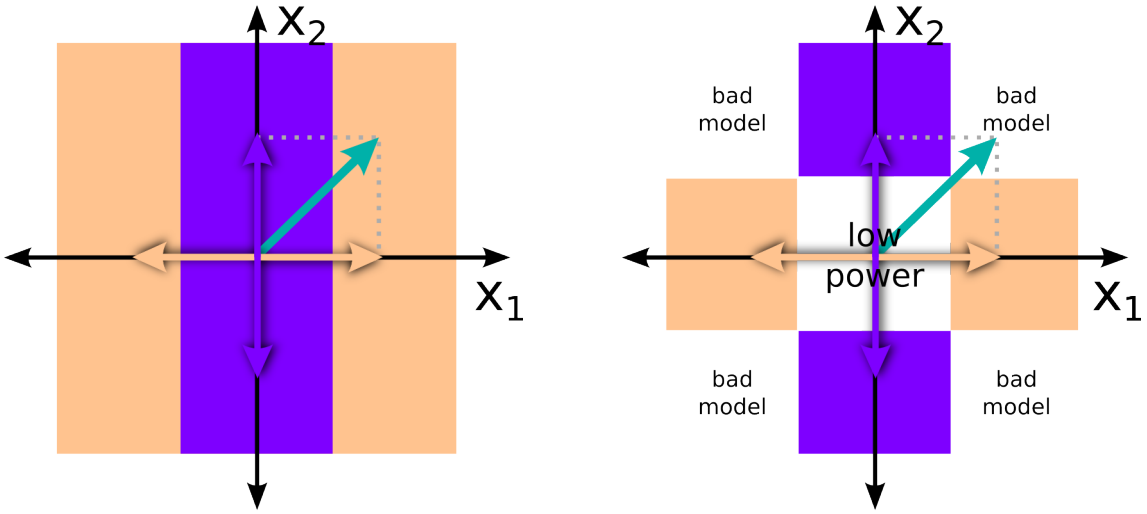
Table 4.1: The decision space of a two-sided PAIN classifier.

The two-sided PAIN classifier tests each hypothesis ($\mathcal{H}_0$ and $\mathcal{H}_1$) independently, using the two statistics t_0 and t_1. The first statistic, t_0, assumes that the null hypothesis $\mathcal{H}_0$ is true, and aims to test the event hypothesis $\mathcal{H}_1$. The second statistic, t_1, assumes the event hypothesis is true and aims to test the null hypothesis.

Note that the two-sided PAIN classifier is under-specified; behavior is not defined for scenarios where $t_0(\boldsymbol{x}) \geq \eta_0$ but $t_1(\boldsymbol{x}) < \eta_1$ or $t_0(\boldsymbol{x}) < \eta_0$ but $t_1(\boldsymbol{x}) \geq \eta_1$. While the parameter-invariant approach utilizes two models (null, $\mathcal{H}_0$ and event, $\mathcal{H}_1$), most real-world scenarios are not binary, and thus it is unlikely that these two models always accurately explain all possible scenarios. In scenarios where test statistics do not agree, the two-sided PAIN classifier cannot provide a definitive outcome, as either both models fit (which means the test has insufficient power for a distinction between models to be made) or neither one does (which means neither model accurately describes the data). Table 4.1 shows the decision space for the two-sided PAIN classifier. The benefit of the two-sided testing approach embodied by the two-sided PAIN classifier is that an inaccurate classification will neither occur from a lack of power in the test, nor because models were inaccurate.

Revisiting the sample system in Section 4.2.1 and switching the hypotheses (visualized in Figure 4.3a) produces a second test over the θ_1 dimension.

$$\hat{y}_2(x) = \begin{cases} 1, & \frac{x_1^2}{2} \geq \eta \\ 0, & \frac{x_1^2}{2} < \eta \end{cases}$$



(a) Sample system solution provided by PAIN classification with hypotheses switched.

(b) Full two-sided PAIN classification solution.

Figure 4.3: Graph of the sample system illustrating result of application of two-sided PAIN classification.

Then, a two-sided test can be constructed. This final test is visualized in Figure 4.3b.

$$\hat{y}(x) = \begin{cases} 1, & \frac{x_2^2}{2} \geq \eta_1 \wedge \frac{x_1^2}{2} \geq \eta_2 \\ 0, & \frac{x_2^2}{2} < \eta_1 \wedge \frac{x_1^2}{2} < \eta_2 \\ ?, & \text{otherwise.} \end{cases}$$

As with the PAIN classifier, we can show that a two-sided PAIN classifier satisfies certain performance guarantees:

Lemma 4.2.2 (Performance of Two-sided PAIN Classifier)**.** *Statistics satisfying Equation 4.3 have bounded rates of false positive (α) and true positive ($1 - \beta$).*

Proof. For the false positive rate, we assume $\mathcal{H}_0$ is true and

$$\begin{aligned} &\hat{y}(\boldsymbol{x}) = 1 \\ \longrightarrow &t_0(\boldsymbol{x}) \geq \eta_0 \\ \longrightarrow &P_{FP} \leq \alpha. \end{aligned}$$

For the true positive rate, we assume $\mathcal{H}_1$ is true and

$$\begin{aligned} &\hat{y}(\boldsymbol{x}) = 1 \\ \longrightarrow &t_1(\boldsymbol{x}) \geq 1 \\ \longrightarrow &P_{TP} \geq 1 - \beta. \end{aligned}$$

□

The PAIN and two-sided PAIN classifiers presented in this section provide performance bounds despite variance in the population by using a parameter invariant statistic, which has the effect of basing classification on the projection of measurements into a non-nuisance subspace, eliminating the effect of nuisance transformation groups. Figure 4.4 illustrates this strategy.

In the following section, we develop a PAIN statistic for the specific model in Equation 4.1, and prove it to be a PAIN statistic invariant to the nuisance transformation group of interest in this work, $\mathcal{G}_y$.

4.3 Establishing Invariance to the $\mathcal{G}_y$ Transformation Set

In this section, we use the general definitions for PAIN statistics in Section 4.2 and apply them to the specific groups of transformations we described in Section 3.3.1: translations and rotations/scalings. By doing so, we create the PAIN statistic for populations with varying parameters that we use in the remainder of the work.

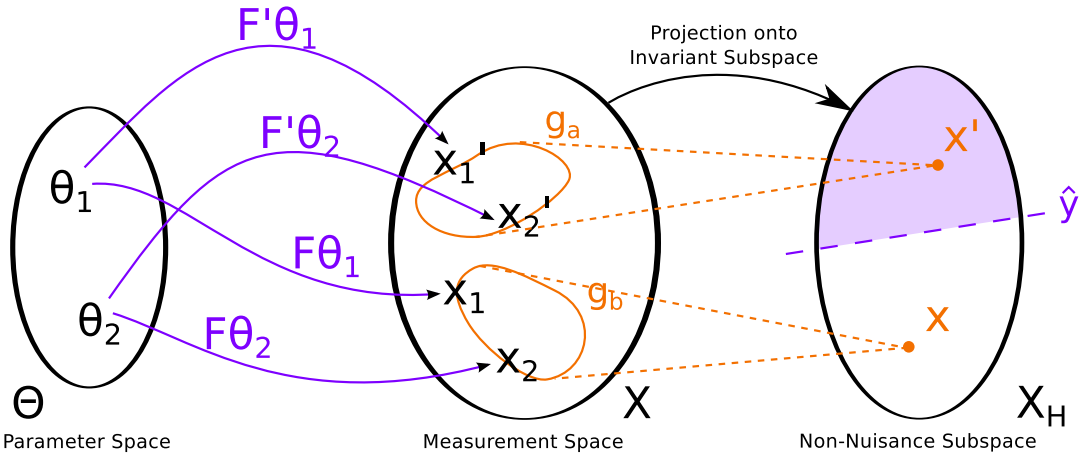


Figure 4.4: An expanded version of Figure 3.3, showing how a classifier $\hat{y}$ can distinguish between model parameter classes while being invariant to nuisance transformations. Eliminating nuisance transformations allows $\hat{y}$ to be unaffected by the irrelevant perturbations in the measurement space caused by variance in parameters.

4.3.1 Statistic Definition

Recall that $\mathcal{I}_0$ and $\mathcal{I}_1$ are the parameter-indexing sets for each of the classes $y \in \{0, 1\}$ defined in Section 3.1.1. Consider the following statistic:

$$t(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0) = \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{r} \right\|}{\left\| \left(\boldsymbol{I} - \hat{\boldsymbol{P}} \right) \boldsymbol{r} \right\|} \tag{4.4}$$

where, assuming $\boldsymbol{F}_{v,j}$ denotes the j-th column of $\boldsymbol{F}_v$,

$$\begin{aligned}
\boldsymbol{r} &= \boldsymbol{P}_0 \boldsymbol{x}_v \\
\hat{\boldsymbol{F}}_v &= \boldsymbol{P}_0 \boldsymbol{F}_v \\
\boldsymbol{P}_0 &\in \{\boldsymbol{P} \mid \boldsymbol{P}\boldsymbol{F}_{v,j} = \boldsymbol{0},\ (\boldsymbol{I} - \boldsymbol{P})\,\boldsymbol{F}_{v,j'} = \boldsymbol{0},\ \forall j \in \mathcal{I}_0,\ \forall j' \in \mathcal{I} \backslash \mathcal{I}_0\} \\
\hat{\boldsymbol{P}} &\in \{\boldsymbol{P} \mid \boldsymbol{P}\hat{\boldsymbol{F}}_{v,j} = \boldsymbol{0},\ (\boldsymbol{I} - \boldsymbol{P})\,\hat{\boldsymbol{F}}_{v,j'} = \boldsymbol{0},\ \forall j' \in \mathcal{I} \backslash \mathcal{I}_1 \backslash \mathcal{I}_0,\ \forall j \in \mathcal{I}_1 \backslash \mathcal{I}_0\}.
\end{aligned}$$

Here, $\boldsymbol{P}_0$ and $\hat{\boldsymbol{P}}$ are *projection matrices*. A projection matrix $\boldsymbol{P}$ has the property that $\boldsymbol{P}^2 = \boldsymbol{P}$. $\boldsymbol{P}_0$ is a null space projection matrix for the columns of $\boldsymbol{F}_v$ selected by the indexing set $\mathcal{I}_0$. $\boldsymbol{r}$ represents the data $\boldsymbol{x}$ projected into that null space, and $\hat{\boldsymbol{F}}_v$ represents the parameter relation matrix projected into that null space. Thus, the statistic is the normalized product of the data projected into the null space of the test parameters under $\mathcal{H}_0$ with the parameter relation matrix projected into the null space of the test parameters under $\mathcal{H}_0$.

In addition to the candidate test statistic, consider the following statistic:

$$t_{n0}(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0) = \frac{\left(\boldsymbol{I} - \hat{\boldsymbol{P}}\right)\boldsymbol{r}}{\left\|\left(\boldsymbol{I} - \hat{\boldsymbol{P}}\right)\boldsymbol{r}\right\|} \tag{4.5}$$

where all terms are defined as above. All discriminatory dimensions for testing the candidate hypothesis is projected out through the $(\boldsymbol{I} - \hat{\boldsymbol{P}})$ term, and thus we refer to t_{n0} as a "nuisance" statistic. While this statistic will provide no benefit for use as part of a PAIN classifier, when combined with t it forms a maximally invariant statistic, and thus satisfies the definition of a PAIN statistic. In order to demonstrate the PAIN property, we first introduce two helpful lemmas:

Lemma 4.3.1 (Equivalence of Cosine and Tangent Forms)**.** *For all $\boldsymbol{x}$ and $\boldsymbol{x}'$ of appropriate dimension, and given some projection matrix $\boldsymbol{P}$,*

$$\frac{\|\boldsymbol{P}\boldsymbol{x}\|}{\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}\|} = \frac{\|\boldsymbol{P}\boldsymbol{x}'\|}{\|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}'\|} \iff \frac{\|\boldsymbol{P}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \frac{\|\boldsymbol{P}\boldsymbol{x}'\|}{\|\boldsymbol{x}'\|}.$$

Proof. Geometrically, we may consider the ratio $\frac{\|\boldsymbol{P}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$ as the cosine of the angle formed between the vector $\boldsymbol{x}$ in the space which $\boldsymbol{P}$ projects into and the remainder of the space, as seen in Figure 3.4c. Similarly, $\frac{\|\boldsymbol{P}\boldsymbol{x}\|}{\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\|}$ is the tangent of this angle. The values of the cosines of the angles between subspaces formed by the vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ are equal if and only if their tangent values are also equal. □

Lemma 4.3.2 (Invariance of Rotations to Null Space Projections)**.** *Given any null space projection matrix $\bar{\boldsymbol{P}} = \bar{\boldsymbol{P}}^2, \bar{\boldsymbol{P}} = \boldsymbol{I} - \boldsymbol{P}$ for $\boldsymbol{P}$ a projection matrix in $\langle \boldsymbol{H} \rangle$ (thus, $\bar{\boldsymbol{P}}$ is in $\langle \boldsymbol{H}^{\perp} \rangle$), and given a rotation endomorphism $g_r \in \mathcal{G}_r$, we can show*

$$\bar{\boldsymbol{P}} g_r(\boldsymbol{x}) = g_r(\bar{\boldsymbol{P}}\boldsymbol{x}).$$

Proof.

$$\begin{aligned}
\bar{\boldsymbol{P}} g_r(\boldsymbol{x}) &= \bar{\boldsymbol{P}}(\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top})\boldsymbol{x} \\
&= \bar{\boldsymbol{P}}(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^{\top} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top})\boldsymbol{x} \\
&= \bar{\boldsymbol{P}}(\boldsymbol{I} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top} - \boldsymbol{U}\boldsymbol{U}^{\top})\boldsymbol{x} \\
&= \bar{\boldsymbol{P}}(\boldsymbol{I} + \boldsymbol{U}(\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{U}^{\top})\boldsymbol{x} \\
&= (\bar{\boldsymbol{P}} + \bar{\boldsymbol{P}}\boldsymbol{U}(\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{U}^{\top})\boldsymbol{x} \\
&= (\bar{\boldsymbol{P}} + \boldsymbol{U}(\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{U}^{\top}\bar{\boldsymbol{P}})\boldsymbol{x} \\
&= (\boldsymbol{I} + \boldsymbol{U}(\boldsymbol{Q} - \boldsymbol{I})\boldsymbol{U}^{\top})\bar{\boldsymbol{P}}\boldsymbol{x} \\
&= (\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^{\top})\bar{\boldsymbol{P}}\boldsymbol{x} \\
&= g_r(\bar{\boldsymbol{P}}\boldsymbol{x})
\end{aligned}$$

□

Using these two lemmas, we can show that the combined statistic conforms to the definition of a PAIN statistic with $\boldsymbol{G}_{\hat{\Theta}_N} = \mathcal{G}_y$.

Theorem 4.3.3 (PAIN Statistic)**.** *The combined statistic $t_c = \begin{bmatrix} t(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0) \\ t_{n0}(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0) \end{bmatrix}$, made up of the two statistics from Equation 4.4 and Equation 4.5, is a PAIN statistic with respect to $\mathcal{G}_y$ for the Problem defined in Section 3.1.*

Proof. We begin by showing $invariant(t(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y)$. For any $g_y \in \mathcal{G}_y$,

$$
\begin{aligned}
t(g_y(\boldsymbol{x}_v); \mathcal{I}_1, \mathcal{I}_0) &= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 g_y(\boldsymbol{x}) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 g_y(\boldsymbol{x}) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 g_t g_s g_r(\boldsymbol{x}) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 g_t g_s g_r(\boldsymbol{x}) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 (g_s g_r(\boldsymbol{x}) + \boldsymbol{H}\boldsymbol{\zeta}) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 (g_s g_r(\boldsymbol{x}) + \boldsymbol{H}\boldsymbol{\zeta}) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 g_s g_r(\boldsymbol{x}) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 g_s g_r(\boldsymbol{x}) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 (\sigma g_r(\boldsymbol{x})) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 (\sigma g_r(\boldsymbol{x})) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} g_r(\boldsymbol{P}_0 \boldsymbol{x}) \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) g_r(\boldsymbol{P}_0 \boldsymbol{x}) \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} (\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top) \boldsymbol{P}_0 \boldsymbol{x} \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) (\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top) \boldsymbol{P}_0 \boldsymbol{x} \right\|} \\
&= \frac{\left\| \boldsymbol{U}^\top \boldsymbol{P}_0 \boldsymbol{x} \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 \boldsymbol{x} \right\|} \\
&= \frac{\left\| \hat{\boldsymbol{P}} \boldsymbol{P}_0 \boldsymbol{x} \right\|}{\left\| (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{P}_0 \boldsymbol{x} \right\|} \\
&= t(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0)
\end{aligned}
$$

and thus $invariant(t(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y)$.

Similarly, we can show $invariant(t_{n0}(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y)$. For any $g_y \in \mathcal{G}_y$,

$$
\begin{aligned}
t_{n0}(g_y(\boldsymbol{x}_v); \mathcal{I}_1, \mathcal{I}_0) &= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_y(\boldsymbol{x})}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_y(\boldsymbol{x})\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_t g_s g_r(\boldsymbol{x})}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_t g_s g_r(\boldsymbol{x})\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 (g_s g_r(\boldsymbol{x}) + \boldsymbol{H}\boldsymbol{\zeta})}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 (g_s g_r(\boldsymbol{x}) + \boldsymbol{H}\boldsymbol{\zeta})\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_s g_r(\boldsymbol{x})}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_s g_r(\boldsymbol{x})\right\|} \\
&= \frac{\sigma(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_r(\boldsymbol{x})}{\left\|\sigma(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 g_r(\boldsymbol{x})\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}}) g_r(\boldsymbol{P}_0 \boldsymbol{x})}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}}) g_r(\boldsymbol{P}_0 \boldsymbol{x})\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})(\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top)\boldsymbol{P}_0 \boldsymbol{x}}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})(\boldsymbol{V} + \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^\top)\boldsymbol{P}_0 \boldsymbol{x}\right\|} \\
&= \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 \boldsymbol{x}}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0 \boldsymbol{x}\right\|} \\
&= t_{n0}(g_y(\boldsymbol{x}_v); \mathcal{I}_1, \mathcal{I}_0)
\end{aligned}
$$

and thus $invariant(t_{n0}(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y)$.

In order to show maximality, we consider both statistics, and let

$$
\begin{aligned}
t_c(\boldsymbol{x}) = t_c(\boldsymbol{x}') &\Longrightarrow t(\boldsymbol{x}) = t(\boldsymbol{x}') \wedge t_{n0}(\boldsymbol{x}) = t_{n0}(\boldsymbol{x}') \\
&\Longrightarrow \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}\right\|} = \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'\right\|} \\
&\wedge \frac{\left\|\hat{\boldsymbol{P}}\boldsymbol{P}_0\boldsymbol{x}\right\|}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}\right\|} = \frac{\left\|\hat{\boldsymbol{P}}\boldsymbol{P}_0\boldsymbol{x}'\right\|}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'\right\|} \\
&\Longrightarrow \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}\right\|} = \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'\right\|} \\
&\wedge \frac{\|\boldsymbol{P}_0\boldsymbol{x}\|}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}\right\|} = \frac{\|\boldsymbol{P}_0\boldsymbol{x}'\|}{\left\|(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'\right\|} \\
&\Longrightarrow \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}}{\|\boldsymbol{P}_0\boldsymbol{x}\|} = \frac{(\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{P}_0\boldsymbol{x}'}{\|\boldsymbol{P}_0\boldsymbol{x}'\|} \\
&\Longrightarrow (\boldsymbol{I} - \hat{\boldsymbol{P}})\left(\frac{\boldsymbol{P}_0\boldsymbol{x}}{\|\boldsymbol{P}_0\boldsymbol{x}\|} - \frac{\boldsymbol{P}_0\boldsymbol{x}'}{\|\boldsymbol{P}_0\boldsymbol{x}'\|}\right) = \boldsymbol{0} \\
&\Longrightarrow \exists g_r \in \mathcal{G}_r \mid \frac{\boldsymbol{P}_0\boldsymbol{x}}{\|\boldsymbol{P}_0\boldsymbol{x}\|} = g_r\left(\frac{\boldsymbol{P}_0\boldsymbol{x}'}{\|\boldsymbol{P}_0\boldsymbol{x}'\|}\right) \\
&\Longrightarrow \exists g_r \in \mathcal{G}_r, \exists g_s \in \mathcal{G}_s \mid \boldsymbol{P}_0\boldsymbol{x} = g_s g_r\left(\boldsymbol{P}_0\boldsymbol{x}'\right) \\
&\Longrightarrow \exists g_r \in \mathcal{G}_r, \exists g_s \in \mathcal{G}_s \mid \boldsymbol{P}_0\boldsymbol{x} = \boldsymbol{P}_0 g_s g_r\left(\boldsymbol{x}'\right) \\
&\Longrightarrow \exists g_r \in \mathcal{G}_r, \exists g_s \in \mathcal{G}_s, \exists g_t \in \mathcal{G}_t \mid \boldsymbol{x} = g_t g_s g_r\left(\boldsymbol{x}'\right) \\
&\Longrightarrow \exists g_y \in \mathcal{G}_y \mid \boldsymbol{x} = g_y(\boldsymbol{x}').
\end{aligned}
$$

Thus,

$$
\begin{aligned}
maximal(t_c(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y) &\wedge invariant(t_c(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y)) \\
&\Leftrightarrow maxinv(t_c(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y) \\
&\Leftrightarrow PAIN(t_c(\boldsymbol{x}_v; \mathcal{I}_1, \mathcal{I}_0), \mathcal{G}_y, \boldsymbol{G}_{\hat{\Theta}_0}).
\end{aligned}
$$

□

4.3.2 Performance when Data is Gaussian

When data is Gaussian, we can show that a PAIN classifier based on the statistic defined above achieves a constant false positive rate, as originally desired.

Lemma 4.3.4 (Performance of PAIN classifier when data is Gaussian)**.** *The PAIN classifier achieves a constant false positive rate when data is Gaussian.*

Proof. First, we rearrange the form of the statistic above when compared to an appropriate threshold[19]:

$$\begin{aligned}
\frac{\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r}}{\boldsymbol{r}^\top \boldsymbol{r}} &\lessgtr \frac{\nu}{1+\nu} \\
(1+\nu)\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\lessgtr \nu(\boldsymbol{r}^\top \boldsymbol{r}) \\
\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} + \nu \boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\lessgtr \nu(\boldsymbol{r}^\top \boldsymbol{r}) \\
\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\lessgtr \nu \boldsymbol{r}^\top \boldsymbol{I} \boldsymbol{r} - \nu \boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} \\
\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\lessgtr \nu(\boldsymbol{r}^\top \boldsymbol{I} \boldsymbol{r} - \boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r}) \\
\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\lessgtr \nu \boldsymbol{r}^\top (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{r} \\
\frac{\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r}}{\boldsymbol{r}^\top (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{r}} &\lessgtr \nu.
\end{aligned}$$

Consider the distributions of the numerator and denominator of the statistic. Given that $x \sim N[\boldsymbol{H}\theta, \sigma I] \in \mathbb{R}^N$ and letting $\hat{\boldsymbol{P}} = UU^\top$ to avoid zero eigenvalues, we can write

$$\begin{aligned}
\boldsymbol{r} &\sim N[0, I] \in \mathbb{R}^N \\
\implies U^\top \boldsymbol{r} &\sim N[0, U^\top U = \boldsymbol{I}_N] \\
\implies \boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r} &\sim \chi^2[N]
\end{aligned}$$

[19] This form of the statistic, the tangent form, has a closed-form solution, an unbounded output value. The original form, the cosine form, has no closed-form solution, but its value is bounded. Note that there is a one-to-one mapping between the two forms of the statistic; thus, we use them interchangeably in this proof.

as the null-space projection re-centers the distribution around zero, and the product of two Gaussians is chi-squared distributed. Likewise, letting $(\boldsymbol{I} - \hat{\boldsymbol{P}}) = VV^\top$,

$$\begin{aligned} \boldsymbol{r} &\sim N[0, I] \in \mathbb{R}^N \\ \implies V^\top \boldsymbol{r} &\sim N[0, V^\top V = \boldsymbol{I}_M] \\ \implies \boldsymbol{r}^\top (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{r} &\sim \chi^2[M]. \end{aligned}$$

By design, $\hat{\boldsymbol{P}}$ and $(\boldsymbol{I} - \hat{\boldsymbol{P}})$ contain no overlapping dimensions, and thus $\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r}$ and $\boldsymbol{r}^\top (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{r}$ are independent. Thus,

$$\frac{\boldsymbol{r}^\top \hat{\boldsymbol{P}} \boldsymbol{r}}{\boldsymbol{r}^\top (\boldsymbol{I} - \hat{\boldsymbol{P}}) \boldsymbol{r}} \sim F[N, M]$$

as the ratio of two independent chi-squared distributions is F distributed, with degrees of freedom equal to the dimension of the numerator and denominator of the ratio. Thus, selecting a value for ν guarantees a constant false positive rate based on the value of the F distribution at that point.

□

As more data is collected, it is not unreasonable to assume that data will asymptotically approach a Gaussian distribution, and thus achieve CFAR. Even when data does not conform precisely to a Gaussian distribution, the statistic tends to produce reasonable bounds on false positive rate.

4.4 Summary

In this chapter, we described linear time-invariant systems. We then described the general form of a PAIN statistic, defined two types of classifiers that utilize these statistics, and described their performance guarantees. We then showed how to build a specific PAIN

statistic for linear time-invariant systems that is invariant to translation, scaling, and rotation.

While parameter invariance is a useful technique for developing detectors on its own, especially in situations when per-patient parameter tuning is impossible, its utility sometimes diminishes in comparison to other possible techniques as the amount of available data grows and interpatient differences can be explicitly quantified through various machine learning approaches. In the next chapter, to improve the overall performance of the PAIN classifier, we combine PAIN statistics with existing machine learning techniques, introducing an approach that identifies columns of $\boldsymbol{F}$ (based on training data) that are likely to be useful for classification while maintaining the performance guarantees provided in this section.

Chapter 5

Parameter Invariant Subspace Selection

In the previous chapter, we showed how to build parameter invariant classifiers that could classify with bounded false positive rates across all individuals in a population. They achieve this through the use of a parameter invariant statistic, which eliminates information in dimensions affected by nuisance transformations, and uses a null space projection to re-center the data around the null hypothesis. This approach is useful because it makes the resulting statistic robust to noisy transformations in undesirable dimensions, makes all the individuals in the population follow the same distribution, and leads to a guaranteed constant rate of false positives that holds over each individual in the population, not just in aggregate (that is, each individual's false positive rate is bounded).

However, this false positive guarantee comes at a cost; for many individuals, their true positive rate is not maximized, as potentially useful signal information may be removed in the process of establishing invariance. Moreover, the parameter invariant statistic does not weight individual dynamical components (that is, the columns of $\boldsymbol{F}$) differently based on the quality of information contained within them. This leads to an averaging of signal-containing modes with noisy modes, resulting in a test with lower power.

In this chapter, we focus on a way to regain some of this lost information using machine learning. We show how parameter invariant statistics over multiple subspaces can be systematically generated, and then propose the use of feature selection to choose which of these statistics "selects" the best subspace. We show that doing so often improves the true positive rate of a classifier, while still providing false positive rate performance guarantees.

5.1 Potential PAIN Performance Improvements

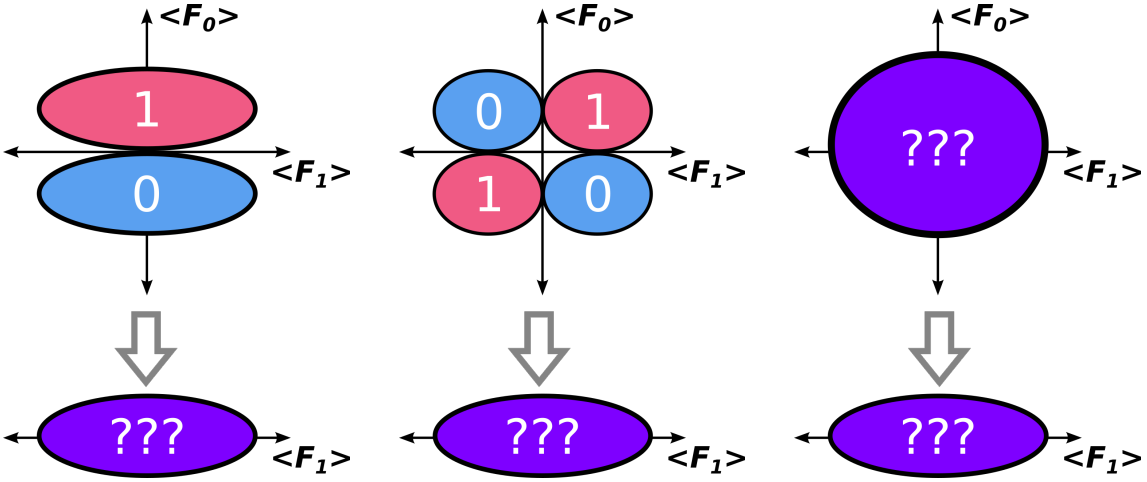
The parameter invariant statistic $t(\boldsymbol{x}_v)$ is defined in Section 4.2 based on the transformation groups $\boldsymbol{G}_0$ and $\boldsymbol{G}_N$ induced by parameter sets $\hat{\Theta}_N$, $\hat{\Theta}_0$, and $\hat{\Theta}_1$. For each individual v with a LTI model as defined in Equation 3.1, these parameter sets define subspaces of the dynamics:

$$\langle \boldsymbol{F}_{v,N} \rangle = \{ \boldsymbol{F}_v \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \Theta_1 \cap \Theta_0 = \hat{\Theta}_N \}$$
$$\langle \boldsymbol{F}_{v,0} \rangle = \{ \boldsymbol{F}_v \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \Theta_0 \backslash \Theta_1 = \hat{\Theta}_0 \}$$
$$\langle \boldsymbol{F}_{v,1} \rangle = \{ \boldsymbol{F}_v \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \Theta_1 \backslash \Theta_0 = \hat{\Theta}_1 \}.$$

In the remainder of the section, we drop the v subscript for simplicity. Note that $\langle \boldsymbol{F}_0 \rangle$ and $\langle \boldsymbol{F}_1 \rangle$ are expected to contain all information that can be used to distinguish between the two candidate hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Also for simplicity, in further discussion, we assume each of these subspaces is an orthogonal basis; if not, we may use QR decomposition to make them so.

There are three cases in which the "re-centering" null space projection over $\langle \boldsymbol{F}_0 \rangle$ may result in a suboptimal test:

- Dimensions in $\langle \boldsymbol{F}_N \rangle$ that contain signal information under $\mathcal{H}_1$ but contain no signal information under $\mathcal{H}_0$ will be eliminated by the null space projection of $\langle \boldsymbol{F}_0 \rangle$, though they would otherwise have improved the power of the test statistic.



(a) Loss of information from $\langle \boldsymbol{F}_0 \rangle$ dimensions.

(b) Loss of information from $\langle \boldsymbol{F}_1 \rangle$ dimensions.

(c) Failure to remove noise from $\langle \boldsymbol{F}_1 \rangle$ dimensions.

Figure 5.1: An illustration of the three scenarios in which a null space projection may result in a suboptimal test. Useful hypothesis-distinguishing signal is visualized using ovals in certain subspace regions. Dimensions with signal in null and candidate hypothesis subspaces are displayed as blue and pink, respectively. The purple question-mark regions indicate noise (a mixture of signal in both directions).

- Dimensions in $\langle \boldsymbol{F}_N \rangle$ that contain signal information under $\mathcal{H}_0$ but contain no signal information under $\mathcal{H}_1$ will be eliminated by null space projection of $\langle \boldsymbol{F}_0 \rangle$. This information could have improved the power of the test statistic if the roles of $\mathcal{H}_0$ and $\mathcal{H}_1$ were reversed, which is relevant for the two-sided PAIN classifier.

- Dimensions in $\langle \boldsymbol{F}_1 \rangle$ that are not included in $\langle \boldsymbol{F}_N \rangle$ but that do not contain any signal will contribute noise to the test statistic that will be incorrectly considered signal.

These three cases are illustrated in Figure 5.1. All of these cases lead to noisy modifications of the value of the statistic, leading to degradation in performance. In this chapter, we

focus on the third problem, and investigate methods for increasing the likelihood that $\langle \boldsymbol{F}_1 \rangle$ contains only dimensions with high signal-to-noise ratios.

5.2 Choosing a Subspace using Generated Features

In light of the suboptimalities in using $\hat{\Theta}_N = \Theta_1 \cap \Theta_0$ enumerated in the previous section, it may be tempting to try to define Θ_N by hand. As it is often difficult to know *a priori* exactly which dimensions of $\langle \boldsymbol{F}_1 \rangle$ do not contain signal information, attempting to appropriately choose nuisance parameters by hand can be difficult.

In the remainder of this chapter, we propose *learning* which nuisance parameters to eliminate directly from the available data. These parameters thus define the eliminated subspace $\langle \boldsymbol{F}_N \rangle$. To do so, we make a simple modification to the usual technique of building a PAIN classifier: in lieu of calculating a single statistic across all the dimensions of the data, we create a set of different statistics by individually projecting out each dimension of the space. Then, we show how to use these statistics as raw features in a machine learning classifier. Feature selection, or a classifier with a regularization parameter, can then be used to preserve only those features that lead to good performance, corresponding to the inclusion of only those dimensions of $\langle \boldsymbol{F}_1 \rangle$ that contain a high signal-to-noise ratio in the statistic.

5.2.1 Generating PAIN Features

Let $\mathcal{P}(\mathcal{I}_1)$ represent the power set of the index set $\mathcal{I}_1$, and let $\mathcal{P}(\mathcal{I}_1)_j$ represent the j-th set in the power set. We generate a set $T_{full} = \{t_1 \dots t_j \dots t_{\|\mathcal{P}(\mathcal{I}_1)\|}\}$ of features where

$$t_j = t(x_v, \mathcal{P}(\mathcal{I}_1)_j, \mathcal{I}_0).$$

This feature set can then be used with a feature selection technique. A naïve option would be to select those features which achieve a true positive rate greater than some specified value γ over a test data set when used alone as part of a PAIN classifier $\hat{y}_t(\boldsymbol{x})$, *i.e.,*

$$T_{select} = \{t | t \in T_{full} \wedge P_{TP}(\hat{y}_t) > \gamma\}. \tag{5.1}$$

Alternately, the full set of features could conceivably be used with a classifier that imposes a regularization penalty on parameter weights. An appropriate penalty metric, such as the L1 norm, would drive the coefficients of those $t_j \in T_{full}$ that contain less useful signal information toward zero as the amount of available training data increases.

The major problem with this technique occurs when $\langle \boldsymbol{F} \rangle$ has a large number of dimensions, causing the feature set to become intractably large. Searching all combinations of $\theta_j \in \boldsymbol{\theta}$ requires producing $\|\mathcal{P}(\mathcal{I}_1)\| = 2^N$ statistics. Due to the exponential increase in features, generating all of them is infeasible for problems with more than a few dimensions in $\langle \boldsymbol{F} \rangle$.

Greedy PAIN Feature Generation

Instead, to generate a feature set of a more manageable size, we propose using a greedy column selection method to generate a number of features equal to the number of columns in $\boldsymbol{F}$ (and thus the number of dimensions in $\langle \boldsymbol{F} \rangle$). To do so, we iteratively construct an index set J, beginning with an empty set and adding, at each step, the j' where

$$j' = \max_{j \in \mathcal{I}_1 / J} P_{TP}(\hat{y}_t | t = t(x_v, \{J, j\}, \mathcal{I}_0)).$$

In effect, this process "selects" the column of $\boldsymbol{F}$ that produces the most powerful statistic when added to the current set of selected columns. That statistic $t(x_v, \{J, j'\}, \mathcal{I}_0)$ is then added to the feature set T.

Algorithm 1 PAIN Greedy Feature Generation

procedure GENERATEPAINFEATURES($V, \mathcal{I}_0, \mathcal{I}_1$)
    Initialize J to be an empty array
    Initialize T to be an empty array
    for $i \leftarrow 1, size(\mathcal{I}_1)$ **do**
        for $j \in \mathcal{I}_1 / J$ **do**
            Let $J' = [J, j]$
            Let $c[j] = power(V, t(x_v, \mathcal{I}[J'], \mathcal{I}_0))$
        end for
        Let $\hat{j} = \max_j(c)$
        Let $J = [J, \hat{j}]$
        Let $t_i = t(x_v; J, \mathcal{I}_0)$
        Let $T[i] = t_i$
    end for
end procedure

As before, the P_{TP} of each column is calculated over the population as the feature that maximizes the number of true positives constrained to a set false positive maximum. Note that once a column is selected, it is stored and used in all future power estimates (*i.e.,* the first to j-th selected columns are included in the PAIN statistics when calculating the $j + 1$ most powerful column). We iteratively include columns in this way to prevent the selection of multiple columns that share a similar "direction" in the subspace, unless the increase in power is significant in comparison to the other available columns. This procedure is summarized in Algorithm 1. In both the PAIN and two-sided PAIN scenarios, we select only one column at a time to avoid scenarios in which mutual information shared between columns inflates a column's individual power. Selecting columns that have large power

levels primarily due to mutual information could result in decreased performance.

In order to create the t_0 statistics required for the two-sided PAIN classifier, the same algorithm can be applied to generate PAIN statistics by switching the hypotheses and calculating

$$j' = \max_{j \in \mathcal{I}_0 / J} P_{\boldsymbol{\theta}}(\hat{y}_t = 0 | t = t(x_v, \{J, j\}, \mathcal{I}_1))$$

when selecting indexes. In the two-sided PAIN scenario, we avoid selecting $\boldsymbol{F}$ columns simultaneously for both hypotheses, as this could result in decreased performance.

Both the standard feature selection technique and the greedy feature generation algorithm described above are equivalent to searching the space of statistics to discover a subspace projection that achieves good performance. In the next section, we show that this reduced feature set is asymptotically guaranteed to include the maximally invariant statistic, if it exists.

5.2.2 Performance and Properties

In this section, we discuss the computational complexity of the PAIN feature generation algorithm, and discuss in what sense the sets of PAIN features T_{full} and T are optimal.

Computational Complexity

Given there are J columns in $\boldsymbol{F}$, the worst-case complexity of PAIN feature generation is $O(J^5)$. The worst-case complexity arises from calculating the null space of $\boldsymbol{F}$ in the PAIN statistic, an $O(J^3)$ operation, and must be completed $O(J^2)$ times during the feature generation loop. While the worst-case complexity is $O(J^5)$, this complexity can be reduced by exploiting the structure of $\boldsymbol{F}$ in certain scenarios (*e.g.,* when $\boldsymbol{F}$ corresponds to a linear system). Moreover, we note that the number of columns is often significantly fewer than

the number of dimensions in the measurement $\boldsymbol{x}$; thus, the computational complexity is likely to be acceptable in most applications.

Optimality

In this section, we aim to demonstrate that under appropriate conditions (specifically, when used in conjunction with an appropriate classifier or feature selection technique), the previously described PAIN feature set will meet or exceed the performance of a UMPI test, if one exists.

Consider, first, that the existence of a UMPI classifier implies the existence of an invariant statistic that is maximal. Trivially, this statistic is included in T_{full}. We can show that it must also be among those included in T.

Lemma 5.2.1 (Asymptotic Optimality of PAIN features)**.** *If a UMPI classifier exists, then a UMPI classifier can be constructed using some statistic in the greedily constructed PAIN feature set T:*

$$\exists \hat{y}_t, \;\; UMPI(\hat{y}_t, \boldsymbol{G}) \longrightarrow \exists \eta, \exists t_j \in T, UMPI(t_j \geq \eta, \boldsymbol{G}).$$

Proof. The iterative column selection ensures that the parameters θ_j (and thus the columns of $\boldsymbol{F}$) representing dimensions that contain discriminatory information (that is, those that are non-noise) are selected first. As more individuals are included in the population, discriminatory dimensions will be more definitively selected before the non-discriminatory dimensions. The statistic that includes all discriminatory dimensions while excluding all non-discriminatory dimensions is the maximally invariant statistic, and thus it will be contained in the feature set. □

Therefore, with enough data, if a UMPI statistic exists, it will be included in the PAIN feature set. Our second step shows that a chosen feature selection and machine learning

technique will achieve the performance of the UMPI statistic, if that statistic is included in the feature set. This step is dependent on the choice of feature selection technique and classifier, but, to illustrate, consider the simple feature selection technique in Equation 5.1 applied to the greedy PAIN feature set T.

Lemma 5.2.2 (Asymptotic Optimality of Naïve Feature Selection over PAIN features)**.** *If the UMPI statistic $\hat{t}$ is in T, the UMPI statistic is in T_{select}, or T_{select} is empty.*

Proof. By contradiction, assume $\hat{t}$ is in T, but

$$T_{select} \neq \varnothing \wedge \hat{t} \notin T_{select}$$

which implies that

$$T_{select} \neq \varnothing \longrightarrow \exists t \in T_{select} | t \neq \hat{t} \wedge \hat{y}_t \in \mathcal{Y}_\alpha \wedge P_{TP}(\hat{y}_t) > \gamma$$

and

$$\begin{aligned} \hat{t} \notin T_{select} \longrightarrow & P_{TP}(\hat{y}_{\hat{t}}) \leq \gamma \\ \longrightarrow & P_{TP}(\hat{y}_{\hat{t}}) < P_{TP}(\hat{y}_t). \end{aligned}$$

But, from the definition of UMPI,

$$UMPI(\hat{y}_{\hat{t}}) \Longleftrightarrow P_{TP}(\hat{y}_{\hat{t}}) \geq P_{TP}(\hat{y}') \; \forall \hat{y}' \in \mathcal{Y}_{\alpha, invariant}$$

which presents a contradiction. Thus, either $T_{select} = \varnothing$ or $\hat{t} \in T_{select}$. □

A feature selection scheme that attempts to choose features that perform well will likely select the UMPI statistic as a feature. However, as a UMPI statistic rarely exists for most problems, use of the suggested feature set is likely to provide better performance than use of any one single PAIN statistic. In the next section, we use a synthetic data set to demonstrate the performance boost that the PAIN features provide.

5.3 Simulated Performance

In this section, we create a synthetic data set by creating two classes of LTI systems, and instantiating a number of "individuals" with random parameters from each class. We then use the described parameter invariant features to build a classifier over this data, and compare the results to the GLRT and to an ARMAX model automatically fitted to the data using MLE. We show that though both ARMAX and GLRT achieve good population-level performance, they can cause a very large number of false positives over some individuals in the population. The proposed parameter invariant techniques achieve slightly lower detection rates when tuned to match the average false positive rate of the other techniques, but achieve this rate of false positive over every individual in the population. Additionally, our proposed data-driven PAIN feature set only boosts the detection rate of the PAIN-based classifier.

5.3.1 Problem and Model Description

In order to create data that could serve as a reasonable stand-in for patient data, we chose to use linear models. As described in Section 4.1, linear models serve as good approximations for most real-world systems over relatively small windows and have been analyzed extensively, with strong results developed for various classes of *linear time invariant* (LTI) systems.

We note that while the LTI system can map to the general form in Equation 4.1, it is more classically considered an autoregressive moving average model with exogenous inputs (ARMAX); however, since the PAIN approach is motivated by geometric isolation of nuisance parameters rather than their statistical estimation, the condensed "regression" representation of the ARMAX model is sufficient.

The $\boldsymbol{F}_v$ matrix generated from the LTI system contains inputs which must be specified

by the user. In this formulation, the problem defined by Equation 3.2 requires distinguishing between two different parameter combinations representing (in the LTI systems case) different inputs and dynamics.

5.3.2 Data Set Details

To create the data set we generated 100 random sets of model parameters for a third-order LTI model. To explain how this was done, recall the form of the LTI model in Equation 4.1:

$$\boldsymbol{x}_v = \boldsymbol{A}_v \boldsymbol{a}_v + \sum_{l=1}^{L} \boldsymbol{B}_{v,l} \boldsymbol{b}_{v,l} + \sigma_v \boldsymbol{n}.$$

For a third-order model, parameter matrices contain three columns, with rows of the matrices representing measurements over a time window. This is the *time domain* representation of the system. We can equivalently characterize the system in the *frequency domain* by considering the *transfer function*, the relationship between the inputs (captured in the matrix $\boldsymbol{B}$) and the outputs ($\boldsymbol{x}$) assuming zero initial conditions and zero-point equilibrium. The transfer function $H(z)$ is the linear mapping of the Laplace transform of the input $\boldsymbol{B}(z) = \mathcal{L}\{\boldsymbol{B}\}$ to the Laplace transform of the output $\boldsymbol{X}(z) = \mathcal{L}\{\boldsymbol{x}\}$:

$$H(z) = \frac{\boldsymbol{X}(z)}{\boldsymbol{B}(z)} = \frac{\mathcal{L}\{\boldsymbol{x}\}}{\mathcal{L}\{\boldsymbol{B}\}}$$

where z is the z-transform. The resulting function is a ratio of polynomials which can each be factored, which will reveal *zeros* in the numerator and *poles* in the denominator. A system is fully characterized by the values of its poles and zeros; this is the *pole-zero* representation of the system.

In order to generate patients from two distinct classes, we randomly generated two sets of poles and zeros to represent null and candidate hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Then, we added random covariance noise to create new sets of parameters from these baselines. We varied

the magnitude of this noise to represent different amounts of interpatient variation; variance was set to either low (0.1) or high (1.0). Each of these generated parameter sets would be treated as a different individual patient. Randomly generating systems using the pole-zero representation was advantageous because it allowed us to ensure in advance that systems created (in particular, their poles) were *stable*; that is, every bounded input to the system would produce a bounded output, a desirable property for a system intended to mimic a biological patient.

Then, for each of these new "patients," we generated 1000 "tests," with each test made up of 25 sequential samples generated by the parameterized model. Ten percent of patients were used for training the classifiers. Class distribution was balanced between the two classes. We learned a threshold for each classifier corresponding to a population-level false positive rate of 0.2. To generate the MLE/ARMAX, we used Matlab's built-in ARMAX function to fit a second-order ARMAX model to the training data.

In all cases, learned coefficients were then used as features in a logistic regression classifier. We chose logistic regression as a convex surrogate for 0-1 loss with good convergence properties (Rosasco et al., 2004). We also compared against the GLRT, as described in Section 3.2. Note that the GLRT is an optimal test when parameters do not vary across the population.

5.3.3 Performance Results

In Figures 5.2a and 5.2b, we see the population-level ROC curves for each of the three types of classifiers applied to a population with low intersubject variance and high intersubject variance, respectively. The MLE achieves the highest true positive/false positive trade-off in both cases, followed by GLRT. In Figures 5.2c and 5.2d, however, we see the false positive rate for each individual plotted. In the case of low variance (5.2c), the MLE and

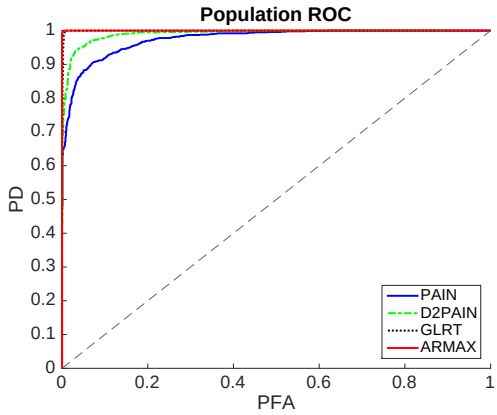
Classifier	Low Variance	High Variance
ARMAX/MLE	0.000974	0.0341
GLRT	0.002	0.0462
PAIN	**0.00015833**	**0.0001958**
dd-PAIN	0.0001667	0.000214

Table 5.1: The variance in false positive rate for four different classifiers over a simulated data set. Variance in interpatient parameters was set to either low (0.1) or high (1.0) for each of the two subject classes.

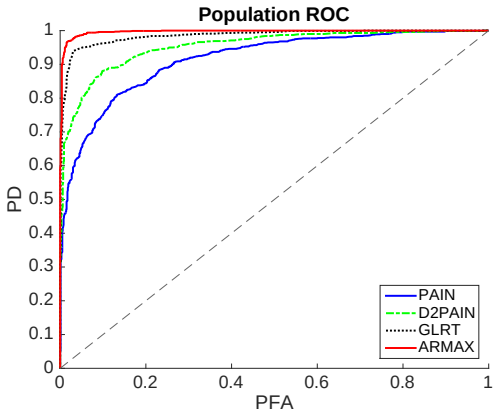
GLRT perform the best and are able to maintain a relatively low level of false positives. However, in the case of high inter-patient variance (5.2d), the variance of the false positive rate for the MLE classifier expands dramatically, meaning a number of patients experience very high rates of false positives. The PAIN approach is able to maintain its detection rate along with a consistent false positive rate. The data-driven selection technique we propose improves that detection rate while still maintaining the limited false positive rate.

Though MLE performs better on a population level, at the level of the individual, it achieves this performance at the expense of some individuals in the population. On these individuals, it achieves extremely poor performance (producing huge numbers of false positives). Table 6.2 shows the variance of the false positive rates for each of the three classification techniques. Note that the classifier based on the PAIN statistic consistently exhibits far less variability in false positive rate than the ARMAX/MLE- or GLRT-based classifiers.

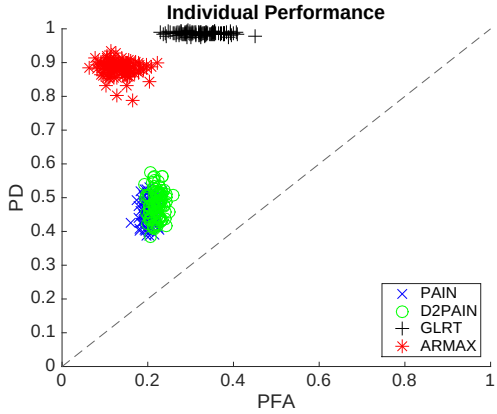
The PAIN approach maintains performance over all individuals even when patient variability increases, sacrificing some detection performance for a near-guaranteed rate of false



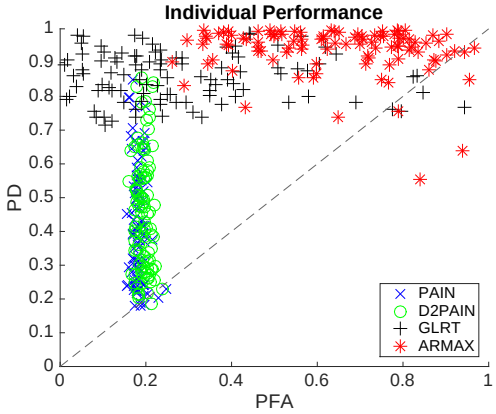
(a) ROC Curve comparing classifier performance over 100 subjects with low variance.



(b) ROC Curve comparing classifier performance over 100 subjects with high variance.



(c) Individual performance (false positive rate plotted against true positive rate) for 100 patients with 1000 tests each with low variance.



(d) Individual performance (false positive rate plotted against true positive rate) for 100 patients with 1000 tests each with high variance.

Figure 5.2: Graphs comparing performance of ARMAX, GLRT, and PAIN classifiers over a simulated subject population with low/high intersubject variance.

positives across all patients. The low false positive variability in the PAIN approach illustrates the performance claim in Lemma 4.3.4, where the PAIN statistic asymptotically

converges to a constant false alarm rate. The PAIN approach does not achieve exactly constant false alarm rate in Figure 5.2 as our simulation method involves repeated sampling of many different runs.[20]

Small sample sizes and high interpatient variance are common conditions in many real-world applications, in particular when data is collected from sensors measuring a physical system. In these situations, dramatic jumps in false positive rate for certain individuals are often highly undesirable. In the medical domain, large numbers of false alarms can lead to alarm fatigue, and erratic changes in the performance of a system can make it difficult for clinicians to make informed decisions about the care of the patient. PAIN classification, therefore, with its predictable performance, is highly applicable to real-world scenarios. In the following chapter, we apply PAIN classification to three such real-world medical scenarios.

5.4 Summary

In this chapter, we described the scenarios in which information is lost through the parameter invariant process, leading to suboptimal detection rates. We proposed the application of a simple machine learning technique to greedily generate a subspace feature set that would allow some of this lost information to be regained, boosting classifier detection performance. We then described a simulation which demonstrated this improvement in performance.

[20]It is worth noting that the low rate of detection for the PAIN and data-driven PAIN classifiers in the low-variance case shown in Figure 5.2a is dependent on the parameters randomly chosen in the simulation. Note that in Figure 5.2b, the detection rates vary widely as the parameter values do; undoubtedly, if the parameters for all patients in the low-variance simulation were centered around one of the patients for whom the PAIN classifier happened to perform well, the classifier would appear to perform better.

Chapter 6

Applications

Each section of this chapter demonstrates the utility of the proposed PAIN-statistic-based classification technique (highlighting in particular the maintenance of a constant false positive rate) by applying the technique to a real, challenging detection problem from the medical domain. Table 6.1 summarizes some key details of each of these testing scenarios to allow for easier comparison.

In Section 6.1, we develop a technique for using photoplethysmogram signals to detect when a patient begins to experience hypovolemia (low blood volume). We use data from a large publicly available database of physiologic signals, the Physionet MIMIC II database (Goldberger et al., 2000; Saeed et al., 2011) to show that a parameter invariant classifier can distinguish between hypovolemic and non-hypovolemic patients at or before the onset of hypovolemia as determined by clinicians.

In Section 6.2, we use an FDA-approved diabetic patient model to generate glucose data including periodic meals for a large number of simulated patients. We apply the PAIN-classifier technique to the glucose data, and show that it can reliably detect the changes in glucose associated with a meal with fewer false alarms than state-of-the-art detection

Domain	Signal	Assumed System Order	Subjects
Hypovolemia	PPG	3	25
Diabetic Meals	Glucose	5	200
Pulmonary Shunt	S_pO_2	2	292

Table 6.1: The details of each of the applications of PAIN to medical domains presented in this work.

techniques.

In Section 6.3, we use respiratory data collected from infants undergoing pulmonary lobectomies to detect unintended pulmonary shunts. We are able to detect shunt events in a timely fashion, before more traditional methods of detection would alert clinicians, while producing only limited numbers of false alarms.

For each of these sections, we first describe the problem and unique features of the problem that pertain to the design of the model in use. We then provide a section on simulation details (including values of parameters used) and, finally, describe analysis results.

6.1 Detection of Hypovolemia in Critically Ill Patients

In this section, we present a technique for utilizing PPG waveforms to monitor patients' fluid status, in particular for detecting hypovolemia. We apply preprocessing to remove signal artifact, then apply novel parameter invariance techniques to create a statistic that is invariant to other common forms of signal noise. We evaluate the detector over a set of both hypovolemic and non-hypovolemic ICU patients from the Physionet MIMIC II database (Goldberger et al., 2000; Saeed et al., 2011).

6.1.1 Medical Context

The photoplethysmogram (PPG) is an optical measurement used to detect changes in blood volume in the microvascular tissue bed. Devices that measure PPG (such as pulse oximeters) contain a light emitting diode and an optical sensor. Light is emitted into flesh and either reflected off bone and back to the sensor, or transmitted directly through the flesh and into the sensor. The scattering, absorption, reflection, and fluorescence of the biological tissue impacts the amount of light reabsorbed by the sensor (Allen, 2007).

PPG has seen widespread clinical application, as it is noninvasive and can be used to measure many different aspects of cardiovascular function, most commonly pulse rate and tissue oxygenation (Allen, 2007). The signal also contains information about vascular distensibility, cardiac arrhythmia, systolic blood pressure, respiratory variability (Shelley, 2007), and, notably, blood volume. Recorded pulses bear a direct relationship with perfusion (the delivery of blood to the capillary bed), as larger blood volumes produce larger attenuation in the light source (Allen, 2007).

Patients who present to emergency rooms with trauma, patients undergoing surgery, and post-operative patients in intensive care units frequently suffer from hemorrhage. Persistent internal hemorrhage can, over time, cause a decrease in the volume of blood in the circulatory system, a condition known as *hypovolemia* (Gutierrez et al., 2004). Hypovolemia is common among post-operative patients. Bleeding-related complications (such as rapidly fatal hypovolemic shock) are a major cause of prolonged length of stay and death in hospitals (Moscucci et al., 2003; Stokes et al., 2011).

Assessment of decreasing blood volume is one of the most difficult tasks in current clinical medicine (Marik et al., 2009), as the body's hemodynamic compensation mechanisms can mask changes in most of the vital signs that would traditionally be used to assess volume status (Convertino et al., 2009, 2011; Gutierrez et al., 2004; McGee et al., 1999;

Zöllei et al., 2013). Patient fluid inputs and outputs are closely monitored for changes (Bartels et al., 2013), as common medical practice holds that these metrics may reflect related changes in blood volume. These changes only occur, however, after significant blood is lost (Stewart et al., 2014).

There have been numerous attempts to use the PPG waveform to noninvasively monitor fluid status and detect hypovolemia, with the goal of providing earlier, more accurate, less invasive detection; see Alian et al. (2011); Cannesson et al. (2008); Forget et al. (2010); Loupec et al. (2011); Pizov et al. (2010); Sahni (2012); Stewart et al. (2014) for a number of examples. These studies have used compensatory reserve index (CRI), variations in pulse oximeter waveform amplitude (ΔPOP), and/or pleth variability index (PVI). Results show promise, but predictive values seem to vary substantially between studies (Antonsen and Kirkebøen, 2012; Cannesson et al., 2010), and few studies gauge performance in critically ill patients.

While PPGs contain large amounts of information about a patient's cardiovascular function, they can be difficult to use because they often experience large amounts of artifact. The PPG sensor is sensitive to movement and orientation against the skin; small shifts can significantly impact the measured intensity of light, and subsequently the accuracy of the data. Detection is made still more challenging by interpatient variability. Patients' blood volumes and compensatory mechanisms vary, and the body's response to blood loss varies based on the severity and location of the bleed. Both of these problems make utilizing PPG for fluid management a non-trivial challenge ripe for the application of PAIN statistics.

6.1.2 Methodology

This section describes how we construct a PAIN-statistic-based detector over a sampled average of the PPG waveform over time to test for fluid loss. We first addresses modeling

the PPG waveform, describe design of the PAIN test, then provide a description of the algorithm used for PPG waveform preprocessing.

PPG Trend Modeling

In this subsection we develop models representing the PPG trends under normal ($\mathcal{H}_0$) and hypovolemic ($\mathcal{H}_1$) scenarios. The PPG waveform is composed of a static DC signal related to the absorption of light by the non-blood components of the body (*i.e.,* bone, muscle, skin, etc.) and a dynamic AC signal corresponding to the blood-related absorption. In McGrath et al. (2011), it is shown that immobilized healthy patients experiencing central blood volume loss have AC PPG signals which tend to decrease in amplitude and pulse width such that the average value of the PPG waveform, over the respiratory cycle, decreases.

Let $PPG(t)$ be the value of the PPG waveform at time $t \geq 0$, and let $\overline{PPG}(n)$ with $n \in \mathbb{N}$ represent the n-th sampled average of the PPG waveform over a time window of T seconds. (Note that $\overline{PPG}(n)$ is a function of $PPG(t)$ over the domain $(n-1)T \leq t < nT$) Then we can model the trend of $\overline{PPG}$ in a hypovolemic scenario as

$$\mathcal{H}_1 : \overline{PPG}(k+1) = \alpha_1 \overline{PPG}(k) + \beta_1 + \sigma_1 n(k) \tag{6.1}$$

with $\beta_1 > 0$ and $0 < \alpha_1 < 1$ proportional to the fluid loss volume and fluid loss rate parameters, respectively. σ_1 represents the variance of the noise. Due to varying patient physiology and condition, the parameters $\alpha_1, \beta_1, \sigma_1$ are unknown.

Intensive care patients are rarely immobilized and healthy; thus, the average of the PPG waveform of non-hypovolemic patients tends to drift over time. Rather than attempt to model all possible physiological scenarios that explain drifts in the PPG waveform, we model $\overline{PPG}$ under non-hypovolemic conditions (the null hypothesis) as a Brownian mo-

tion,

$$\mathcal{H}_0 : \overline{PPG}(k+1) = \overline{PPG}(k) + \sigma_0 n(k) \tag{6.2}$$

where $\sigma_0 n(k) \sim \mathrm{N}[0, \sigma_0^2]$ denotes the input noise. Consistent with the event model in Equation 6.1, the noise parameter σ_0 is unknown.

The models developed in this subsection use medical trends to describe the dynamics of the PPG mean. Note that the PAIN design approach only requires models that capture the general trend of the PPG signal; they need not be an accurate first-principles representation of true hemodynamics.

Parameter-Invariant Test Design

As described in the previous section, the parameters of the models Equation 6.1 and Equation 6.2 are unknown, and vary over each patient. In this section, we calculate the PAIN statistic that will produce our PAIN classifier.

To develop the statistic, we assume a testing window of K samples and write $y(k) = \overline{PPG}(k) - \overline{PPG}(k-1)$. We rearrange to obtain a time-concatenated model under each hypothesis that can be written as

$$\begin{aligned} \mathcal{H}_0 &: \boldsymbol{y}_k = \sigma_0 \boldsymbol{n} \\ \mathcal{H}_1 &: \boldsymbol{y}_k = \boldsymbol{f}_k(\alpha_1 - 1) + \mathbf{1}\beta_1 + \sigma_1 \boldsymbol{n} \end{aligned} \tag{6.3}$$

where

$$\boldsymbol{y}_k = \begin{bmatrix} y(k-K) \\ \vdots \\ y(k) \end{bmatrix} \text{ and } \boldsymbol{f}_k = \begin{bmatrix} \overline{PPG}(k-K-1) \\ \vdots \\ \overline{PPG}(k-1) \end{bmatrix}.$$

We then construct a sufficient statistic for the hypothesis testing problem in Equation 6.3 that is invariant to the effect of the unknown parameters as

$$t(\boldsymbol{y}_k) = \frac{\mathbf{1}^\top \boldsymbol{P}_k \boldsymbol{y}_k}{\sqrt{\mathbf{1}^\top \boldsymbol{P}_k \mathbf{1}} \sqrt{\boldsymbol{y}_k^\top \boldsymbol{P}_k \left(\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right) \boldsymbol{P}_k \boldsymbol{y}_k}}$$

where

$$\boldsymbol{P}_k = \boldsymbol{I} - \frac{\boldsymbol{f}_k \boldsymbol{f}_k^\top}{\boldsymbol{f}_k^\top \boldsymbol{f}_k}.$$

In words, we design invariant to the effect of α by projecting onto the null space of $\boldsymbol{f}_k$ (*i.e.,* multiplying by $\boldsymbol{P}_k$).

The sufficient statistic t represents the ratio of the signal affected by β_1 to the signal unaffected by β_1 such that the scaling imposed by σ_i is canceled between the numerator and denominator, as described in Chapter 4. This eliminates the effect of the noise parameter σ_i under each hypothesis.

A threshold test ϕ is then employed to decide between the hypotheses:

$$\phi(\boldsymbol{y}_k) = \begin{cases} \mathcal{H}_0 & \text{if } t(\boldsymbol{y}_k) \geq \eta \\ \mathcal{H}_1 & \text{else} \end{cases} \tag{6.4}$$

$\phi(\boldsymbol{y}_k)$ maintains a constant false positive rate as the distribution of the statistic t is invariant to the unknown parameters under the null hypothesis.

PPG Waveform Preprocessing

The PPG waveform is known to contain artifact associated with movement, spontaneous breathing, clipping, and missing data. The removal of PPG waveform artifact is an open area of research (Cannesson et al., 2008, 2010). Consistent with work by Convertino et al. (2011); McGrath et al. (2011); Stewart et al. (2014), we observe that without artifacts, the dominant non-DC frequencies of the PPG waveform correspond to the fundamental

frequency of the heart rate and its harmonic frequencies. As the test in Equation 6.3 only requires the sampled average PPG waveform, $\overline{PPG}$, we can employ this observation to generate the sampled average PPG waveform at each time step k, $\overline{PPG}(k)$, corresponding to a T second time window by dividing the T second window into J sub-windows of equal length. We then perform a spectral analysis via the Fourier transform of each sub-window. The sub-window's data is only included in the sampled average if the maximum non-DC frequency in that sub-window is likely to correspond to the heart rate. In the event that too many sub-windows do not meet the criteria, we treat the sampled average at that time as a missing measurement. Formally, this process is described in Algorithm 2 (assuming $\omega = \exp\{\frac{-2\pi i}{N}\}$, and N_0 and J_0 correspond to the minimum heart rate frequency and minimum number of sub-windows which must be included in the average, respectively):

Algorithm 2 PPG preprocessing algorithm

procedure PPG–PREPROCESSING
  $\overline{PPG}(k+1) = -1,\ S = 0,\ J = 0$
  for each sub window $j \in \{0, \ldots, J-1\}$ **do**
    for each frequency $l \in \{0, \ldots, N-1\}$ **do**
      $X_{l,j} = \sum_{n=0}^{N-1} PPG\left(\tau\left(kJN + jN + n\right)\right)\omega^{nl}$
    end for
    $\hat{l} = \arg\max_{l \in \{1,\ldots,N-1\}} X_{l,j}$
    if $\hat{l} > N_0$ **then** $S \mathrel{+}= \frac{1}{N} X_{w,0}$ and J ++
    end if
  end for
  if $J > J_0$ **then** $\overline{PPG}(k+1) = \frac{S}{J}$
  end if
end procedure

6.1.3 Results and Discussion

In this section we describe the retrospective data used to validate our PAIN detector, and provide a summary of results.

Evaluation Data Set

We evaluated the proposed technique over hypovolemic and non-hypovolemic patients drawn from the matched subset of the Physionet MIMIC II Waveform Database (Goldberger et al., 2000; Saeed et al., 2011). The matched subset allowed us to use nursing notes to find annotated times of suspected hypovolemia and subsequent fluid administration. To select non-hypovolemic patients, we chose patients from the database who had PPG waveforms, did not die, and had four or fewer ICD9 codes.

To select patients with a high likelihood of hypovolemia, we searched for patients with documented hypovolemia ICD9 codes on discharge with accompanying notes documenting approximate time of suspected hypovolemia. From these patients, we selected those who did not die and who had available PPG waveforms. Time of hypovolemia was annotated as the timestamp of the note describing suspected hypovolemia in the patient's file. We ran the algorithm only on data from the ICU stay with documented hypovolemia. Probability of false alarm was set to four false alarms per day, a level we deemed reasonable for a system that would continuously be in use.

Parameter Invariant Results

The results of running the proposed technique on each patient's PPG data can be found in Table 6.2. For all seven hypovolemic patients, our detector presented a higher-than-

Patient	LoS (hrs)	% data "good"	Hypovolemic	Alarms		
				Number	Mean Length (min)	Within 24 hrs of onset
03617	34.6	28.4	No	0	NaN	N/A
03640	22.0	6.4	No	0	NaN	N/A
05345	21.4	85.4	No	0	NaN	N/A
08949	34.5	90.3	No	5	17.8	N/A
11622	14.8	75.7	No	0	NaN	N/A
11727	24.3	38.4	No	0	NaN	N/A
14251	25.4	26.0	No	0	NaN	N/A
19309	6.3	67.2	No	0	NaN	N/A
19608	32.9	90.7	No	11	10.5	N/A
21986	18.0	74.1	No	0	NaN	N/A
27539	12.4	28.7	No	0	NaN	N/A
28706	40.8	41.2	No	1	6.0	N/A
29116	4.7	24.9	No	0	NaN	N/A
29126	4.7	76.2	No	0	NaN	N/A
30243	55.7	26.4	No	0	NaN	N/A
31015	5.2	25.7	No	0	NaN	N/A
31140	8.1	96.1	No	0	NaN	N/A
32249	22.3	11.1	No	0	NaN	N/A
00618	68.5	91.2	Yes	7	6.6	1
06085	37.6	53.9	Yes	1	1.0	1
07251	54.0	87.2	Yes	5	6.0	4
12351	160.8	62.9	Yes	7	6.4	4
16139	117.5	30.4	Yes	4	4.5	3
17582	11.6	90.3	Yes	1	8.0	1
22585	24.8	55.5	Yes	1	61.0	1

Table 6.2: Summary of patients selected from the Physionet MIMIC II database and results of the application of the PAIN-based hypovolemia detector on these patients. Patient IDs are from the matched subset. Length of stay (LoS) includes single relevant ICU stay.

average number of alarms within a 24-hour envelope of first documented hypovolemia.[21] One such patient's data, calculated statistic, and alarm values are plotted in Figure 6.1. A number of the hypovolemic patients considered had alarms prior to documented suspicion of hypovolemia. Several also had alarms long after. We consider these other alarms inconclusive, as patient notes provided no clear indication of when hypovolemia ended for these patients. Total alarm duration for these seven patients was 189 minutes (3.15 hours) out of a total time in the ICU of 474.81 hours (19.8 days).

The detector generated alarms for only three of the 18 non-hypovolemic patients. In total, these patients had a total 388.18 hours of ICU time, and experienced 18 alarms. These alarms had a duration of 196.8 minutes (3.28 hours). One of these patients (patient 19608) suffered respiratory issues which we believe may have triggered the false alarms. The data for one of the non-hypovolemic patients who did not trigger any alarms can be seen in Figure 6.2. Note that as the proportion of missing PPG data in the lower graph increases, the required threshold cutoff for alarming in the upper graph (the dotted line) also increases, preventing missing data from improperly triggering an alarm.

Overall, these results indicate that PAIN statistics and PPG waveforms can be used to create a hypovolemia detector robust to artifact and noise. The described preliminary tests on retrospective patient data suggest the proposed detector produces alarms near and before time of diagnosed hypovolemia while producing few false alarms in healthy patients, and seems to perform better than PVI, a state-of-the-art method, over these patients. The proposed approach seems to perform well over noisy data with artifact, making it particu-

[21] We felt a 24-hour envelope was reasonable, as in many patients gradual volume loss is not traditionally apparent for several hours or days (Shamir et al., 2011). No standard currently exists for the acceptable maximum amount of time to alert care teams to hypovolemia. For gradual volume loss, clinicians expect longer time to detection, as they currently depend on non-urgent detection methods such as postural hypotension or low urine output (Sladen, 2000).

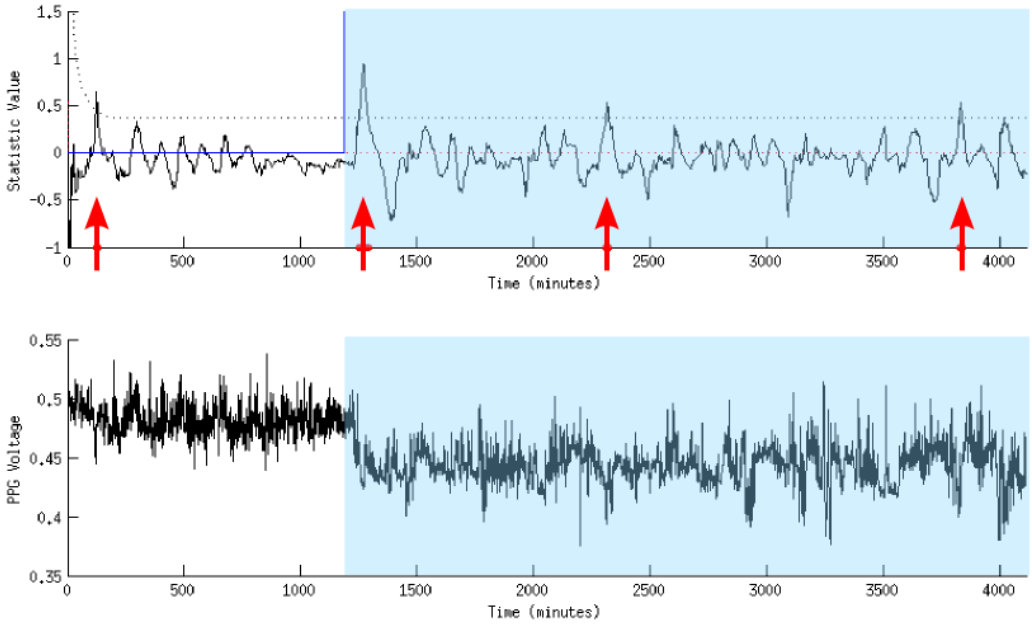


Figure 6.1: Section of PPG waveform (bottom) and accompanying statistic value (top) for patient 00618 in the MIMICII database. The blue highlighted section on the upper graph shows time post suspicion of hypovolemia. The upper graph shows the value of the parameter-invariant statistic, and red arrows indicate times of proposed hypovolemia alarms. The dotted line on the top graph indicates the threshold cutoff for the detector, which varies with amount of available data in the sliding window.

larly applicable to clinical intensive care settings. Though the size of the patient data set used was too small to assess definitive sensitivity or specificity, the results are promising. Because of the small size of the dataset, we did not apply the feature selection procedure outlined in Chapter 5, but hope to do so in future work.

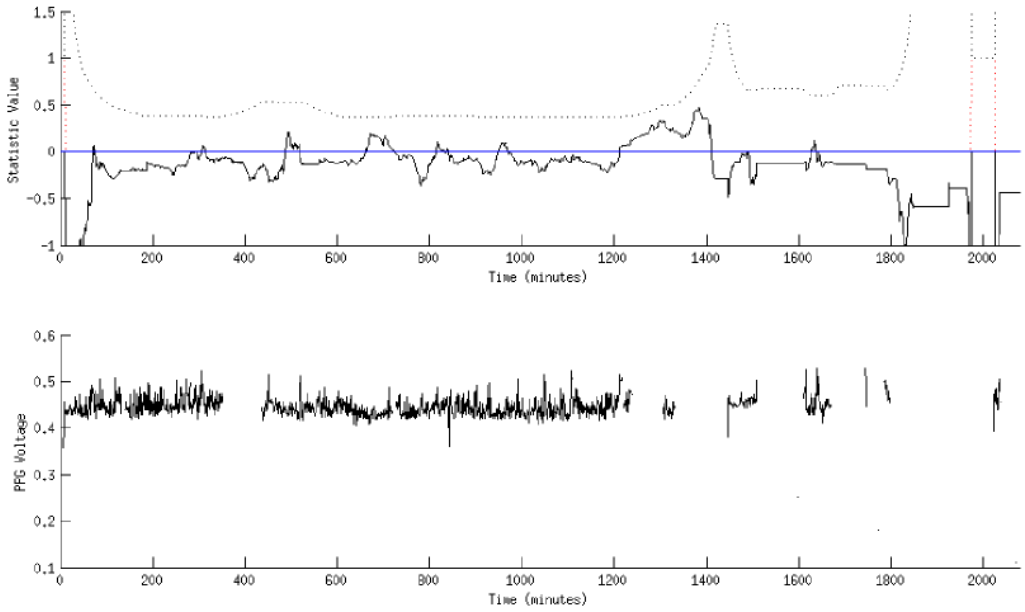


Figure 6.2: Section of PPG waveform for patient 03617 in the MIMICII database. This patient did not have hypovolemia. The upper graph shows the value of the parameter-invariant statistic. The lower graph shows the PPG waveform. No alarms would have been produced over this patient's ICU stay.

6.2 Meal Detection in Diabetic Patients

In this section, we describe the problem of meal detection for type I diabetics. Specifically, we discuss how a PAIN classifier can be built to analyze blood glucose levels and, at any point in time, classify the patient as having ingested a meal or not. The majority of the work in this section is not original, having been developed by Chen et al. (2015a). Our contribution is to extend that work using the PAIN feature set described in Chapter 5, and demonstrate an improvement in performance.

6.2.1 Medical Context

Type 1 diabetics depend on daily insulin infusion or injection to keep their glucose level within an acceptable range. Too much insulin can cause life-threatening hypoglycemia (extremely low glucose level) and too little insulin can cause nerve-damaging hyperglycemia (high glucose level) (Association, 2016). Ingested carbohydrates from meals cause a major disturbance to blood glucose levels, and therefore every type 1 diabetic faces a life-long control challenge: he or she has to carefully titrate insulin doses for every meal so that post-meal hyperglycemia is effectively controlled, while avoiding administering too much insulin and risking hypoglycemia. Currently, many type 1 diabetics use continuous glucose monitor sensors and wearable insulin pumps. These devices allow the user to manually input the time and estimated carbohydrate count of each meal into the device, which then calculates a suggested insulin dose. Unfortunately, self-reported meal information is known to be inherently unreliable (Dassau et al., 2008). More dependable meal detection methods are necessary to ensure patient safety. One possible solution is detecting meal events directly from physiologic data.

6.2.2 Methodology

In Section 4.2.1, we describe how to build PAIN classifiers on the occurrence (or non-occurrence) of an event, and how to build a two-sided PAIN classifier which tests for both occurrence and non-occurrence and checks to ensure that these tests agree. The benefit of such an approach is that good performance results when the hypotheses are accurate. When one hypothesis is inaccurate, however, or accurately annotated data is unavailable, these classifiers suffer significant performance loss.

As an alternative, Chen et al. (2015a) suggest monitoring for *sequentially differential* events. To do so, the classifier hypothesizes that an "event" (in this scenario, a meal) oc-

curred δ steps back from the current time (meals are treated as impulses in the detector model).[22] The null hypothesis $\mathcal{H}_0$ states that a meal indeed happened in a time window around the hypothesized meal time (the d_0 window in Figure 6.3). The event hypothesis $\mathcal{H}_1$ states that a meal actually happened in an even earlier time window (the d_1 window in Figure 6.3). Two separate statistics can then be calculated, as in the two-sided PAIN classifier (see Definition 4.2.1). The sequential differential monitor is illustrated in Figure 6.3. Here, we assume a window w, containing subwindows, d_0 and d_1, denoting the hypothesized time when each class $y \in 0, 1$ of event is assumed to occur. The detector works in a sliding window fashion; an accurately trained PAIN classifier will generate statistics similar to the bottom plot of Figure 6.3. In the figure, we observe a sequential rise and fall of the statistics, which provides a strong indication that an event has occurred. In practice, training sequential differential classifiers is difficult when the hypothesized windows are very close to one another, but as illustrated in Section 6.2.3, it can yield powerful results.

Simulation Details and Parameters

We maintain the parameter values chosen in Chen et al. (2015a): we used a sliding testing window w of 45 minutes. The subwindow size in which we test for the event's effect (illustrated in the figure by d_y) is set to be five minutes (five samples at one sample per minute), and these windows are considered five minutes before the end of the window (δ equal to five minutes).

The underlying system for the PAIN classifier is chosen to be a fifth-order LTI system, as consistent with the commonly accepted minimal glucose/insulin physiological model by Bergman et al. (1979). To generate data for evaluation, we use the FDA-accepted UVa/Padova Type 1 Diabetes Mellitus Metabolic Simulator (T1DMS) built upon a com-

[22] δ is a detector parameter that stays constant at run-time once it is chosen.

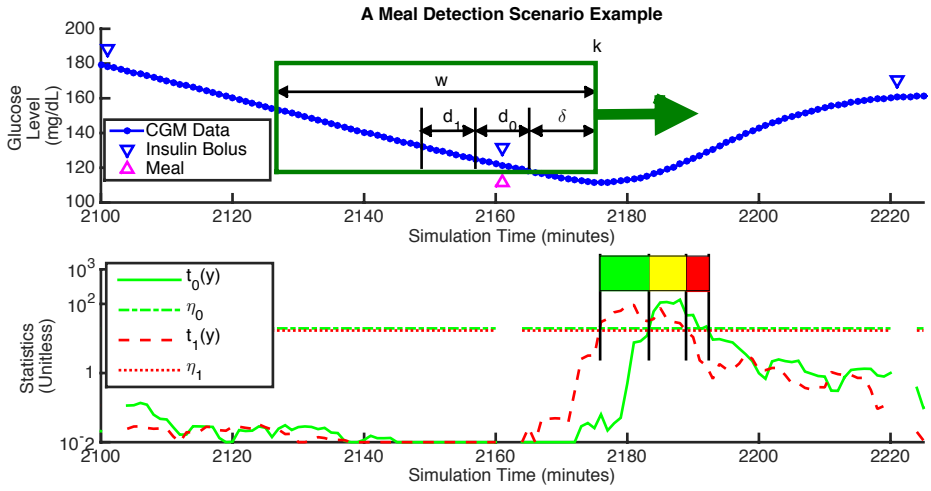


Figure 6.3: An illustration of the CFAR sliding window diabetic patient meal detector. The top graph shows movement of the sliding window over the patient's glucose level. The bottom graph showing the values of the statistics for each of the two-sided tests. Illustration taken from Chen et al. (2015a).

plex non-linear model that describes the glucose-insulin responses with meals as proposed by Magni et al. (2009) and Man et al. (2007). This model serves as an FDA-accepted substitute for animal testing in pre-clinical trials when evaluating glucose control algorithms. The simulator consists of continuous-time differential equations with thirteen state variables and 32 physiological parameters. The model includes three subsystems: the insulin subsystem, the meal glucose absorption, and the glucose kinetics. The academic version of the FDA-accepted type 1 Diabetic Simulator (Kovatchev et al., 2009) (T1DMS)

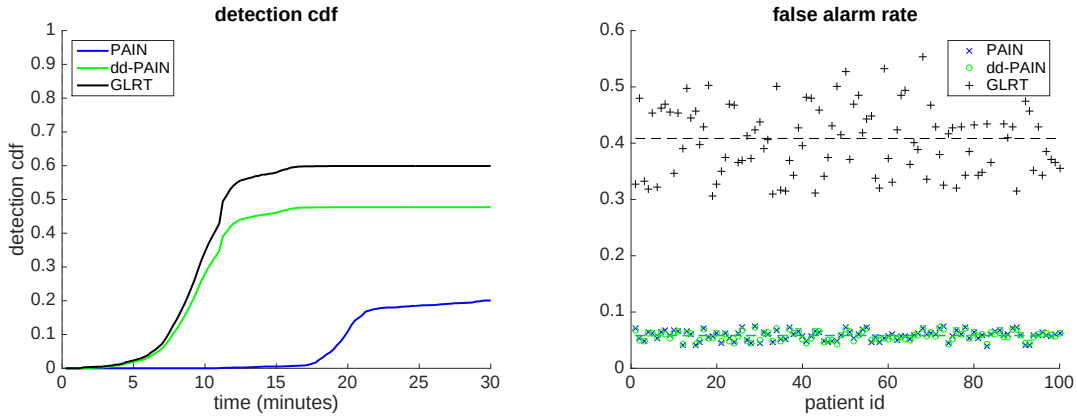
comes with ten virtual subjects that are drawn from the same parameter distribution with the FDA-accepted virtual subject population. Each virtual subject is a realization of the 32 parameters of the physiological model. To thoroughly test the performance of our detector across a wide range of possible patient parameters, we randomly sampled 200 new virtual subjects from the parameter space *spanned* by the ten provided T1DMS virtual subjects.

As in Chen et al. (2015a), we ran simulations of the FDA-accepted maximal model (with the virtual subjects' parameters) by "feeding" meals and insulin inputs that mimic the real-life scenario of a type 1 patient. Three meals with random amounts of carbohydrates are fed to a virtual subject every simulation day. Whenever a meal is given, a meal bolus based on a randomized meal insulin ratio is also fed into the simulator. In addition, we set a checkpoint once every hour except overnight (22:00–6:00) to mimic the scenario that a patient may check their own blood sugar and take correction insulin boluses when the glucose reading is too high. The dose of each correction bolus is calculated based on a randomly drawn insulin sensitivity value.

In addition to recreating the PAIN approach used by Chen et al. (2015a), we also tested the greedy PAIN feature set approach described in Chapter 5; we applied the GLRT as baseline. We omit the ARMAX classifier baseline due to the high-dimensionality of the problem.

6.2.3 Results and Discussion

Results for the experiment are shown in Figure 6.4. Because meal carbohydrates take time to have an effect on a patient's blood glucose, successful "detection" is defined with respect to time since the ingestion of the meal. In Figure 6.4a, we illustrate performance by plotting the classifier's cumulative detection rate over time. Note that all tested classifiers do not begin to detect meal events until approximately five minutes after they occur, suggesting



(a) Cumulative rate of detection over time.

(b) Individual false alarm rates for each simulated diabetic patient.

Figure 6.4: Graphs comparing the performance of the PAIN approach, a logistic regression classifier trained on PAIN features ("dd-PAIN"), and GLRT-based classifiers, all tested over simulated diabetic patients.

that patients' blood glucose only begins to be affected by metabolized carbohydrates five minutes after food is consumed. At this point, we see a sharp increase in the detection rates of both the GLRT and the "data-driven" PAIN approach (the PAIN feature set), whereas the PAIN approach alone does not begin to detect until fifteen minutes after the meal event and achieves a very low performance rate. These results show a significant improvement between PAIN and data-driven PAIN, likely due to the fact that the fifth-order system contained parameters that did not help to distinguish between patient groups. These parameters were removed through the machine learning process over the PAIN features, as intended. As the order of the model used increases, so too do the number of parameters in the model, and thus there is a greater likelihood for improvement between the PAIN and greedy PAIN feature set.

Overall, the GLRT achieves a higher detection rate than the data-driven PAIN approach.

Classifier	False Alarm Rate
GLRT	0.0439
PAIN	0.00090265
PAIN features+ML	$\mathbf{0.00086923}$

Table 6.3: The variance in individual false positive rate for three different classifiers over the diabetic patient data.

However, Figure 6.4b provides a more complete picture. Here, the false positive rate over each simulated patient is plotted. Under the GLRT-based classifier, patients experience an average false alarm rate of 0.4, with a high rate of variance (0.04) over each individual, as shown in Table 6.3. In contrast, both the PAIN and data-driven PAIN approaches achieve false alarm rates below 0.1 and, with a variance less than 0.001, maintain this rate over all patients in the population.

6.3 Pulmonary Shunt Detection in Infants

In this section, we consider the problem of monitoring the oxygen (O_2) content in a patient's blood during surgery, in particular to detect when an infant experiences a pulmonary shunt. The majority of the work in this section is not original, having been developed by Ivanov et al. (2015). Our contribution is to replicate that work, extend it using the PAIN feature set described in Chapter 5, and demonstrate an improvement in performance.

6.3.1 Medical Context

Blood O_2 content is perhaps the most closely monitored physiological variable, as values that are too low can lead to organ failure (*e.g.,* brain damage), and values that are too high can cause atelectasis (*i.e.,* collapse of the lungs). In this section, we focus on efforts to detect drops in blood O_2 content caused by pulmonary shunts in infants. A pulmonary shunt occurs when a patient is breathing with only one lung. Shunts can be caused by a physical disorder, such as pulmonary edema, or may occur accidentally in patients being mechanically ventilated, if ventilation tubes are improperly placed. Occasionally, one-lung ventilation is intentionally induced via a shunt at the request of a surgeon to keep a lung still for operation (Hammer, 2004). Infants are especially vulnerable to accidental shunts because they have very short circulation times and underdeveloped lungs (Oberst and La Roche, 1954). In these patients, breathing with one lung may not supply enough O_2 to the body, and the O_2 content may quickly drop to dangerously low levels.

Though it is one of the most closely monitored physiological variables, blood O_2 content is also one of the most challenging to monitor, as it cannot currently be measured non-invasively or in real time. Instead, clinicians must monitor proxy variables. One popular proxy is the hemoglobin oxygen saturation in the peripheral capillaries, denoted by S_pO_2. While it is a good noninvasive measure of the O_2 content in the location at which it is measured (usually a fingertip, or the foot in small infants), S_pO_2 is a delayed measure of the O_2 content in other parts of the body (*e.g.,* the arteries), as blood takes time to circulate (Shafer et al., 2014). For this application, we aim to predict drops in a patient's O_2 content before these drops are observed through the current low S_pO_2 approach by using other available physiological measurements. In particular, we consider multiple pulmonary measurements that are available through an anesthesia machine, namely the partial pressures of O_2 and carbon dioxide (CO_2) as well as the tidal volume and respiratory rate.

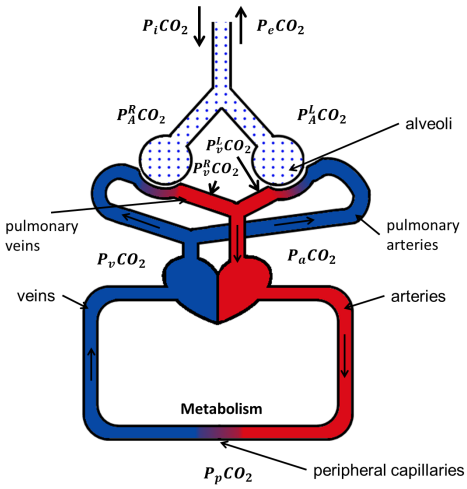
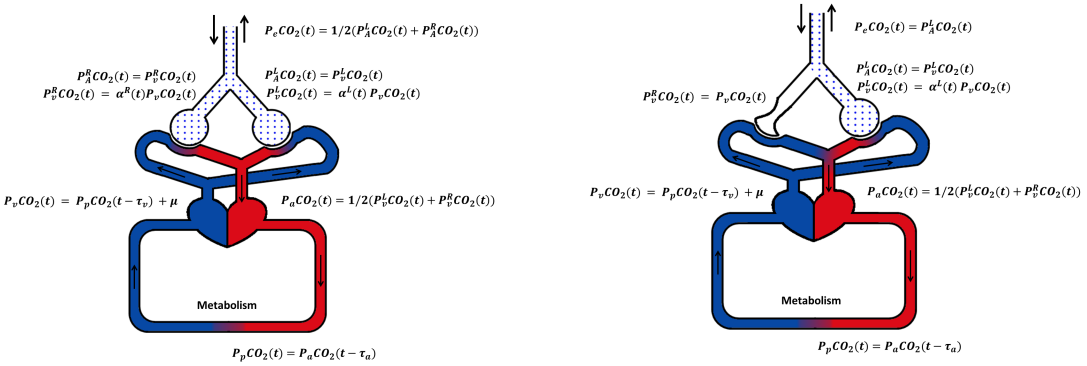


Figure 6.5: A simplified schematic model of CO_2 partial pressures in the respiratory and cardiovascular systems. Taken from Ivanov et al. (2015).

In some medical scenarios, linear physiological input-output models are known and readily available. Ivanov et al. (2015) notes that modeling the entire circulation of O_2 is not possible since there is no known closed-form expression for its diffusion (*i.e.,* the movement of gases from lungs to the air and vice versa). CO_2, on the other hand, has diffusive properties that are better suited for the development of such a model. In addition, there is a correlation between the partial pressure of CO_2 and the O_2 content: sharp increases in the CO_2 partial pressure are correlated with decreases in blood O_2 content. With this information, Ivanov et al. (2015) developed a model of the circulation of the partial pressures of O_2 and CO_2 around the cardiovascular and pulmonary systems for the purposes of critical pulmonary shunt detection in infants, as illustrated in Figure 6.5 and Figure 6.6.

In scenarios where physiological models such as these exist, they can be used instead of the LTI formulation previously described. The ultimate trade-off for classification based on physiological models is the cost associated with developing the model (*e.g.,* time, expertise,



(a) System with two functioning lungs.

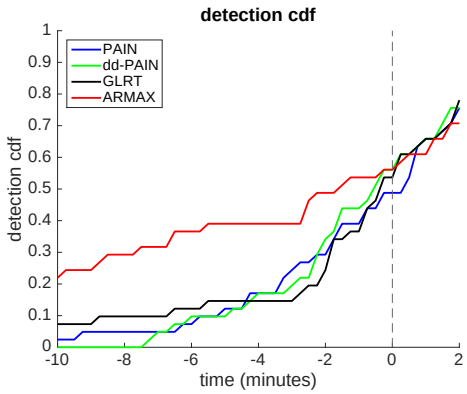
(b) System with a shunt.

Figure 6.6: An illustration of the response of the respiratory and cardiovascular partial pressures to a shunt. Taken from Ivanov et al. (2015).

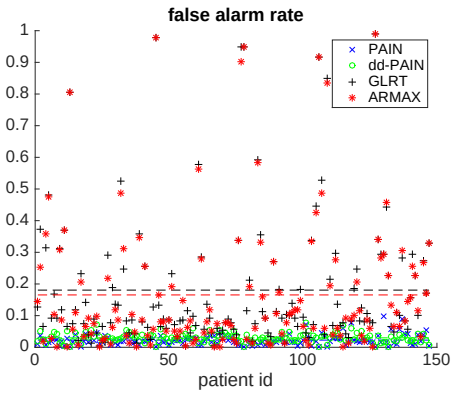
etc.) and a risk of overfitting versus the potential for performance improvement. This trade-off is investigated in the context of the experimental results for the critical pulmonary shunt monitor in Section 6.3.3.

6.3.2 Methodology

We maintain the parameter values chosen in Ivanov et al. (2015), generating a time-series based PAIN classifier using a second-order LTI model, and one based on their derived physiologic model on the lobectomy dataset they use. In addition, for both models, we applied our proposed greedy PAIN feature set as described in Chapter 5, learning a logistic regression classifier over the features. As in the previous work, because the number of patients in the dataset with confirmed shunts was relatively small, we use leave-one-out cross validation to train the logistic regression classifier over the greedy PAIN feature set, as well as the other maximum likelihood–based approaches.



(a) The cumulative detection rate for each of the classifiers in the time leading up to an annotated drop in patient SpO_2.



(b) False alarm rate for each patient in the test set for each of the tested classifiers.

Figure 6.7: Graphs comparing the performance of the candidate PAIN classifiers and the GLRT and ARMAX classifiers over the lobectomy/shunt data set.

6.3.3 Results and Discussion

Results are presented in Figure 6.7. In Figure 6.7a, the cumulative rate of detection for each of the classifiers is plotted over time. The x-axis shows time, in minutes, leading up to the annotated drop in patient S_pO_2 (which acts as time 0). We use this drop in S_pO_2 as the gold-standard marker of a shunt having occurred; however, we know that this is a late-stage marker, which happens a number of minutes after the actual shunt. While the precise time of shunt is uncertain in the data, we know (based on the time it takes for blood to circulate the body) that the true event time must occur in the minutes leading up to the annotated severe drop in S_pO_2.

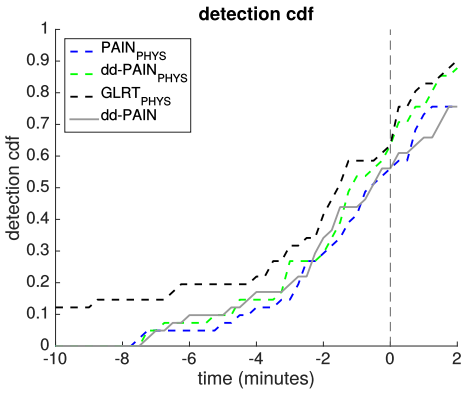
In the figure, it can be seen that all of the classifiers see a dramatic increase in detection between four and one minutes before the drop in S_pO_2, with the data-driven PAIN approach and the ARMAX model achieving the highest overall accuracy by the time of the S_pO_2

drop. Overall, the proposed data-driven PAIN approach outperforms the original PAIN classifier in detection rate. When we consider false alarm rate over each individual in the population (Figure 6.7b), we see that the parameter invariant approach achieves a very low rate of false alarms over all individuals in the population, and the false alarm rates of the ARMAX and GLRT classifiers, while low overall, have very high variabilities, with some patients achieving a false alarm rate near 1.

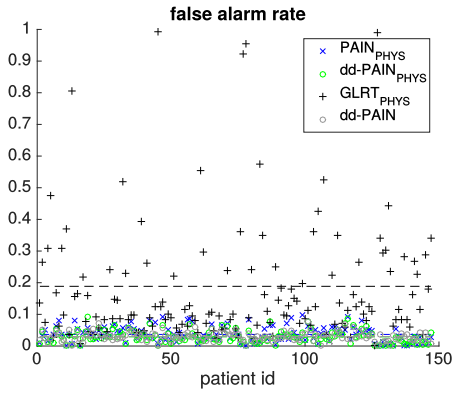
Next, we considered the classifiers developed using the physiologic model. We ran the original parameter invariant classifier described in that previous work, then extended it to include the proposed greedy PAIN feature set. We also tested a GLRT classifier based on the physiologic model. As the model was already defined, an ARMAX/MLE approach was not appropriate.

Results are shown in Figure 6.8, with the classifier trained on the non-physiologic greedy PAIN feature set technique (dd-PAIN) included for reference (in solid gray line). Note that the classifier trained over greedy PAIN features and the GLRT approach both achieve a small boost in detection performance with the use of a tailored physiologic model. Though the GLRT continues to achieve higher detection cdf, the PAIN classifiers both continue to maintain their relatively constant, low rates of false alarm.

All of the experimental results we have presented have demonstrated the advantages of the proposed PAIN approach, and the improvement in detection offered by the use of PAIN features.



(a) Cumulative detection rate for classifiers built upon circulatory physiologic model in the time leading up to the annotated drop in patient SpO_2.



(b) False alarm rate for each patient in the test set for each of the tested classifiers built upon circulatory physiologic model.

Figure 6.8: Graphs comparing the performance of the candidate PAIN and data-driven PAIN classifiers, applied to an LTI model based in physiology, with the performance of a GLRT classifier. All classifiers were tested over the lobectomy/shunt data set. dd-PAIN performance over standard non-physiologic learned LTI model from Figure 6.7 included for reference.

Chapter 7

Conclusion

In this work, we have described parameter invariant statistics, and demonstrated how to use them as part of several different methods of classification. We have shown that parameter invariant classifiers, if used in this way, achieve robust classification when a population is made up of individuals whose parameters vary, as parameter invariance allows us to achieve good classification accuracy while maintaining a bound on false positive rate over every individual in a population. While the parameter invariant approach does not achieve population-level detection rates as high as those of the GLRT or an MLE/ARMAX approach, a bounded rate of false positive is a useful result when a high number of false positives could cause a patient to receive substandard care. We showed how parameter invariant statistics can be calculated for any linear time-invariant system over groups of nuisance transformations, and calculated the statistic for cases in which the transformation group includes translation, rotation, and scaling. We described how to perform classification using a parameter invariant statistic, and how a "dual" classifier extension allows for lower bounds on detection performance. We showed that generating a number of parameter invariant statistics over different sets of parameters and then applying feature selection

maintains a constant false positive rate while boosting the true positive rate, as it allows for the implicit selection of the best subspace to eliminate. Finally, we applied parameter invariant classification to a number of different medical scenarios: detection of hypovolemia, meal detection in diabetic patients, and detection of pulmonary shunts in infants.

Parameter invariant statistics can be used to classify time series data generated by LTI systems with a guaranteed minimum level of performance without tuning over individual patients, even when the parameters of the individuals in the population have high variance. We have also shown that generating a set of parameter invariant statistics over different subspaces and using them as a feature set for a machine learning classifier results in a higher detection rate than using a single statistic alone. We have begun to successfully apply the proposed approach to real-world medical applications. We hope that this technique can be refined further and eventually allow for the implementation of CDS systems with bounded rates of false alarms. Such systems could be deployed even in scenarios where the amount of patient data available to learn from was small, and would be more robust than traditional systems.

Note that while this work focuses on the ICU, as medical devices become more sophisticated and less expensive, continuous monitoring has begun to spread outward to the general floor of the hospital and to home care scenarios. This work thus has applicability anywhere clinical monitoring is employed. More broadly, while this thesis focuses on applying PAIN statistics to enable clinical decision support in the medical domain, their application is not limited to this domain. Indeed, the developed methodology should have applicability in numerous other spaces. Some of the characteristics that make physiologic signals challenging for clinicians to use—real-time, high-frequency data; variations in data's provenance; intersource variation—are common characteristics of much of the data produced by cyber-physical, financial, and social systems today. We hope that our work can achieve broad applicability to building decision support systems in scenarios where having a certain

bound on false alarm rates is desirable.

7.1 Future Work

There are a number of possible directions for future work. A major area involves exploring solutions to the other two information-loss problems mentioned in Section 5.1. Regaining information lost in the null space projection process would boost the performance of the PAIN statistic.

The technique described in this work uses regression to build the final classifier over the parameter invariant statistics. Another area of future work is incorporating parameter invariance into other more robust machine learning techniques. Ideally, the concept of parameter invariance could be encoded into a kernel function, which would allow it to be used in a wide variety of classifiers.

Though the PAIN statistics in this work are calculated using a sequence of measurements, this work does not consider how learning could occur over sequences of PAIN statistics themselves. The behavior of the PAIN statistic as it changes over time has the potential to provide interesting information. There are a number of possible ways of modifying the proposed feature set (*e.g.,* using sliding windows) to incorporate this information.

In this work, we only calculate PAIN statistics over individuals who are not missing data. Missing data, however, is a common occurrence in the clinical care environment, as sensors fall off or are moved on patients frequently. The PAIN statistic approach could be extended to cope with periods of missing data, ideally to allow the statistic to still be calculated if the gap in data is small enough. Additionally, knowledge that data is missing often itself indicates something about the behavior of the system. Future work might involve incorporating this information into the calculation of the statistic itself.

This work describes the convergence of the PAIN statistic to a constant false alarm

rate under certain conditions. Similarly, we have shown that the greedy PAIN features we introduced converge to the more traditional PAIN statistic built over an optimal dimensional subset if incorporated into an appropriate classifier. In both cases, we do not establish convergence rate. We hope to do so in future work.

In Section 4.2, we describe a generalized form of a parameter invariant classifier to approximate a maximally invariant statistic when one does not exist. However, there are numerous other possible parameter invariant classifiers that can be shown to be maximally invariant. Future work may involve developing these alternate classifiers and identifying their advantages and disadvantages.

PAIN statistics could be applied to a number of other medical problems where inter-patient variability makes detection challenging. Areas of interest include acute respiratory distress syndrome and sepsis. Ideally, the PAIN classifiers we have described would be incorporated into full clinical decision support systems for most clinical ailments, with underlying infrastructure supporting easy clinical use. Building such systems would require overcoming numerous other technological and legal challenges, including improving hospital data acquisition infrastructure and developing more advanced data visualization mechanisms, and so we leave such implementation matters to future work. We hope, however, that the work of this thesis will provide one more small step toward the proliferation of scalable, reliable CDS systems that can improve patient health and save lives in hospitals worldwide.

Bibliography

A Ml Albisser, BS Leibel, TG Ewart, Z Davidovac, CK Botz, W Zingg, H Schipper, and R Gander. Clinical control of diabetes by the artificial pancreas. *Diabetes*, 23(5):397–404, 1974. 18

Gregory L. Alexander. Issues of Trust and Ethics in Computerized Clinical Decision Support Systems. *Nursing Administration Quarterly*, 30(1):21–29, January 2006. 5

Aymen A Alian, Nicholas J Galante, Nina S Stachenfeld, David G Silverman, and Kirk H Shelley. Impact of central hypovolemia on photoplethysmographic waveform parameters in healthy volunteers part 2: frequency domain analysis. *Journal of clinical monitoring and computing*, 25(6):387–96, December 2011. 94

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1–39, March 2007. 93

Lars Prag Antonsen and Knut Arvid Kirkebøen. Evaluation of fluid responsiveness: is photoplethysmography a noninvasive alternative? *Anesthesiology research and practice*, January 2012. 94

American Diabetes Association. Hypoglycemia (low blood glucose), 2016. URL `http://www.diabetes.org/living-with-diabetes/`

treatment-and-care/blood-glucose-control/hypoglycemia-low-blood.html. [Online; accessed 8-July-2016]. 104

Karl Johan Aström and Peter Eykhoff. System identification survey. *Automatica*, 7(2): 123–162, 1971. 17

Karsten Bartels, Robert H Thiele, and Tong J Gan. Rational fluid management in today's ICU practice. *Critical Care*, 17(Suppl 1):2–7, 2013. 94

Casey Bennett and Thomas W Doub. Data mining and electronic health records: Selecting optimal clinical treatments in practice. In *Proceedings of the 6th International Conference on Data Mining*, pages 313–318, 2011. 3

Peter M Bentler and Theo Dijkstra. Efficient estimation via linearization in structural models. *Multivariate analysis VI*, pages 9–42, 1985. 55

B Wayne Bequette. A critical assessment of algorithms and challenges in the development of a closed-loop artificial pancreas. *Diabetes technology & therapeutics*, 7(1):28–47, 2005. 18

Richard N Bergman, Y Ziya Ider, Charles R Bowden, and Claudio Cobelli. Quantitative estimation of insulin sensitivity. *American Journal of Physiology-Endocrinology And Metabolism*, 236(6):E667, 1979. 18, 105

Eta S Berner. Clinical Decision Support Systems : State of the Art. Technical Report 09, Agency for Healthcare Research and Quality, 2009. 2

P. Bogdan, S. Jain, K. Goyal, and R. Marculescu. Implantable pacemakers control and optimization via fractional calculus approaches: A cyber-physical systems perspective. In *Proceedings of the Third International Conference on Cyber-Physical Systems*, pages 23–32, 2012. 18

Torsten Bohlin and Stefan F Graebe. Issues in nonlinear stochastic grey box identification. *International Journal of Adaptive Control and Signal Processing*, 9(6):465–490, 1995. 17

Nicolas Bon, Ali Khenchaf, and Rene Garello. GLRT subspace detection for range and Doppler distributed targets. *IEEE Transactions on Aerospace and Electronic Systems*, 44(2):678–696, April 2008. 9

George E P Box and Gwilym M Jenkins. *Time Series Analysis: Forecasting and Control*, volume Third. 1994. ISBN 0130607746. 20

Marc Breton, Anne Farret, Daniela Bruttomesso, Stacey Anderson, Lalo Magni, Stephen Patek, Chiara Dalla Man, Jerome Place, Susan Demartini, Simone Del Favero, et al. Fully integrated artificial pancreas in type 1 diabetes modular closed-loop glucose control maintains near normoglycemia. *Diabetes*, 61(9):2230–2237, 2012. 18

M Cannesson, O Desebbe, P Rosamel, B Delannoy, J Robin, O Bastien, and J-J Lehot. Pleth variability index to monitor the respiratory variations in the pulse oximeter plethysmographic waveform amplitude and predict fluid responsiveness in the operating theatre. *British journal of anaesthesia*, 101(2):200–6, August 2008. 94, 97

Maxime Cannesson, Mateo Aboy, Christoph K Hofer, and Mohamed Rehman. Pulse pressure variation: where are we today? *Journal of Clinical Monitoring and Computing*, 25 (1):45–56, April 2010. 94, 97

Christopher K Carter and Robert Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994. 21

Leo Anthony Celi, Roger G Mark, David J Stone, and Robert A Montgomery. “Big Data”

in the Intensive Care Unit. *American journal of respiratory and critical care medicine*, 187(11):1157–1166, 2010. 2

B Chaudhry, J Wang, and S Wu. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*, 144: 742–752, 2006. 2, 4

Sanjian Chen, James Weimer, Michael Rickels, Amy Peleckis, and Insup Lee. Towards a model-based meal detector for type I diabetics. In *Medical Cyber-Physical Systems Workshop 2015*, 2015a. 103, 104, 105, 106, 107

Sanjian Chen, James Weimer, Michael R Rickels, Amy Peleckis, and Insup Lee. Towards a model-based meal detector for type i diabetics. In *Medical Cyber-Physical Systems Workshop 2015*, 2015b. 9

T Clark, Y David, M Baretich, and T Bauld. Impact of clinical alarms on patient safety. Technical report, ACCE Healthcare Technology Foundation, 2006. 6

Lei Clifton, David A Clifton, Peter J Watkinson, and Lionel Tarassenko. Identification of Patient Deterioration in Vital-Sign Data using One-Class Support Vector Machines. In *Proceedings of the Federated Conference on Computer Science and InformationSystems*, number ii, pages 125–131, 2011. 22

Claudio Cobelli and Ewart Carson. *Introduction to modeling in physiology and medicine*. Academic Press, 2008. 17

Claudio Cobelli, Eric Renard, and Boris Kovatchev. Artificial pancreas: past, present, future. *Diabetes*, 60(11):2672–2682, 2011. 18

Committee On Quality Of Healthcare In America. Crossing the Quality Chasm: A New Health System for the 21st Century. Summary. *Health San Francisco*, pages 1–358, 2001. 3

Victor a. Convertino, Caroline a. Rickards, Keith G. Lurie, and Kathy L. Ryan. Hyperventilation in Response to Progressive Reduction in Central Blood Volume to Near Syncope. *Aviation, Space, and Environmental Medicine*, 80(12):1012–1017, December 2009. 93

Victor a Convertino, Steven L Moulton, Gregory Z Grudic, Caroline a Rickards, Carmen Hinojosa-Laborde, Robert T Gerhardt, Lorne H Blackbourne, and Kathy L Ryan. Use of advanced machine-learning techniques for noninvasive monitoring of hemorrhage. *The Journal of trauma*, 71(1 Suppl):S25–32, July 2011. 93, 97

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA Type 1 Diabetes Simulator new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014. 18

Eyal Dassau, B Wayne Bequette, Bruce A Buckingham, and Francis J Doyle. Detection of a meal using continuous glucose monitoring implications for an artificial β-cell. *Diabetes care*, 31(2):295–300, 2008. 104

TG Dietterich. Machine Learning for Real-Time Decision Making. Technical report, 2001. 23

Patrice Forget, Fernande Lois, and Marc de Kock. Goal-directed fluid management based on the pulse oximeter-derived pleth variability index reduces lactate levels and improves fluid management. *Anesthesia and analgesia*, 111(4):910–4, October 2010. 94

Emily Fox, Erik B Sudderth, Michael I Jordan, and S Alan. Bayesian Nonparametric Infer-

ence of Switching Dynamic Linear Models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, 2011. 21

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Nonparametric Bayesian Learning of Switching Linear Dynamical Systems. *Proceedings of Neural Information Processing Systems*, 21, 2008. 21

Frost & Sullivan. Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations. Technical report, 2012. 2

Amit X Garg, Neill K J Adhikari, Heather McDonald, M Patricia Rosas-Arellano, P J Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, 293(10):1223–38, March 2005. 3

Andrew Gelman. *Bayesian data analysis*. 1995. 21

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. 91, 92, 99

Jen J Gong, Thoralf M Sundt, James D Rawn, and John V Guttag. Instance weighting for patient-specific risk stratification models. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 369–378. ACM, 2015. 25

Robert A. Greenes. *Clinical Decision Support: The Road to Broad Adoption: Second Edition*. 2014. ISBN 9780123984760. 3

F.S. Grodins. *Control theory and biological systems*. Columbia University Press, 1963. 16

Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. The 'big data' revolution in healthcare. *McKinsey Quarterly*, (January), 2013. 2

Guillermo Gutierrez, H David Reines, and Marian E Wulf-Gutierrez. Clinical review: Hemorrhagic Shock. *Critical care (London, England)*, 8(5):373–81, October 2004. 93

James D. Hamilton. *Time Series Analysis.*, volume 24. March 1995. 19

G. B. Hammer. Single-lung ventilation in infants and children. *Pediatric Anesthesia*, 14: 98–102, 2004. 110

Dereck L. Hunt, R. Brian Haynes, Steven E. Hanna, and Kristina Smith. Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes. *Jama*, 280(15):1339, October 1998. 3

R. Ivanov, J. Weimer, A. Simpao, M. Rehman, and I. Lee. Early detection of critical pulmonary shunts in infants. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems (ICCPS)*, pages 110–119, 2015. 9, 109, 111, 112

Tommi Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999. 24

Tony Jebara and Andrew Howard. Probability Product Kernels. *The Journal of Machine Learning Research*, 5:819–844, 2004. 24

Z. Jiang, M. Pajic, and R. Mangharam. Cyber-physical modeling of implantable cardiac medical devices. In *Proceedings of the IEEE*, pages 122–137, 2012. 18

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999. 21

Anatoli Juditsky, Hakan Hjalmarsson, Albert Benveniste, Bernard Delyon, Lennart Ljung, Jonas Sjöberg, and Qinghua Zhang. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31(12):1725–1750, 1995. 16

R E Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering*, 82(Series D):35–45, 1960. 20

Robert E. Kass and Duane Steffey. Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *Journal of the American Statistical Association*, 84(407):717–726, 1989. 21

Aaron S Kesselheim, Kathrin Cresswell, Shobha Phansalkar, David W Bates, and Aziz Sheikh. Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. *Health affairs (Project Hope)*, 30(12):2310–7, December 2011. 5

Ellen Kilsdonk, Linda W Peute, Rinke J Riezebos, Leontien C Kremer, and Monique W M Jaspers. From an expert-driven paper guideline to a user-centred decision support system: a usability comparison study. *Artificial intelligence in medicine*, 59(1): 5–13, September 2013. 4

Minyoung Kim and Vladimir Pavlovic. Discriminative Learning of Dynamical Systems for Motion Tracking. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. 21

LT Kohn, JM Corrigan, and MS Donaldson. *To Err Is Human: Building a Safer Health System*. 2000. ISBN 0309068371. 3

Daphne Koller, Nir Friedman, Lise Getoor, and Ben Taskar. Graphical Models in a Nutshell. In *Statistical Relational Learning*, pages 13–55. MIT press, 2007. 21

Ross Koppel, Abigail Cohen, Brian Abaluck, A Russell Localio, Stephen E Kimmel, and Brian L Strom. Role of computerized physician order entry systems in facilitating medication errors. *Journal of the American Medical Association*, 293(10):1197–1203, 2005. 6

Boris P. Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *Diabetes Sci Technol*, 3(1):44–55, 2009. 106

S. Kraut and L.L. Scharf. The CFAR adaptive subspace detector is a scale-invariant GLRT. *IEEE Transactions on Signal Processing*, 47(9):2538–2541, 1999. 9, 39

Niels Rode Kristensen, Henrik Madsen, and Sten Bay Jørgensen. Parameter estimation in stochastic grey-box models. *Automatica*, 40(2):225–237, 2004. 17

I. Lee, O. Sokolsky, S. Chen, et al. Challenges and research directions in medical cyber–physical systems. *Proceedings of the IEEE*, 100(1):75–90, 2012. 26

Insup Lee and Oleg Sokolsky. Medical cyber physical systems. In *Proceedings of the 47th Design Automation Conference*, pages 743–748. ACM, 2010. 26

T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11): 1857–1874, 2005. 23

Lennart Ljung and Svante Gunnarsson. Adaptation and tracking in system identificationa survey. *Automatica*, 26(1):7–21, 1990. 17

Thibault Loupec, Hodanou Nanadoumgar, Denis Frasca, Franck Petitpas, Leila Laksiri, Didier Baudouin, Bertrand Debaene, Claire Dahyot-Fizelier, and Olivier Mimoz. Pleth variability index predicts fluid responsiveness in critically ill patients. *Critical Care Medicine*, 39(2):294–299, February 2011. 94

L. Magni, D.M. Raimondo, C. Dalla Man, G. De Nicolao, B. Kovatchev, and C. Cobelli. Model predictive control of glucose concentration in type i diabetic patients: An in silico trial. *Biomedical Signal Processing and Control*, 4(4):338 – 346, 2009. 106

Lalo Magni, Davide M Raimondo, Luca Bossi, Chiara Dalla Man, Giuseppe De Nicolao, Boris Kovatchev, and Claudio Cobelli. Model predictive control of type 1 diabetes: an in silico trial. *Journal of diabetes science and technology*, 1(6):804–812, 2007. 18

Chiara Dalla Man, Robert A. Rizza, and Claudio Cobelli. Meal simulation model of the glucose-insulin system. *Biomedical Engineering and IEEE Transactions on*, 54(10): 1740–1749, oct. 2007. 106

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbes, Charles Roxburgh, and Angela Hung Byers. Big data : The next frontier for innovation , competition , and productivity. Technical Report June, 2011. 2

Paul E Marik, Rodrigo Cavallazzi, Tajender Vasu, and Amyn Hirani. Dynamic changes in arterial waveform derived variables and fluid responsiveness in mechanically ventilated patients: a systematic review of the literature. *Critical care medicine*, 37(9):2642–7, September 2009. 93

Steven McGee, William B. III Abernethy, and David L. Simel. Is This Patient Hypovolemic? *Journal of the American Medical Association*, 281(11):1022, March 1999. 93

Susan P McGrath, Kathy L Ryan, Suzanne M Wendelken, Caroline a Rickards, and Victor a Convertino. Pulse oximeter plethysmographic waveform changes in awake, spontaneously breathing, hypovolemic volunteers. *Anesthesia and analgesia*, 112(2):368–74, February 2011. 95, 97

Howard T Milhorn. Application of control theory to physiological systems. 1966. 16

J.H. Milsum. *Biological control systems analysis*. McGraw-Hill electronic science series. McGraw-Hill, 1966. 16

Harry G Mond and Alessandro Proclemer. The 11th world survey of cardiac pacing and implantable cardioverter-defibrillators: Calendar year 2009–a world society of arrhythmia's project. *Pacing and Clinical Electrophysiology*, 34(8):1013–1027, 2011. 18

Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. *Advances in Neural Information Processing Systems*, 2003. 24

M Moscucci, K.A.A. Fox, Christopher P. Cannon, W. Klein, Jose Lopez-Sendon, G Montalescot, K. White, and R.J. Goldberg. Predictors of major bleeding in acute coronary syndromes: the Global Registry of Acute Coronary Events (GRACE). *European Heart Journal*, 24(20):1815–1823, October 2003. 93

Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*, 2015:132, 2015. 25

DENIS Noble. The surprising heart: a review of recent progress in cardiac electrophysiology. *The Journal of Physiology*, 353(1):1–50, 1984. 18

Denis Noble. Modeling the heart–from genes to cells to the whole organ. *Science*, 295 (5560):1678–1682, 2002. 18

B. B. Oberst and F. La Roche. Circulation time in the newborn infant, using the fluorescein dye method. *Journal of Pediatrics*, 45(5):580–582, 1954. 110

Jerome A Osheroff, Eric A Pifer, Jonathan M Teich, Dean F Sittig, and Robert A Jenders. Improving outcomes with clinical decision support: an implementer's guide. HIMSS, 2005. 3

Johnny T Ottesen, Mette S Olufsen, and Jesper K Larsen. *Applied mathematical models in human physiology*. Siam, 2004. 16

M Peleg and S Tu. Decision Support , Knowledge Representation and Management in Medicine. *IMIA Yearbook of Medical Informatics*, 45:72–80, 2006. 3

Reuven Pizov, Arieh Eden, Dmitri Bystritski, Elena Kalina, Ada Tamir, and Simon Gelman. Arterial and Plethysmographic Waveform Analysis in Anesthetized Patients with Hypovolemia. *Anesthesiology*, 113(1):83–91, 2010. 94

John A Quinn, Christopher K.I. Williams, and Neil McIntosh. Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009. 20

Lawrence R Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 20

Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):3, 2014. 2

Ruty Rinott, Boaz Carmeli, Carmel Kent, Yonatan Maman, Yoav Rubin, and Noam Slonim. Utilizing assigned treatments as labels for supervised machine learning in clinical decision support. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*, page 493, 2012. 3

Alexander Roederer, James Weimer, Joseph DiMartino, Jacob Gutsche, and Insup Lee. Robust monitoring of hypovolemia in intensive care patients using photoplethysmogram signals. In *Proceedings of the 37th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. EMBC, 2015. 9

Lorenzo Rosasco, E Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004. 87

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011. 91, 92, 99

Rakesh Sahni. Noninvasive monitoring by photoplethysmography. *Clinics in perinatology*, 39(3):573–83, September 2012. 94

Suchi Saria, Daphne Koller, and Anna Penn. Learning individual and population level traits from clinical temporal data. In *Proceedings of Neural Information Processing Systems*, pages 1–9, 2010. 21

Suchi Saria, Andrew Duchi, and Daphne Koller. Discovering deformable motifs in con-

tinuous time series data. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 22(1):1465, 2011. 23

L. L. Scharf and C. Demeure. *Statistical Signal Processing*. Addison-Wesley Publishing Company, 1991. 29, 33, 34, 39, 48

L.L. Scharf and B. Friedlander. Matched subspace detectors. *IEEE Transactions on Signal Processing*, 42(8):2146–2157, 1994. 9, 47

F Sha and LK Saul. Large margin hidden Markov models for automatic speech recognition. *Advances in neural information processing systems*, pages 1249—-1256, 2006. 21

S. Shafer, J. P. Rathmell, and R. Stoelting. *Stoelting's Pharmacology & Physiology*. Wolters Kluwer, 2014. 110

Micha Y Shamir, Leonid Kaplan, Rachel S Marans, Dafna Willner, and Yoram Klein. Urine flow is a novel hemodynamic monitoring tool for the detection of hypovolemia. *Anesthesia & Analgesia*, 112(3):593–596, 2011. 101

Kirk H Shelley. Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate. *Anesthesia and analgesia*, 105(6 Suppl):S31–6, tables of contents, December 2007. 93

Ali Shoeb. *Patient-specific seizure onset detection*. PhD thesis, 2003. 23

Ali Shoeb, Herman Edwards, Jack Connolly, Blaise Bourgeois, S Ted Treves, and John Guttag. Patient-specific seizure onset detection. *Epilepsy & behavior : E&B*, 5(4): 483–98, August 2004. 25

Jonas Sjöberg, Qinghua Zhang, Lennart Ljung, Albert Benveniste, Bernard Delyon, Pierre-Yves Glorennec, Hakan Hjalmarsson, and Anatoli Juditsky. Nonlinear black-box mod-

eling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995. 16

Robert N Sladen. Oliguria in the icu: systematic approach to diagnosis and treatment. *Anesthesiology Clinics of North America*, 18(4):739–752, 2000. 101

AFM Smith and M West. Monitoring renal transplants: An application of the multiprocess kalman filter. *Biometrics*, pages 867–878, 1983. 21

Oleg Sokolsky, Insup Lee, and Mats Heimdahl. Challenges in the regulatory approval of medical cyber-physical systems. In *Embedded Software (EMSOFT), 2011 Proceedings of the International Conference on*, pages 227–232. IEEE, 2011. 26

Camille L Stewart, Jane Mulligan, Greg Z Grudic, Victor a Convertino, and Steven L Moulton. Detection of low-volume blood loss: compensatory reserve versus traditional vital signs. *The journal of trauma and acute care surgery*, 77(6):892–7; discussion 897–8, December 2014. 94, 97

Michael E Stokes, Xin Ye, Manan Shah, Katie Mercaldi, Matthew W Reynolds, Marcia F T Rupnow, and Jeffrey Hammond. Impact of bleeding-related complications and/or blood product transfusions on hospital costs in inpatient surgical patients. *BMC health services research*, 11(1):135, January 2011. 93

Johan AK Suykens and Joos PL Vandewalle. *Nonlinear Modeling: advanced black-box techniques*. Springer Science & Business Media, 2012. 16

Zaiyong Tang, C. de Almeida, and P. A. Fishwick. Time series forecasting using neural networks vs. Box- Jenkins methodology. *Simulation*, 57(5):303–310, November 1991. 20

The Epsilon Group. UVa/Padova T1DM Metabolic Simulator, 2016. URL `http://www.tegvirginia.com/T1DM.htm`. [Online; accessed 8-July-2016]. 18

Natalia A Trayanova. Whole-heart modeling applications to cardiac electrophysiology and electromechanics. *Circulation Research*, 108(1):113–128, 2011. 18

Robert Trowbridge and Scott Weingarten. Clinical decision support systems. *Making health care safer: a critical analysis of patient safety practices. Evidence Report/Technology Assessment*, (43), 2001. 4

Herbert JAF Tulleken. Grey-box modelling and identification using physical knowledge and bayesian techniques. *Automatica*, 29(2):285–308, 1993. 17

Harry L Van Trees. *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004. 36

Marion Verduijn, Lucia Sacchi, Niels Peek, Riccardo Bellazzi, Evert de Jonge, and Bas a J M de Mol. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. *Artificial intelligence in medicine*, 41(1): 1–12, September 2007. 23

Mats Vibergt. Subspace-based Methods for the Identification of Linear Time-invariant Systems. *Automatica*, 31(12):1835–1851, 1995. 20

Shyam Visweswaran and Gregory F Cooper. Learning instance-specific predictive models. *Journal of Machine Learning Research*, 11(Dec):3333–3369, 2010. 25

Shyam Visweswaran, Derek C Angus, Margaret Hsieh, Lisa Weissfeld, Donald Yealy, and Gregory F Cooper. Learning patient-specific predictive models from clinical data. *Journal of biomedical informatics*, 43(5):669–685, 2010. 25

Saul N Weingart, Maria Toth, Daniel Z Sands, Mark D Aronson, Roger B Davis, and Russell S Phillips. Physicians' decisions to override computerized drug alerts in primary care. *Archives of internal medicine*, 163(21):2625–31, November 2003. 6

Stuart A Weinzimer, Garry M Steil, Karena L Swan, Jim Dziura, Natalie Kurtz, and William V Tamborlane. Fully automated closed-loop insulin delivery versus semiautomated hybrid control in pediatric patients with type 1 diabetes using an artificial pancreas. *Diabetes care*, 31(5):934–939, 2008. 18

Jenna Wiens and JV Guttag. Active learning applied to patient-adaptive heartbeat classification. In *Advances in neural information processing systems*, pages 1–9, 2010. 23, 25

Peter Young. Parameter Estimation for Continuous-Time Models-A Survey. *Automatica*, 17(1):23–39, 1981. 19

Lazaros Zafeiriou, Mihalis a. Nicolaou, Stefanos Zafeiriou, Symeon Nikitidis, and Maja Pantic. Learning Slow Features for Behaviour Analysis. *2013 IEEE International Conference on Computer Vision*, pages 2840–2847, December 2013. 23

Ying Zhang and Peter Szolovits. Patient-specific learning in real time for adaptive monitoring in critical care. *Journal of biomedical informatics*, 41(3):452–60, June 2008. 25

Ying Zhu. Automatic detection of anomalies in blood glucose using a machine learning approach. *Journal of Communications and Networks*, 13(2):125–131, 2011. 21

Eva Zöllei, Viktória Bertalan, Andrea Németh, Péter Csábi, Ildikó László, József Kaszaki, and László Rudas. Non-invasive detection of hypovolemia or fluid responsiveness in spontaneously breathing subjects. *BMC anesthesiology*, 13(1):40, January 2013. 94